

Collaborative Technologies and Data Science in Artificial Intelligence Applications

A.Hajian, N. Baloian, T. Inoue, W. Luther (Eds.)

Proceedings from the 2nd Codassca Workshop
Yerevan, Armenia 2020



λογος

Aram Hajian, Nelson Baloian, Tomoo Inoue, Wolfram Luther (Eds.)

Collaborative Technologies and Data Science in Artificial Intelligence Applications

Logos Verlag Berlin



Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Logos Verlag Berlin GmbH 2020

ISBN 978-3-8325-5141-4



Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10,
12681 Berlin

Tel.: +49 (0)30 / 42 85 10 90

Fax: +49 (0)30 / 42 85 10 92

<http://www.logos-verlag.de>

**Aram Hajian, Nelson Baloian, Tomoo Inoue,
Wolfram Luther (Eds.)**

**Collaborative Technologies and Data Science in
Artificial Intelligence Applications**

**2nd International Workshop at the American University of
Armenia, College of Science & Engineering**

September 14 to September 17, 2020

Co-organized with IEEE Computer Society Armenia Chapter

Revised contributions

Volume Editors

Aram Hajian

American University of Armenia, College of Science and Engineering
40 Marshal Baghramyan Ave, Yerevan 0019, Armenia
Email: ahajian@aua.am

Nelson Baloian

Department of Computer Science, Universidad de Chile
Blanco Encalada 2120, Santiago 6511224, Chile
E-mail: nbaloian@dcc.uchile.cl


Tomoo Inoue

University of Tsukuba, Faculty of Library, Information and Media Science
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
E-mail: inoue@slis.tsukuba.ac.jp

Wolfram Luther

University of Duisburg-Essen, Scientific Computing, Computer Graphics, and Image Processing
Lotharstraße 63, 47057 Duisburg, Germany
E-mail: wolfram.luther@uni-due.de

ACM Subject Classification (1998): H.1.1, H.1.2, H.5.2, H.5.3, I.2.x, I.6, J.1, J.5

This work has been published in print by Logos Verlag, Berlin. Some rights are reserved, especially the right to distribute printed copies. The material and parts of it may be used in accordance with the creative commons licence stated:  CC BY-NC-SA. This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.

Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from the editors or authors. Violations are liable for prosecution under the German Copyright Law.

The Open Access publication of this title under CC BY-NC-SA 4.0 licence was made possible with funds from the Publication Fund of the University of Duisburg-Essen.

PREFACE

Following the successful CODASSCA2018 which was held two years ago at the American University of Armenia, we have the pleasure to invite interested experts and scientists to contribute to the topics of the conference and to participate in the 2nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2020). This event was originally planned to take place from 25 to 28 August 2020 at the American University of Armenia (AUA) in Yerevan directly after a DAAD-funded summer school organized by the AUA and the University of Duisburg-Essen on Enhancements of Deep Learning Systems for Intelligent Applications and the Connected Society. Due to the CoViD19 pandemic and the worldwide restrictions, it was still not clear after the submission deadline whether the two events would take place at the AUA in the usual form and the guests from all over the world, would be able to come to Yerevan, meet the students and colleagues from Armenia, and return home safely. At the beginning of June, the editors therefore decided to postpone the date until the end of the year or, alternatively, to organize the workshop online and publish the proceedings simultaneously in electronic form only during September.

Research in AmI and SmE in Urban and Rural Areas presents great challenges: AmI depends on advances in sensor networks, artificial intelligence, ubiquitous and persuasive computing, knowledge representation, spatial and temporal reasoning. SmE builds upon embedded systems, smart integration, and an increasing fusion of real and virtual objects in the IoT. Customized sensor networks are used to detect human behavior and activities, evaluation logic and process mining are needed to replace people's cognitive abilities in Ambient Assisted Living (AAL) applications, detecting recurring activities without being noticed and hurting their privacy. As digitization has become an integral part of everyday life, data collection has resulted in the accumulation of huge amounts of data that can be used in various beneficial application domains.

Effective analysis, quality assessment and utilization of big data is a key factor for success in many business and service domains, including the domain of smart systems. However, a number of challenges must be overcome to reap the benefits of big data. As big data handles large amounts of data with varying data structures and real-time processing, one of the most important challenges is to maintain data security and adopt proper data privacy policies. In general, there is a strong need to gain information of interest from big data analysis and at the same time, prevent misuse of data so that people's trust in digital channels is not broken. To ensure data quality, accurate results and reliable analysis support in health care applications, additional collaboration issues, privacy and security requirements are addressed within a throughout verification and validation management.

This workshop has attracted paper submissions which deal with the challenges mentioned above. The studies are in specialized areas and show novel solutions. Especially interesting are approaches based on existing theories suitably applied.

The succeeding talks in this volume are organized thematically: Technical Challenges and Edge-Cutting Technologies for Smart Environments, Smart Human Centered Computing, Data Science and Information-Theoretic Approaches for Smart Systems, and Artificial Neural Networks and Deep Learning.

The editors and the organizers Aram Hajian, Nelson Baloian, Gregor Schiele would like to express their gratitude to the German Research Foundation (DFG) and the German Academic Exchange Service (DAAD) for funding our common activities. Finally, we want to thank Yanling Chen and Ashot Harutyunyan for their ongoing encouragement and support and all participants for their presentations and contributions to the workshop and this proceedings volume.

Yerevan, Tsukuba, Duisburg, September 2020

The Editors: Aram Hajian, Nelson Baloian, Tomoo Inoue, and Wolfram Luther

CONTENTS

2 nd CODASSCA Workshop: Collaborative Technologies and Data Science in Smart City Applications	
<i>A.Hajian, N. Baloian, T. Inoue, and W. Luther</i>	v
Accepted contributions	vii
TECHNICAL CHALLENGES AND EDGE-CUTTING TECHNOLOGIES FOR SMART ENVIRONMENTS	
Armenian Khachkars: Towards an Automated Handling of Segmentation and Classification	
<i>N. Baloian, D. Biella, W. Luther, B. Panay, S. Peñafiel, J. A. Pino</i>	1
GPS Drawing on Street Networks: Extracting Routes from Polygonal Coverings	
<i>N. Baloian, D. Biella, W. Luther</i>	11
A Batching Cloaking Scheme for Continuous Location-Based Services	
<i>C. Faúndez, P. Galdames, and C. Gutierrez-Soto</i>	24
Protecting query privacy through semantic caching in location-based services	
<i>F. Vera-Catricura, P. Galdames, C. Gutierrez-Soto, and A. Curiel</i>	30
Automatic image classification supported by expert knowledge	
<i>S. Peñafiel, B. Panay, N. Baloian, J. A. Pino, and W. Luther</i>	35
Small is Beautiful - Temporal Accelerators for Embedded FPGAs	
<i>C.Cichiwskyj and G. Schiele</i>	38
SMART HUMAN CENTERED COMPUTING	
Online Collaborative Question Refinement Method to Increase Students’ Posed Question Quality	
<i>A. Nugraha, I. A. Wahono, J. Zhanghe, and T. Inoue</i>	42
Promotion of continuous use of a self-guided mental healthcare system by using a chatbot	
<i>T. Kamita, A. Matsumoto, T. Ito, and T. Inoue</i>	49

Validation and Verification Assessment of Genetic Counseling and Testing <i>E. Auer and W. Luther</i>	61
Explaining and visualizing recurrent neural network decisions <i>D. Qaramyan and H. Khachatrian</i>	73
Revisiting the Promotion Effectiveness Measurement in Retail <i>N. Baloian, J. Frez, C. Fuenzalida, B. Panay, S. Peñafiel, J. A. Pino, and H. Sanson</i>	76

DATA SCIENCE AND INFORMATION-THEORETIC APPROACHES FOR SMART SYSTEMS

Enriching Word Vectors with Morphological Information <i>M. Mirakyan and H. Khachatrian</i>	86
The role of alignment of multilingual contextualized embeddings in zero- shot cross-lingual transfer for event extraction <i>K. Hambardzumyan, H. Khachatrian, and J. May</i>	97
On the Tradeoff Between Accuracy and Fairness in Representation Learning <i>T. Galstyan and H. Khachatrian</i>	101
Excess-Risk consistency of group-hard thresholding estimator in Robust Estimation of Gaussian Mean <i>A. Minasyan</i>	105
Current approaches and challenges for the two-party privacy-preserving record linkage (PPRL) <i>Y. Chen</i>	108
Inner Bound of E-capacity-Equivocation Region for the Generalized Wiretap Channel <i>M. Haroutunian</i>	117

ARTIFICIAL NEURAL NETWORKS AND DEEP LEARNING

A survey on deep semi-supervised learning algorithms <i>A. Vanyan and H. Khachatrian</i>	123
On Machine Learning Powered Theorem Prover for Propositional Fragment of Minimal Logic <i>A. Baghdasaryan and H. Bolibekyan</i>	135
Estimating Efficient Sampling Rates of Metrics for Training Accurate Machine Learning Models <i>T. A. Bunarjyan, A. N. Harutyunyan, A. V. Poghosyan, A.J. Han Vinck, Y. Chen, and N. A.Hovhannisyan</i>	143
W-TSF: Time Series Forecasting with Deep Learning for Cloud Applications <i>A. Poghosyan, A. Harutyunyan, N. Grigoryan, C. Pang, G. Oganesyany, S. Ghazaryan, and N. Hovhannisyan</i>	152
Fingerprinting Data Center Problems with Association Rules <i>A. N. Harutyunyan, N. M. Grigoryan, and A. V. Poghosyan</i>	159
Intelligent Troubleshooting in Data Centers with Mining Evidence of Performance Problems <i>A. N. Harutyunyan, N. M. Grigoryan, A. V. Poghosyan, S. Dua, H.Antonyan, K. Aghajanyan, and B. Zhang</i>	169
Learning Data Center Incidents for Automated Root Cause Analysis <i>A. V. Poghosyan, A. N. Harutyunyan, N. M. Grigoryan, and N. Kushmerick</i>	181

Armenian Khachkars: Towards an Automated Handling of Segmentation and Classification

Nelson Baloian¹, Daniel Biella², Wolfram Luther²,
Belisario Panay¹, Sergio Peñafiel¹, José A. Pino¹

¹ Department of Computer Science, Universidad de Chile, Santiago, Chile
{nbaloian, bpanay, spenafie, jpino}@dcc.uchile.cl

² Centre for Information and Media Services, Computer Science and
Applied Cognitive Science, University of Duisburg-Essen UDE, Germany
{daniel.biella, wolfram.luther}@uni-due.de

Keywords: Virtual khachkar museum · metadata · segmentation · classification · deep learning

Extended Abstract

The DiKEViMA project [1] seeks to develop a virtual khachkar museum with the aid of engaged volunteers. UNESCO [2] describes cross stones as follows: “Khachkars reach [human dimensions of] 1.5 meters in height, and have an ornamentally carved cross in the middle, resting on the symbol of a sun or wheel of eternity, accompanied by vegetative-geometric motifs, carvings of saints and animals. Khachkars are created usually using local stone and carved using chisel, die, sharp pens and hammers.” Important resources can be found in the books *Armenia sacra* [3], *l’art des khachkars* [4] and the contributions of Patrick Donabédian [5], who helped us work out a historical classification of khachkar styles, periods and locations. Table 1 summarizes the periods and styles in the development of khachkar craftsmanship.

Table 1: Khachkar styles, periods and locations

Timespan	Location & Time	Style
5 th –8 th centuries	Garnahovit (Գառնահովիտ), Arutj (Արուճ), Gougark, Shirak, Talin and Mren	Early Christian vertical monuments, widespread in both Armenia and Georgia, composed successively of a pedestal, a cubic base, a four-sided stele, a capital, and a cross. From a historical point of view this type is a “predecessor” of the khachkar. More than 200 examples of this kind of memorials with four-sided stelae surmounted with a cross, have been studied by G. Grigoryan [6].
9 th –10 th centuries	Ani, Dvin, Talin Etchmiadzin 996; Haghpat arch 1004	Earliest khachkars: Among their characteristic features: the simplicity of their decoration; the absence of ornament on the cross, edge band and bottom; the rounded shape of the upper part of the plate, which progressively becomes rectangular; and the tips of the cross arms shaped as a single ball. Also found on the earliest khachkars: a pair of half palmettes or stylized hands, elegantly curved from the foot of the cross; the stepped pedestal and the round medallion under the cross; the two small medallions on both sides of the upper arm of

		the cross; the leaves or fruits hanging in the upper area. These features are present from the 9 th century until the present day. Several functions / intentions for erection: glorification of the cross, funerary monument, commemorative monument, apotropaic mark, monument of victory, landmark, decoration of churches and civic buildings.
10 th –11 th centuries. Armenian Bagratid kingdom with its capital at Ani; Seljuk Turks invasion, in the 2 nd half of the 11 th century	Ani, Sanahin, Syunik, Artsakh	The upper part of the stone plaque receives its definitive rectangular shape. At the extremities of the cross arms, the earlier single ball is replaced by a triple ball. A horizontally disposed pair of palmettes attached by a large belt to the foot of the cross is one of the characteristic innovations of the beginning of the 11 th century. Ornamented cross, vegetative sprouts surrounding the cross from above and below, and enlarged borders with squares and double or triple filament trellis work are further features.
12 th –14 th centuries. Crusades, Kingdom of Georgia, Kingdom of Cilician Armenia, Mongols, Mameluks	Geghard stones carved in 1213 by masters <i>Timot</i> and <i>Mkhitar</i> ; Haghpat “Amenaprkich” khachkar carved in 1273, probably by master <i>Vahram</i> ; Goshavank, carved in 1291 by master <i>Poghos</i> in the Vayots Valley and Syunik; Master Momik at Noravank	Crosses under a semicircular arch with narrow semicolumns or under an ogee arch, horizontally spread palmetto leaves under the cross, characteristic of the beginning of the 11 th century, now with more and larger decoration. Very abundant decoration. Complex interlaces on the edges with double or triple ornamental ties and rows of “Islamizing” eight-pointed stars. Rich arabesques. At the end of the cross arms, the central element of the triple ball acquires a pointed tip, which transforms it into a bud or a blossom, underlining the vegetable, living nature of the wood from which the cross is made. Introduction of two small lateral crosses in the lower quadrants, often held by a human arm. Under the cross, the traditional round medallion, enlarged and richly adorned, sometimes acquires a bulging aspect. Apparition of Christ in glory flanked by angels on the lintel; apparition of Christ on the cross, on four Amenaprkich cross stones sculpted in 1273, 1279, 1281 and 1285; Adam’s head under the cross; the image of the donor/deceased as a mounted hunter on the lower part of the plate; images of birds even sooner.
15 th –17 th centuries. Ottomans, Safavids	Master Kiram at Noratus; Masters Arakel and Melikset. Khachkars in Old Julfa, completely destroyed in 1998, 2002, and 2005	Khachkars incorporate carvings with wide frames in which crosses in various shapes are included, blended with the ornamentation, three or more smaller crosses, organically included in the ornamentation and blended with it: carvings more stylized and higher in relief, rigid and exact. Carvings bring strong light, shadow and plasticity, and fine carvings of winding stalk-like volutes (widespread since the end of the 12 th century). Creation of the Julfa type of elongated khachkar with crosses under deeply carved niches with stressed ogee arch. On human figures, round faces with almond-shaped eyes. On the lintel, along with triumphal scenes of Christ in glory, image of a double griffin with one human head.

In recent years, we have focused our development on the viable Virtual Museum metadata standard ViMCOX in the context of existing standards like LIDO and the realization of the multipurpose system ViMEDEAS (Virtual Museum Exhibition Designer Using Enhanced ARCO Standard). Smaller editors to design and generate virtual 3D and 2D museum environments or to publish and archive virtual exhibition layouts were developed in parallel [7], [8]. Metadata concerns the following attributes:

- **Encoding:** for machine readability, data types, processing, communication, exchange and storage
- **Structuring/Classification:** Categories, hierarchies, sets, elements, relations, indexing, referencing, linking with similar items
- **Naming:** Headings, types, values, controlled vocabularies, metrics, multilingual support, (fuzzy) search and retrieval support (ontologies)
- **Content:** 3D scene graph modeling, texturing and lighting, assets, objects, identifiers and various attributes, connectors, metaphorical design
- **Presentation:** Various exhibition environments, user support, tour planning, navigation support, co-curation support, interaction, publication, knowledge creation.

Excellent photographs and rich metadata are provided and complemented by their codes as was first proposed by the French engineers and experts Haroutioun Khatchadourian and Michel Basmadjian in their book *L'art des khachkars: Les pierres à croix arméniennes d'Isphahan et de Jérusalem* [4]. The work highlights the iconographic and epigraphic corpus of both locations and proposes alphanumeric reference codes to classify the stones with respect to their

- Inscriptions by handling encoding and ligatures, transcription of toponyms, anthroponyms and uppercases to lowercases, abbreviations and logograms.
- Typology using a repository of anterior partitions, borders, structures, crosses, ornaments, plants, flowers, trellis, symbols and so one
- Epigraphy using a grammar, non-terminals (epigraphy, formula, complement, dedicating, name, title, patronym, origin, dating, etc.), production rules, and terminals
- Ontology for local neighborhood relations.

More precisely, partitioning of the khachkars' surface considers the upper part (coded MFnn, with nn being a number that identifies the pattern used), sides (MMnn) and bottom (MBnn), as well as structures with zero to four side elements (Snn). The inner areas include complex motifs, for example (MCnn), as a superordinate structure the Xoran—the frame, cross and object under the cross—(TXnn, MBnn), interiors with cross type and composition (MXnn, CXnn), frames with cross type (MTnn), frames with crosses and attributes (ATnn), typical plate schemes with fixed and varying attributes Axx, Cxx, Exx, Pxx, cross over base ornamental or symbolic element (MEnn), ornaments with simple motifs (MS-*nnn*), complex motifs (MC-*nnn*), linear compositions (CL-*nnn*), arched compositions (CA-*nnn*), centered compositions (CC-*nnn*), and cruciform compositions (CX-*nnn*).

The book represents an important scientific advance; however, it offers no hierarchical coding of the ornaments and their compositions, no digital tool support (using a computer application) and only a limited variety of khachkars from two places—Isfahan and Jerusalem. Therefore, the repositories need completion: without tool support,

image and pattern segmentation and recognition, no automatic type recognition or classification is possible.

In the following, we propose some modifications to the classification scheme of khachkars' metadata in order to allow automatic segmentation on three levels. The next step enables the motifs to be assigned to their classes: border structures, frames, crosses, geometric, vegetable, and figurative objects, as well as areas of repeating patterns. Further details in the assignment remain possible.

Figure 1a shows an example of classification of a khachkar according to [4] and using the following elements. Figure 1b a first step towards segmentation.

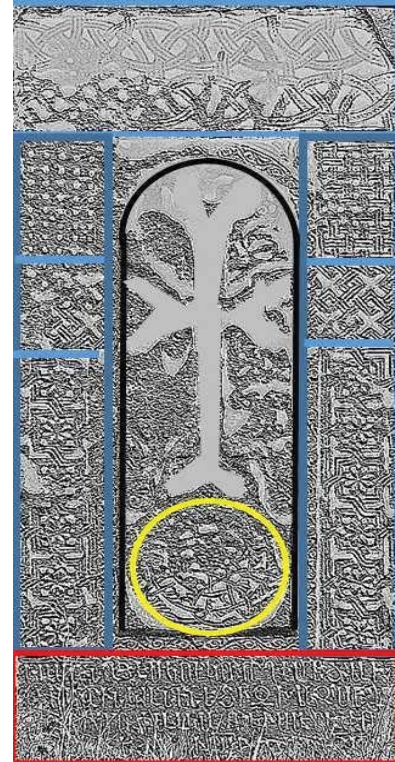


Fig. 1a. Grigor's Khachkar, Etchmiadzin's Saint Gayane Church, partly destroyed (D), late 12th–13th century; 110x60x21 cm³, inscription on the banner below in three lines: Christ, when you come back, remember Grigor and Ohan (P. Donébadian), [3,171]

Fig. 1b. Khachkar in memory of P'alik Noravank monastery, 1285, dimensions 164x70x26 cm³, [3] p. 326, Stbls123rs123 Ftq1q2lsrs XC

Simple motifs MS: geometric objects, stylized parts of plants, fruits, figurative motifs, simple cross shapes, braidlike structures or intricate parallel curves

Complex motifs MC: composition of simple motifs is repeated, (a)symmetrical arrangement, detailed geometric objects, stylized plants and fruits, figurative motifs, wickerwork made of biomaterial or bast

Linear Composition CL: Strip strewn with small geometric objects showing knots in trellis, geometric objects or parts of plants in regular formation

Circular composition CC: Vegetable motifs and wickerwork in a circular arrangement

Cross compositions CX: Crosses consist of two orthogonal narrow rectangles, the horizontal shorter than the vertical, which may be bent and decorated with four simple ends with various simple geometric or vegetable shapes. The body of the cross uses all elements of the CL and MS.

Compositions from columns and arches CA: Columns and arches surround the cross composition and are part of the *xoran*. They are filled with simple geometric patterns. The arch on top can end in a tip or be replaced by smaller arcs.

The *xoran X* includes the cross, its frame, the base and support below the cross. The center of the khachkar can contain several grouped *xorans*. The bordered area of the cross can show various simple motifs around the body of the cross or other smaller crosses. The arms of the cross can be filigree, stylized, delimited by curves or straight lines and may contain further motifs at the ends. The interior of the cross may contain a variety of geometric motifs; the contour is formed by multiple parallel curves that end in various leaf shapes. The interior is often decorated with different geometric objects. Khatchadourian and Basmadjian [4] describe 72 different types of CX-nnn.

Border structures S: Consist of up to four rectangles on the edges of the stone above, below, and on the right and left sides. They record usually simple but sometimes complex figurative or geometric objects (in repetition). Inscriptions (→Epigraphy [4]) are usually in the lower areas.

The automated handling of the collection and classification of cross stones includes the following tasks:

Development of a metadata standard for the structure and content of khachkars based on [4] and the modifications proposed here.

First, spatial structural elements are defined and identified. Second, the structuring is refined, and objects are categorized. Third, substructures are recognized, and objects classified.

The border structure of the four sides of a khachkar—top, bottom, left side, and right side—is as shown in Figure 2:

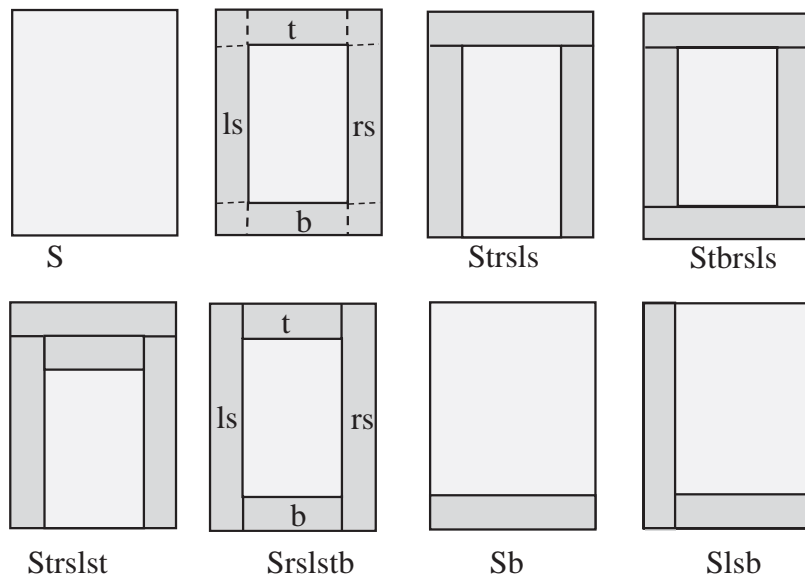


Fig. 2. Border elements of the Khachkar

Rectangular elements are subdivided and sometimes replaced by polygonal forms, as can be seen in Figure 3. Structural elements t, b, ls, rs can be divided and filled with various content.

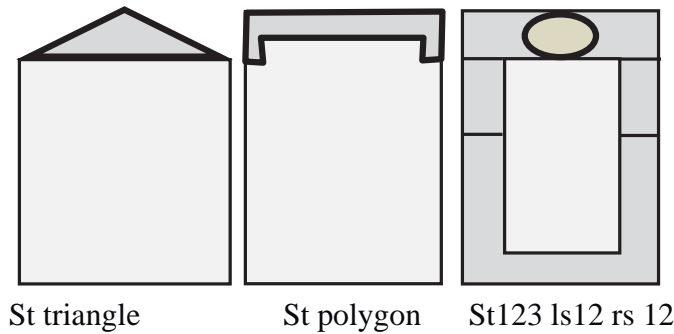


Fig. 3. Three examples of how rectangular elements can be replaced by polygonal forms and their coding, starting with St and adding the name of the changed form

Further frame shapes separate border content and the inner areas of the stone with the cross in the middle: straight lines, circular arcs and rectifiable Jordan arcs. Frames are described by their components above and below, on the right and left. Alternatively, there may be also circles or polygons with circular edges, hemicycles (hc) and quadrants (q) in the inner part of the khachkar. Arched compositions CA allow the quarter circles to culminate in a point, put smaller arches atop each other, use parallel curves, or decorate the frames inside as seen in Figure 4.

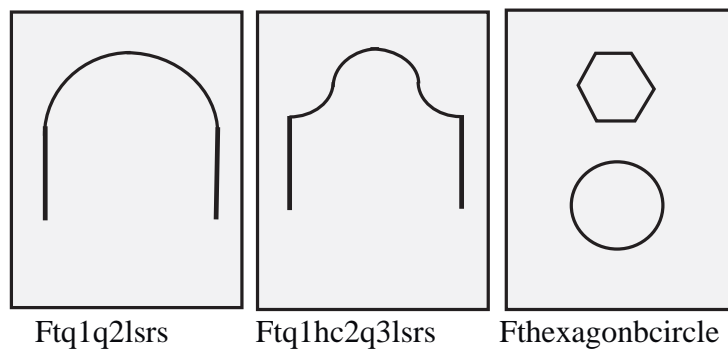


Fig. 4. Examples of additional decorative figures in the inner part of the khachkar surrounding the cross itself with their description code which starts with Ft.

Primitive (P), simple (E), complex (C), and atypical forms (AF) of a khachkar, symbols and their arrangements, compositions of border structure, xoran with frame, cross and base differ in the execution of their three components: the use of simple and complex motifs, their arrangement and the elaboration of the interspaces, up to an asymmetrical arrangement of the attributes (see Figure 5).

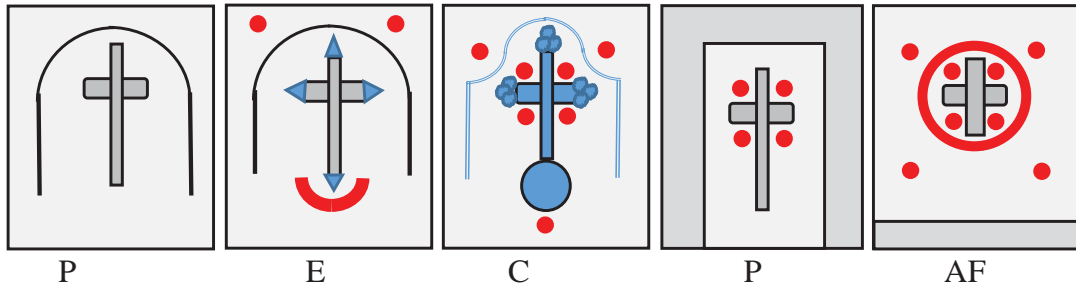


Fig. 5. Examples of types of compositions: primitive (P) simple (E), complex (C) and atypical (AF).

There are ready-to-use software solutions based on deep learning and neural networks for segmenting scenes from various figurative, vegetable, urban or rural contexts. The software can be downloaded. We are currently assessing whether it is possible to use it to segment the khachkars (<http://deepscore.cs.uni-freiburg.de/#demo>). A software implementation of this project (AdapNet ++, SSMA, AdapNet, CMoDE) based on TensorFlow can be found in the Freiburg University GitHub repository for academic use and is released under the GPLv3 license. Another approach providing software is found at <http://deepscore.cs.uni-freiburg.de/#demo>. A visual language for describing khachkar iconography should be developed in parallel.

Implementation proposal

The first challenge in classifying the khachkars automatically is the background noise in these images. Khachkars are generally located in places with a lot of vegetation and other ancient buildings. These elements do not provide any useful information for the classification and thus should be eliminated to increase performance and avoid classifiers behaving incorrectly due to ineffective information or undesirable correlations. For example, in object detection problems, several of the best-known models were unable to distinguish between wolves and Eskimo dogs (huskies) when trained with snow in the background of the images containing wolves [9].

One alternative to removing the background automatically is to use image segmentation techniques [10]. Image segmentation models divide an image into semantically similar structures. For example, for an urban photo, these models can identify which parts of an image are buildings, people or cars. Several models have been proposed to solve image segmentation. The ones that perform best are convolutional neural networks (CNN). The most commonly used architectures for these problems include U-Net [11] and Feature Pyramid Networks [12]. In our case, we can use these trained models to detect the khachkar and the background for different classes with no human intervention. Finally, we can remove all the other categories and get a clean image of each khachkar.

Figure 6 presents an example of applying the segmentation algorithm to subtract the background and extract the khachkar. The image on the left is a sample khachkar with a background that is just noise for classification purposes; on the right is the same image with its background cropped by applying segmentation to the image.



Fig. 6. The image at the right presents the result after applying segmentation to the image on the left.

After analyzing the part of the photo containing the khachkar, another preprocessing task is to convert the image to greyscale. As explained above, the important properties of the khachkars for classification are mainly their shapes, regardless of the material; color of the stone; and textures. Therefore, converting the khachkar images to greyscale forces the model to focus on motifs and shapes. The segmentation approach also focusses on efficient discovery of curved lines and discontinuous edges in order to detect half circles on top of khachkars and surface areas that may have been broken or are missing [17].

For the image classification process itself, we believed that convolutional approaches like CNNs can operate correctly because these algorithms can recognize and classify emotions and their typical facial displays [13].

Figure 7 shows an example of applying a Sharpen filter to the greyscale khachkar image. On the left is the original khachkar image after removing the background, and on the right is the same image after applying the sharpen kernel. One can see that the image on the right has much greater detail in the khachkar's motifs, compositions and shape. Thus, it can be useful for classification.



Fig.7. The image at the right presents the result after applying a sharpening filter to the image on the left.

Our major concern about using CNN is the amount of data available for training the models. These network architectures require thousands of labeled samples per class to fit their inner parameters. In our case, data is limited; only about a hundred records are labeled correctly. An alternative means of tackling this problem is to transfer learning and fine-tuning techniques in the model [14]. These techniques use a pretrained convolutional model (generally in a different dataset) as an initial state for the classification in a new dataset. Findings show that many of the weights of the first layers can be reused regardless of the dataset and only the last layers are specific to the problem, with only a few parameters requiring adjustment. A similar technique is described by using CNN architectures with several different convolution layers to classify isolated glyph images [15].

Another approach that can be used to improve performance is scenes analysis techniques [16]. In this approach we perform the object classification stepwise. First, we detect the elements that compose a khachkar, for example, a border, crosses and writings. Then, using these basic objects and the already introduced layouts (xoran), we combine them and create a greater part of the picture. Doing so will help to find patterns already known and present in certain parts of the image, such as in the border or in the frame. To combine the elements, we can use another machine learning model or a model that supports uncertainty reasoning.

Acknowledgement

We dedicate this work to the author, scientist, and expert in historical architecture Samvel Karapetian, who died far too early on February 27, 2020. His advice and expertise have accompanied us for years. His untiring work for the historical heritage of Armenia, which he has documented in many books and lectures, remains unfinished. We have lost an irreplaceable supporter of our projects and a friend to whom we owe a debt of gratitude.

References

1. DiKEViMA project: http://www.vimedead.com/wordpress/?page_id=203
2. UNESCO <http://www.unesco.org/culture/ich/en/decisions/5.COM/6.1>
3. Durand, J., Rapti, I., Giovannoni, D.: Armenia sacra: Mémoire chrétienne des Arméniens. Musée du Louvre Editions. Somogy (2007).
4. Khatchadourian, H., and Basmadjian, M.: L'art des khachkars: Les pierres à croix arméniennes d'Ispahan et de Jérusalem. Geuthner, Paris (2014).
5. Donabédian, P.: Le khachkar, un art emblématique de la spécificité arménienne. In Augé (I.), Dedeyan (G.) dir. L'Église arménienne entre Grecs et Latins, fin XIe - milieu XVe siècle. Geuthner, Montpellier, France 151-168 (2009).
6. Grigoryan, G.: Early Medieval Four-Sided Stelae in Armenia. History Museum of Armenia, Yerevan (2012).
7. Sacher, D.: A generative approach to virtual museums using a new metadata format. A curators', visitors'; and software engineers'; perspective. Ph.D. Dissertation, University of Duisburg-Essen, Germany, Logos, Berlin (2017).
8. Baloian, N., Zurita, G., Pino, J. A., Peñafiel, S., and Luther, W.: Developing Hyper-stories in the Context of Cultural Heritage Appreciation, in Nakanishi, H., Egi, H., Chounta, I. A., Takada, H., Ichimura, S., Hoppe, H. U. (Eds.): Collaboration Technologies and Social Computing. Proceedings 25th International Conference, CRIWG + CollabTech 2019, Kyoto, Japan, September 4–6. LNCS 11677, Springer 110-128 (2019).
9. Ribeiro, M. T., Singh, S., and Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (2016).
10. Long, J., Shelhamer, E., and Darrell, T.: Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition (2015).
11. Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham (2015). <https://arxiv.org/pdf/1505.04597.pdf>
12. Seferbekov, S., Igloukov, V., Buslaev, A., Shvets, A.: Feature Pyramid Network for Multi-Class Land Segmentation. CVPR Workshops (2018).
13. Ruiz-Garcia, A., Elshaw, M., Altahhan, A., Palade, V.: Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. International Joint Conference on Neural Networks (IJCNN). IEEE 1586-1593 (2017).
14. Yosinski, J., Clune, J., Bengio, Y., and Lipson, H.: How transferable are features in deep neural networks? Advances in neural information processing systems 3320-3328 (2014).
15. Paulus, E., Hadi, S., Suryani, M., Suryana, I., Simanjuntak, Y. D.: Evaluating Ancient Sundanese Glyph Recognition Using Convolutional Neural Network. Journal of Physics: Conference Series. 1235 012063 (2019).
16. Kosir, A., Tasic, J. F.: Formal system based on fuzzy logic applied to digital image scene analysis. 11th IEEE Mediterranean Electro-technical Conference (IEEE Cat. No. 02CH37379), Cairo, Egypt 409-413 (2002).
17. Biella, D., Sacher, D., Weyers, B., Luther, W., Baloian, N., Schreck, T.: Crowdsourcing and Knowledge Co-creation in Virtual Museums, 21st International Conference, CRIWG 2015, Yerevan, Armenia, September 22-25, 2015, Proceedings LNCS 9334 Springer, 1-18 (2015).

GPS Drawing on Street Networks: Extracting Routes from Polygonal Coverings

Nelson Baloian¹, Daniel Biella², Wolfram Luther²

¹ Department of Computer Science, Universidad de Chile UCH, Santiago, Chile
nbaloian@dcc.uchile.cl

² Centre for Information and Media Services, Computer Science and
Applied Cognitive Science, University of Duisburg-Essen UDE, Germany
{daniel.biella, wolfram.luther}@uni-due.de

Abstract. GPS drawing needs basics of digital geometry and topology, road networks, GPS technologies, and metaphorical correspondences. The contribution classifies GPS art, network parameters and route generation algorithms that help to generate examples of GPS drawings in various web applications. Route generation algorithms construct polylines from covering polygon boundaries to approximate a given Jordan arc or compute the chain code of a polyline over a given grid and apply the code to replace the curve by a suitable sequence of crossing points of the street network which is then passed through. An outlook on the selection and finding of suitable route networks as canvas for the realization of a GPS artwork is given.

Keywords: GPS drawing · metaphors · road networks · digital geometry

1 Introduction

Smartphones, modern GPS technology and freely accessible geographical maps allow the planning and recording of trips or journeys by land, water or air. Map sections are enriched with important private and public information, routes can be annotated or highlighted in color and the results can be published online. Conversely, for a given start and end point and specified mode of travel, optimal routes can be determined and predefined geometric figures traced using that route. After a definition of the term, this contribution highlights the technical and algorithmic basics of GPS drawing, acquisition of street maps, creating a canvas, GPS accuracy, presents artwork creation algorithms and shows a variety of possible applications. Finally, the notion of fractal geometry is introduced to understand shapes of cities. We give an explanation why we cannot find regular road networks of any size.

As stated in [2], GPS art consists in drawing on a digital map following a given path and using a GPS device. The route of a journey can automatically be loaded into the GPS receiver's memory and then be shaped, embellished and visualized on a computer or smartphone display. The map image is obtained from various open or commercial sources, such as Google Maps (<https://www.google.com/maps>) or OpenStreetMap (<https://www.openstreetmap.org>).

This definition needs some explanation. Recording points reached sequentially on a map and following trajectories described in standardized three-dimensional coordinate systems or after projection in planar areas is a prerequisite for covering the distances traveled and visualizing them on a canvas as colored artifacts. Color and style of

artwork visualize GPS signal properties (brushstroke), route characteristics (line style), artist's performance, environmental influences, and other properties defined by the metaphorical design of the artwork.

There are various ways to describe geometrical objects, including parameterized rectifiable Jordan arcs, polygons, or artifacts such as letters or geometric objects used as overlays on a map. Routes belong to various categories, allowing for walking, running, cycling, driving or hiking. These categories include footways in a pedestrian zone or alleys crossing a city center, modern quarters with a rectangular road networks or bypasses and arterial roads into rural areas with field and forest paths, as well as lakes and mountain trails requiring 3D representation. Road networks are described using segments between node points with ellipsoidal 3D coordinates or projections into a plane together with various street objects. Segments are polylines, (circular) arcs, clothoids or, more generally, spline curves together with parameters and standardized exchange formats.

Waschk and Krüger [18] present a software system for planning and generating routes to represent apparently 2D GPS artwork, such as curve/polygon drawings or texts entered by the user via a stylus or keyboard. The system supports basic transformations and object conversion in polygons. Street data, such as intersection nodes, are obtained from the Open Street Map project. Users freely choose a location on the globe to create GPS artwork, and the algorithm computes a route fulfilling given quality criteria using appropriate metrics and visualizes the route. The authors do not provide references to basic results of digital geometry, digitization and grid coverings. Also, some examples are not easy to understand (cf. Figure 8). Important aspects, such as the selection of suitable road networks and permitted route categories are missing.

2 Acquisition of Street Maps: Creating the Canvas

The canvas for GPS drawings can be created in two different ways: (1) extracting roads, building footprints and other features in the areas between them based on satellite images and GPS data and mapping them via a suitable projection or (2) deriving a road network incrementally and constructing it based on a camera tour or the trajectories of many vehicles and their uploaded GPS data.

Schindler et al. [14] were the first to introduce various forms of data acquisition for the generation of road networks: laser scanner-based data acquisition, video-based lane detection, processing of aerial images, and input from other data sources such as maps or reference measurements. Babahajiani et al. [1] propose an efficient and accurate two-stage method to segment and semantically label 3D city maps of registered LiDAR point clouds and RGB street view images. Two applications of this approach concern model-based 3D visualization for better user experience and 2D semantic segmentation for 2D applications, such as GPS drawing in a subnet.

Satellite images in high resolution play an important role as mentioned in [11]. These images are kept in governmental, commercial and open repositories of surface maps with spatial object resolution descending from a diameter of 80 meters to 25 cm or less. The authors in [19] use deep convolution networks in an image segmentation approach as a solution for extracting road networks from high resolution GF-2 satellite images. The papers [10,3] share details of new deep-learning and weakly supervised training

models and make data and specialized map-editing, -reviewing and -verifying services available to the global mapping community through Map With AI.

Zhongyi Ni et al. [20] address the inverse problem: Vehicles equipped with positioning devices can generate and upload a huge amount of trajectory data in real time. Based on incremental learning, they propose a road network generation method.

Schüller [15] develops a geographic information system that can record, edit, query and visualize road condition objects and describes the creation of a street network editor with node points, segments, addresses, objects, coordinates and other geometrical data.

In order to approximate geometric objects or characters on a canvas showing foot-paths, streets, roads, hiking and ski trails, lakes, buildings, blocks, parks and recreational areas, various parameters of interest can be extracted from satellite images, aerial photographs or map material. For a rectangular area, the orientation, regularity and mesh size of the grid must be determined. If aligned global quadtrees or other locally aligned hierarchical structures are used, they are expanded as long as all leaves at the bottom level contain a street segment of the desired category. An interesting research question is devoted to the task of how areas can quickly be considered for their suitability as a canvas for a GPS art project.

3 GPS Accuracy

If all clocks are synchronized, with signals from three satellites, a receiver's location can be found as the common point of three spherical surfaces. A fourth satellite is used to include the time variable. In practice, there are various error sources; time, longitude, latitude, and altitude all have statistical error bounds. If we interpret the vectors of the signals emitted by the satellites used to calculate the position and time of the receiver as a brush, we need some information about the location, time and properties of the transmitters, the receiver, the space traversed by the signals and the requirements for visualizing the distances traveled by the user after a standardized projection on a 2D map. It should be borne in mind here that reliable distance calculations must take elliptical coordinates into account for greater distances and, as a third dimension, altitude.

The shortest path between two points on Earth, customarily treated as an ellipsoid of revolution, is called a geodesic. Two geodesic problems are usually considered: the direct problem of finding the end point of a geodesic given its starting point, initial azimuth, and length; and the inverse problem of finding the shortest path between two given points. Algorithms for the computation of geodesics on an ellipsoid of revolution are offered in [9]. These algorithms provide accurate, robust, and fast solutions to the direct and inverse geodesic problems and allow differential and integral properties of geodesics to be computed. Three independent systems will be available soon (USA GPS, RUS Glonass and now EU Galileo). Quality of service is set and described in the Global Positioning System (GPS) standard [6]. Standard Positioning Service (SPS) includes Position/Time Accuracy Standards, Global Average Position and so forth. Governments committed to broadcasting the GPS signal in space with a global average use range error URE smaller than 0.7 m with CI 95% on May 11, 2016 [16].

To describe satellites' and receivers' positions in space and on Earth, standardized national reference systems (e.g., WGS-84 ellipsoid) have been agreed, as have Cartesian and ellipsoidal coordinates and their transformations. Details are reported in Zogg

Collaborative generation	Travel from or to a location, airport or port, carried out simultaneously or with a time delay
Quality	Quality criteria and dimensions such as path length, duration or proximity to the given overlay drawing and attractiveness of the route

Creating a drawing depends on the motif, the objects depicted, the artist, and the painting tools—brush, paint, easel, frame and canvas on which the painting is made. Metaphorical correspondences must be found to the artist, brush and canvas styles.

Who or what is producing the drawing? A person or group may be walking, hiking, running, climbing, snowboarding, flying, or sailing. In any case, they are traveling from a starting point S to an endpoint E following a particular route—ground, air and/or sea. The drawing is produced either by means of transport, leaving marks along the route, or alternatively by an algorithm using a standardized description of the route based on GPS tracking of point coordinates or velocity. A typical task is to describe the impact of the environment (buildings, lakes, mountains) and satellites, signaling pathways, and receiver characteristics resulting in uncertain 3D and 2D GPS spatial coordinates in a ball or plane geometry—which influence the accuracy of the map—represented by color, thickness and the style of points or strokes. We could assign a color to a GPS point by using a color model corresponding to the number of satellites used and assign the point a given diameter corresponding to its uncertainty.

There are several scenarios for crowdsourcing and collaborative work: for an example see <http://www.wandermap.net/en/official/3071561-way-of-st-james/>.

The website shows various routes with data and digitized representations of ground features and associated attributes representing the knowledge of residents and volunteers to trace the Way of St. James from starting points in all European countries and leading to Santiago de Compostela in Galicia (Spain).

Strava [17] is a social fitness network using GPS data primarily to track cycling and running exercises. It enables users to achieve aesthetic and artistic goals by conveying a message and coding information about the route and its surroundings or about the traveler (i.e., data about training or about runners' or bikers' performances [12], global heatmaps etc.).

5 Algorithmic Approach

The creation of GPS art includes interesting algorithmic aspects starting with the description of the geometric art objects and limiting curves as well as text in Euclidean and 2D or 3D digital geometry. If geometric objects are approximated by their digital artifacts in two or three dimensions, terms such as *line*, *circle*, *Jordan arc*, *polygon*, *simply connected domain*, *distance*, and *neighborhood* must be adapted to their digital analogues. Then an appropriate canvas should be chosen—a network of roads to approximate the artwork fulfilling quality criteria under a given metric. Appropriate approaches are needed to find a sufficiently large regular grid with fine meshes in order to scale, position and fit the drawing accordingly. Geometric objects are modeled using rectifiable Jordan arcs or closed Jordan curves as boundaries. The Jordan curve theorem asserts that a closed Jordan curve divides a plane into an interior region bounded by the

curve and an exterior region. There is an analogous statement for the discrete plane using four- and eight-neighbor topologies and digital Jordan arcs.

QuadTiles in OpenStreetMap and quadtrees in GoogleMap API are data structures for map clustering to store markers in a certain area. A marker identifies a location on a map, can be labeled, and is searchable. There are similar octree structures in 3D to localize point clouds. For searching an item, latitude and longitude are encoded into a single integer using the QuadTile algorithm. This is done with 31-bit precision per coordinate, thus providing as many possible levels. Location searches become simple integer range-based queries. Thus arcs and a street network can be geocoded with quadtiles up to a certain level corresponding to the requirement to find the representation point of at least one allowed street segment from the network in each node of the lowest level of the tree structure, the leaves, which can serve as an approximation for the intersecting part of a given drawing.

In a stroke-based setting, distances between segments and their end points must be evaluated. After applying a fast in/out test for two polygonal areas in two dimensions and boxes in three dimensions containing two line segments for relevant streets and parts of the drawing, we apply a formula for distances between straight lines based on the cross product. Then, the nearest points must be found and moved to the endpoints if they are outside the boxes/polygons or segments.

To compute the distance between two line segments, three cases are possible: distance between two end points, distance between one end point and an interior point on the other segment, and distance between two interior points. The minimum value problem has a solution since both sets are compact, and the distance function has a minimal solution.

$L_1: \mathbf{x}(t) := \mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)$, $L_2: \mathbf{y}(u) := \mathbf{y}_1 + u(\mathbf{y}_2 - \mathbf{y}_1)$, $\mathbf{n} := (\mathbf{x}_2 - \mathbf{x}_1) \times (\mathbf{y}_2 - \mathbf{y}_1) / |(\mathbf{x}_2 - \mathbf{x}_1) \times (\mathbf{y}_2 - \mathbf{y}_1)|$
 \mathbf{n} is orthogonal to both straight lines. Then the distance between skewed straight lines in the 3D space is given by $\text{dist}(L_1, L_2) = \text{abs}(\mathbf{x}(t) - \mathbf{y}(u), \mathbf{n})$ for any two points $\mathbf{x}(t)$, $\mathbf{y}(u)$. We also use areas bounded by polygons, polygon clipping and decision algorithms such as a point in a polygon or the position of a point with respect to a plane, or the distance between two straight lines (line segments) in 3D and 2D, all based on the Hesse form. When a polygon intersects a rectangle, it can be decided which parts are inside and which outside—an important determiner for the construction of the closest route or trail. All parts of the figure in the overlay are approximated by connected segments of a road network with certain quality characteristics and specific optimization criteria (shortest or nearest) and distance measures.

When a straight line or a Jordan arc is digitized on a square grid, a sequence of grid points is obtained defining a digital straight-line (DSL)/curve segment. Whereas a regular (orthogonal) grid covering provides square midpoints or parts of the border in the four-neighbor topology, the Bresenham algorithm defines a major and a minor direction in which steps are executed. Beginning at the integer starting point, the algorithm delivers integer grid points on both sides of the line, depending on the underlying eight-neighbor topology. In the case of straight line segments and a circular arc (CA) with integer midpoint, integer radian square and an eight-neighbor topology, the digitized arc is the closest (and shortest) one [8]. The algorithm extends to three dimensions. Using characteristic properties of chain codes, it can be decided if a finite code represents a DSL or a DCA. Similar to our approach, the authors of [5] define a supercover model with a geometry based on irregular isothetic grids by tiling the plane using axis-

parallel rectangles and a digitization framework. As application, a supercover digitization of straight lines with recognition algorithms and a process to reconstruct an invertible polygonal representation of a curve are given while we want to assemble a path built from parts of the boundaries of the polygonal covering.

Algorithm 1: GPS Drawing: street approximation of a rectifiable Jordan arc $\mathbf{c}(t)$ with start and end points S and E in 2D

Take a network with streets of allowed categories, and define optimality criteria.

Construct a nearest polygonal covering $P:=\{P_1, P_2, \dots\}$ in which each member is bordered by segments of allowed streets (straight line segments or circular arcs) of a given category.

Collect an ordered list $L_c = \{\mathbf{c}(t_i), i=1, \dots, n\}$ of intersection points ($t_i < t_{i+1}$) of $\mathbf{c}(t)$ with the covering P (avoid corner points).

Start with $\mathbf{c}(t_1)$, and set $P_1=P_{j(0)}$ that contains a part of $\mathbf{c}(t)$, $t < t_1$.

For $i:=1$ to n : determine $P_{j(i)}$ with border point $\mathbf{c}(t_i)$, which contains a part of $\mathbf{c}(t)$; $t > t_i$, $P_{j(i-1)}$ contains a part of $\mathbf{c}(t)$, $t < t_i$, resulting in a list $L_p = \{(P_{j(i-1)}, P_{j(i)}), i=1, \dots, n, j=j(i)\}$ of adjacent polygons.

Try to assemble a path built from parts of the boundary $\partial P_{j(i)}$, $i=0, 1, 2, \dots$, such that the optimality criteria (boundary parts of each member of the polygonal covering used, double crossed sections eliminated, shortest or closest Jordan path constructed) are fulfilled. If there is no connected (continuous) path, try to spread to adjacent polygons or assume polygons with simply connected inner domains (with respect to the four-neighborhood topology), the edges of which can be traversed in both directions. If the arc passes the same polygon twice or is part of a closed Jordan curve, try to use roads/trails inside or outside.

Figure 2 shows an area of Windsor (UK) (<https://www.gpsvisualizer.com/draw/>), a given digital Jordan arc, an allowed covering also using footpaths and a drawing that uses all polygons as well as several variants resulting in shorter or outer/inner routes, including one that is not a Jordan arc.

Figure 3 on the left handles the letter **B**, which is rendered with three closed Jordan curves as a contour. A suitable road network is located in New York (USA). A canvas with a irregular orthogonal road network results in digitally convex inner paths and an outer path from the edges of the covering. Alternatively, the texture of the satellite map can be used to fill the inner domain of the letter B. A covering with a global quadtree, the squares of the lowest level of which contain both letter boundary and roads of the network, only exists for two of the three curves. An access with multigrids of the same depth and orientation solves the problem. On the left is shown a letter drawing using the nearest double-point free irregular grid digitization and chain code representation of inner and outer Jordan curves on a Manhattan city map. A more general approach starts with a classification of road networks of the selected road category based on extrinsic (orientation and size) and intrinsic parameters (parameters for irregularity: varying mesh size and main directions, intersections with fewer or more than four branches, missing sections and areas without roads).

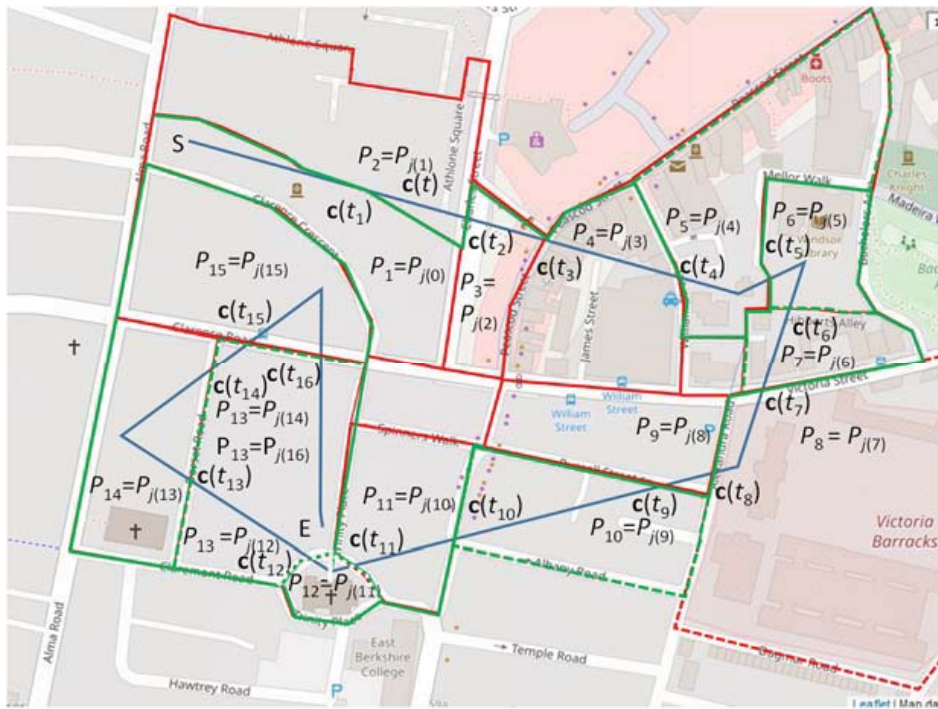


Fig. 2. Various realizations of routes for a given digital Jordan arc and an irregular covering

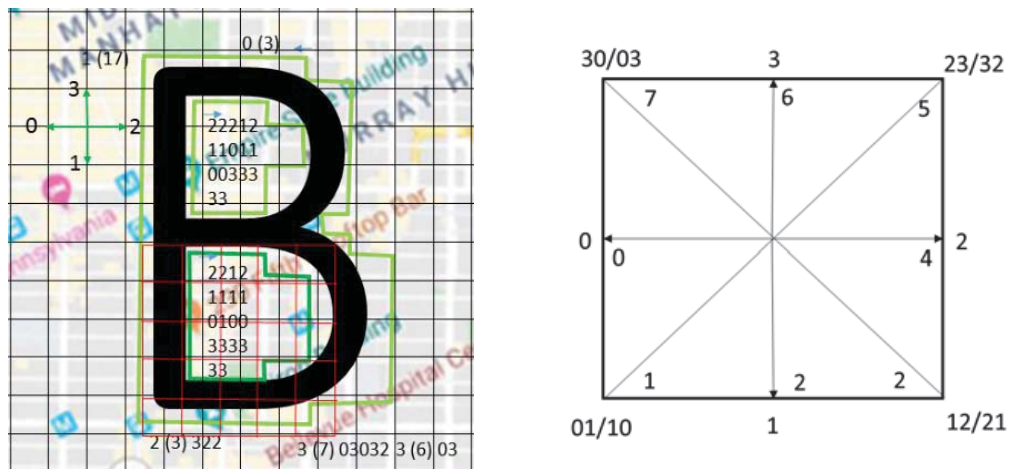


Fig. 3a. Drawing the letter B by walking in Manhattan · **Fig 3b.** 4- and 8-neighborhood topology

Other types of networks consist of circular paths (cf. Figure 4). There are now several important tasks: a) classifying urban and rural areas according to the categories described above; b) classifying the digitized drawings according to polygon types, dividing them into sections and arcs and into filled and unfilled contours; c) searching and fitting (with learning) to a canvas; d) freeform drawing and e) identifying cooperative variants or creating a 3D version.

Next, we review these problems from a software perspective. The relevant data types and algorithms to describe the road networks are graphs, trees (quadtrees) or lists with alphanumeric entries and real or interval-valued scalars or vectors.

The *canvas* is a scaled, aligned rectangular (or polygonal) area with a road network of a category specified in lists of (quasi-) orthogonal streets with the following (regular;

optional) parameters: translation, rotation, map scaling, horizontal street names, vertical street names; list of gaps, minimum and maximum deviation of orthogonality, and minimum and maximum mesh diameter. An irregular canvas contains various aligned or oriented road networks, circular roads, larger meshes or gaps such as parks, lakes, inaccessible or privately used areas.

Alignment and *scaling* increase the clarity of the discussion, particularly when describing the neighboring intersections to the west, south, east and north, but it must be specified whether calculations are made in original or new coordinates.

Road networks consist of streets, intersections, meshes and gaps with the definition

- *Street*:= (name, town/county; start point, end point, category, travel direction);
- *Intersection point*:= (name & name, coordinates; adjacent intersection points AIP0, ..., AIP3, optional AIPs)
- *Mesh* (IP):=(array IP[X,Y] of adjacent intersection points (IP_{x,y}) (IP_{0,0} in the southwest of the area)
- *Gap* (IP):= Polygonal area delimited by roads, not suitable for drawings due to missing paths).

Next we describe polyline rasterization:

Algorithm 2: Text/Polyline rasterization

Input polyline (p_0, \dots, p_n) , $p_i \in \mathbf{Z} \times \mathbf{Z} \forall i$. For each line segment, Bresenham's algorithm is used to rasterize it.

The starting point S and end point E of the segment are specified using integer coordinates, and the dominant axis X and the minor axis Y of the octant containing the line segment are determined.

Each step is coded via an eight-neighbor code 0, ...,7, where only two consecutive numbers (mod 8) are used. This gives the chain code of the segment. If necessary, the direction of the code sequence must be reversed, c changes to $7-c$.

Finally, the individual code sequences are combined to form the polyline code and then transformed into a four-chain code; in the case of two possible pairs, these are selected alternately.

Next, we digitize the drawing on a regular grid, encode it with a four-directional chain code and then visualize it on a suitable canvas of a road network.

Algorithm 3: GPS drawing (cf. Figure 5)

Choose a regular grid (X, Y) and coordinates $0 \leq x \leq X$, $0 \leq y \leq Y$, with a given mesh size.

Digitize the drawing as polyline and describe it with a four-neighbor chain code (Alg. 2, cf. Figure 3b).

Define the start and end points of the curve with grid point coordinates (latitude, longitude) (X_s, Y_s) , (X_e, Y_e) . Place it so that it does not leave the canvas. Mark segments that are not drawn with pen up (pu).

Identify as canvas an area with a street network of the specified category that is as regular as possible with $(X + 1) \cdot (Y + 1)$ crossing points, whose GPS coordinates are given, and determine start and end point intersections. To this end, create two-name lists of all rectangular streets in the (X, Y) -rectangle.

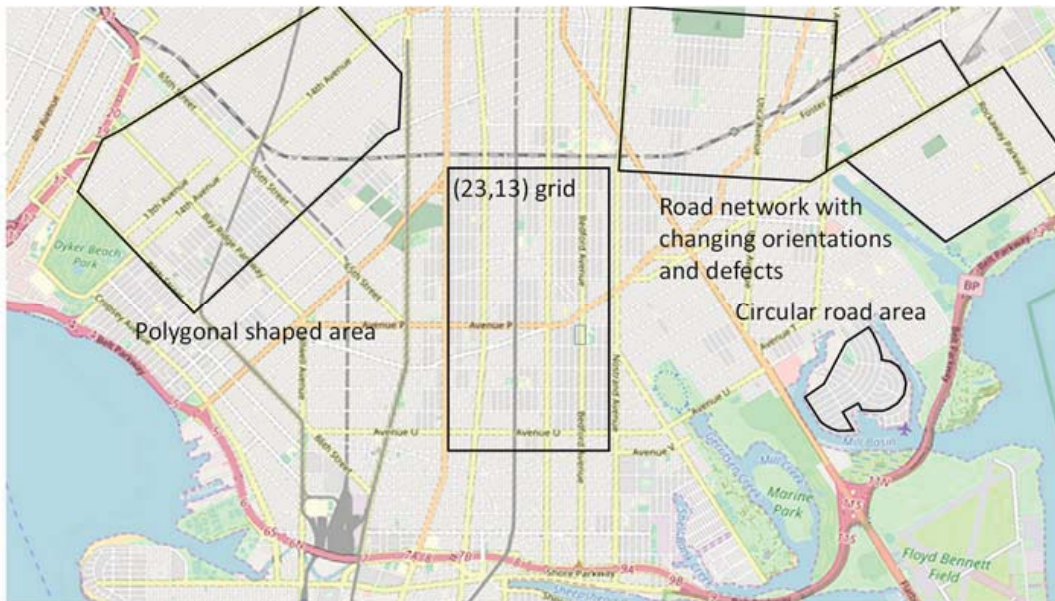


Fig. 4. Brooklyn NY after 8.5° rotation to the right: various road shaped networks

To get an array IP of all intersections of two streets situated in the rectangle in Google Maps, submit the request in the format: “[Street Name A] & [Street Name B], City]”. The API will return latitude and longitude coordinates of the intersection. Then add the four nearest neighbors in the direction 0,1,2,3 to each intersection in IP.



M 33233332111211123323332111111 pen up 222222	(Xs, Ys) = (40.745682,-73.972206) 36th & 1st NY Manhattan
E 00003332222000333222 pu 2(22)111111	(Xe,Ye) = (40.781412,-73.946191) 93th & 1st
R 333333222212101000022212121 pu 2	Midpoint in letter Y: 63th & Lexington Ave, GPS coordinates
Y 332332331101033011 pu 2222	of the neighbor intersections
O 3033332322212111101000 pu 2222	0: 40.764134 -73.966833 2: 40.765361 -73.965953
K 333333 pu 2222 10101002212121 pu 2	1: 40.764061 -73.964784 3: 40.765426, -73.967981

Fig. 5. Digitized text with four-chain code on rectangular (0,57)×(0,6) Manhattan canvas

Describe the artwork according to the affected meshes using the list of GPS intersection coordinates of the constructed path.

Visualize the digitized path on the canvas. The length of the path can be optimized if, in addition to the horizontal and vertical sections of the path, existing cross connections between the intersections (in the eight-neighbor topology) are used.

6 Software

GPSVisualizer [7] is a sketchbook of GPS artists and proposes a standardized input form that automatically draws your GPS data (or various exchange formats (KML/KMZ file, etc.) overlaid upon a variety of background maps and imagery, using either the Google Maps API or Leaflet, an open-source mapping library and freehand drawing utility that allows users to interactively draw on a map creating their own GPX or KML file. We used this software to produce figures 2 through 5.

ExpertGPS (<https://www.expertgps.com>) is a map software for planning outdoor trails. It shows waypoints and tracklogs on any handheld GPS receiver over aerial photos and topographic maps of the United States. It converts any GPS, GIS, or CAD data to or from GPX, Google Earth KML or KMZ, Excel CSV or TXT, SHP shapefiles or AutoCAD DXF drawings and allows users to measure distance, elevation and grade.

Google Maps (<https://maps.google.com>) can be used to build highly customizable and scalable maps with their own content and imagery and to import features from KML files, spreadsheets and other files (CVM, KML, KMZ, GPX, XLSX). It supports the user in creating complex applications and powerful visualizations of the data on a modern web platform with a comprehensive user interface (cf. Figure 6).

OpenStreetMap (<https://www.openstreetmap.org>) offers a free editable map of the whole world. The software is built by a community of volunteers that contribute and maintain data about roads, trails, small businesses and railway stations all over the world. Both can be categorized as Mapping API tools.

Geolocation APIs are mainly used to retrieve geolocation information in a device- and software-agnostic manner. The W3C has published a geolocation API <https://www.w3.org/TR/geolocation-API/> <https://github.com/w3c/geolocation-api>) as draft code that focuses on retrieving WGS84-compliant position information on hosting, especially for mobile devices. Recently, a geolocation sensor API draft has been published as a work in progress, <https://www.w3.org/TR/geolocation-sensor>. It extends the concepts of the aforementioned API in terms of consistency, security, privacy, and extensibility.

Google offers a geolocation API that can also use specified WiFi SSIDs, CellId, area code or mobile network codes to retrieve geolocation information (<https://developers.google.com/maps/documentation/geolocation/intro>). This API can be useful if satellite-based positioning is unavailable or is considered insufficient. Since multiple input parameters are accepted, the assumed location data delivered by the Google API is usually supplemented by an accuracy value. A complete software solution offers an interface to the following functionalities and existing toolboxes:

- 1) Find and extract areas with large road networks, streets and intersections, and search for capital cities on the five continents and road names, such as 63th Street.

- 2) Define canvas, dimensions and mesh size; convert the curves into polygons by such means as applying Chaikin's algorithm to the B-spline control points, (a few times) in order to get (more accurate) directional chain codes; compute the four-chain code of the polyline (<https://stackoverflow.com/questions/36680297/how-to-convert-polyline-or-polygon-into-chain-code>) and the list of GPS coordinates of adjacent intersections; identify the route on the canvas; optimize and visualize the route.

OpenCV proposes a data structure *polyline* with arrays of polygonal curves and vertex counters, the number of curves, a flag indicating whether the drawn polylines are closed or not, and their color and further line style elements.

3) Use GPS visualizer: Input can be done in the form of GPS data (tracks and waypoints), driving routes, street addresses, or simple coordinates.

7 Further Examples, Conclusions and Future Work

Our GPS data were collected in Santiago de Chile on Tuesday, September 3, 2013, between 17 and 18h using several smartphones during a walk in Parque O'Higgins, Tupper, Beauchef, Blanco Encalada, Jose Miguel Carrera, Domeyko, and Almirante Latorre. The walk shows the accuracy of the GPS coordinates (parameter linewidth) using ground truth. The color used represents type of environment—forest, trees, small and high buildings, free areas, streets and crossings (cf. Figure 6).

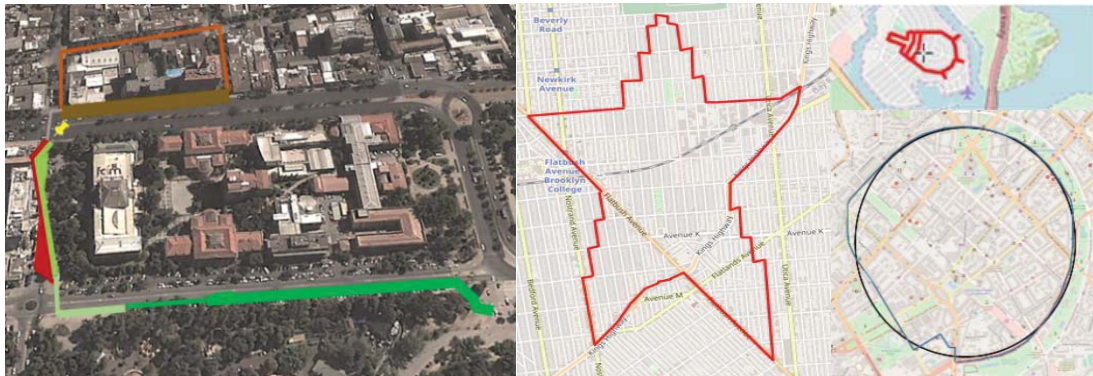


Fig. 6. Walk across the campus of University of Chile: Public domain under Creative Commons 4.0 licence—on the right: star, light bulb (length 3.4 km) in Brooklyn, circle in Yerevan

Rosner et al. [13] mention the challenge to paint a star and a circle which is quite simple with a regular street grid at hand (cf. Brooklyn). For further examples see <https://gpsdoodles.com/>. The grid-covering algorithm can also be applied to the city of Yerevan. A perfectly circular street surrounds the castle in the city of Karlsruhe.

This contribution provides important basics of digital geometry for GPS drawing and two basic algorithms based on polygonal coverage and chain codes. Illustrative examples show the difficulty in finding suitable road networks.

The authors Batty and Longley [4] apply fractal geometry to understand shapes of cities and construction principles developed by city planners and architects in contrast to the historic growth and expansion. They observe that a simple geometric pattern repeated on different scales and self-similarity play an important role. Strong geometric city layouts such as the octagonal Palmanova in Italy are worth mentioning, as are circular towns (e.g., Karlsruhe) or regular cell growth (e.g., Savannah), whereas Beijing has a fractal dimension close to two, but only a roughly regular grid. The observed fractal growth of large cities indicates that no regular street grids of any size can exist and that Euclidean objects cannot be approximated with equal accuracy in any size.

The efficient localization of suitable road networks using intelligent learning approaches, scaling and adaptation of the canvas and an evaluation of the user experience are priority topics for a relevant research question and planned for future work.

References

1. Babahajiani, P., Fan, L., Kämäräinen, J.-K., Gabbouj, M.: Urban 3D segmentation and modelling from street view images and LiDAR point clouds. *Machine Vision and Applications* 28, 679–694 (2017).
2. Balduz, P.: Walk line drawing. BSc Thesis. Faculty of Informatics at the TU Wien (2017).
3. Basu, S., Bonafilia, D. Gill, J., Kirsanov, D., Yang, D.: Mapping roads through deep learning and weakly supervised training. <https://ai.facebook.com/blog/mapping-roads-through-deep-learning-and-weakly-supervised-training/>
4. Batty, M., Longley, P.: *Fractal Cities: A Geometry of Form and Function*. Academic Press, San Diego, CA and London (1994) <http://www.fractalcities.org/>
5. Coeurjolly, D., Zerarga, L.: Supercover model, digital straight line recognition and curve reconstruction on the irregular isothetic grids. *Computers & Graphics* 30(1) 46–53 (2005).
6. Global positioning system standard – Positioning service. Performance standard: Integrity, Service, Excellence - ASD(NII)/DASD (C3, Space and Spectrum) 4th Ed., Sept. 2008.
7. GPS Visualizer <https://www.gpsvisualizer.com/draw/>
8. Janser, A., Luther, W., Otten, W.: *Computergraphik und Bildverarbeitung*, Vieweg 1996.
9. Karney, C.F.F.: Algorithms for geodesics. *J. Geod.* 87, 43–55 (2013).
10. Mattyus, G., Luo, W., and Urtasun, R.: Deep road mapper: Extracting road topology from aerial images. In *The IEEE Inter. Conf. on Computer Vision (ICCV)*, Oct 2017. *Machine Vision and Applications* 28, 679–694 (2017).
11. Pabian, F.: Commercial Satellite Imagery as an Evolving Open-Source Verification Technology: Emerging Trends and Their Impact for Nuclear Nonproliferation Analysis; EUR27687 (2015). <https://ec.europa.eu/jrc>
12. PedBikeInfo: http://www.pedbikeinfo.org/cms/downloads/PBIC_WhitePaper_Crowdsourcing.pdf
13. Rosner, D., K., Saegusa, H., Friedland, J., Chambliss, A.: Walking by Drawing. *Proceedings CHI 2015, Crossings, Seoul, Korea, April 18–23*, 397–406 (2015).
14. Schindler, A., Maier, G., Pangerl, S.: Exploiting Arc Splines for Digital Maps. *Proceedings 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)* 6p. CD-ROM (2011).
15. Schüller, D.: Visualisierung von Straßenzustandsdaten - Ein Nutzerinterface zu ihrer Erfassung und Bearbeitung. Diploma thesis in cooperation with Schniering Ingenieurgesellschaft Essen, University of Duisburg-Essen (2009).
16. Standard Positioning Service (SPS) (2008) <http://www.gps.gov/technical/ps/2008-SPS-performance-standard.pdf>; <https://www.gps.gov/systems/gps/performance/accuracy/> 2011
17. Strava: <https://www.strava.com/>
18. Waschke, A., Krüger, J.: Automatic route planning for GPS art generation. *Computational Visual Media* 5(3) 303–310 (2019).
19. Wei Xia, Yu-Ze Zhang, Jian Liu, Lun Luo and Ke Yang: Road Extraction from High Resolution Image with Deep Convolution Network—A Case Study of GF-2 Image. *MDPI Proceedings* 2, 325 (2018).
20. Zhongyi Ni, Lijun Xie, Tian Xie, Binhua Shi and Yao Zheng: Incremental Road Network Generation Based on Vehicle Trajectories. *International Journal of Geo-Information* 7(10) 382–400 (2018).
21. Zogg, J.-M.: *GPS Essentials of Satellite Navigation Compendium*. U-blox AG (2009).

A Batching Cloaking Scheme for Continuous Location-Based Services

Carlos Faúndez¹, Claudio Gutierrez-Soto^{1,2}, Patricio Galdames^{1,2}
and Pedro G. Campos¹

¹Universidad del Bío-Bío, Concepción 4051385, Chile
carlos.faundez1501@alumnos.ubiobio.cl, pgcampos@ubiobio.cl

²Group of Smart Industries and Complex Systems (gISCOM)
Universidad del Bío-Bío, Concepción 4051385, Chile
{pgaldames, cogutier}@ubiobio.cl

Abstract. Nowadays, the expanded use of LBSs involves opportunities to the adversaries threatening the location privacy of mobile users. Several approaches have been proposed to tackle either location privacy, location safety, and query privacy independently. In this paper, we present a work in progress, which aims to propose a unified framework to protect privacy in all these dimensions simultaneously. The demand for query-privacy protection for many users will be addressed in batch.

Keywords: location privacy, · location safety · query privacy · batch processing.

1 Introduction

Locations Based Services (LBS) are becoming popular due to the high demand for GPS-based services, which have been promoted by the usage of mobile devices and sensors. Geospatial applications can easily access through Google play or Apple store. By using these applications, a user can know which the least congested route is to reach a specific destination, if there is a hospital nearby; and even the user could receive notifications about events that are occurring in its proximity. Furthermore, many LBSs are being developed at present not only for commercial but also for scientific research purposes. This trend is a foreseeable future where our lives will be affected by computing environments, which store and process information not only about our position but also our lifestyle. This computerized future puts our privacy and security at risk.

Nowadays, the expanded use of LBSs involves opportunities to the adversaries threatening the location privacy of mobile users. Through the inference attacks carried out by adversaries [6], there is a chance to know the behavior patterns (i.e., taking into consideration the location and time) for a user. Aiming at tackling this challenge, significant research has been developed to provide users with location privacy and safety.

The current research in the context of location-based services (LBS) have been based primarily on the following three aspects (i.e., these aspects are also techniques): Location Privacy [2, 9, 5, 8, 14, 13, 16], Location Safety [10, 11, 15] and Query Privacy [17, 18, 12]. Location Privacy has as a goal to hide the location of a user to the adversary (i.e., LBS) over a specific region. To achieve this goal, a cloaking region is built for the user considering at least $k-1$ other users who have been in the same area (note that some users could be in the same space at the same time). On the other hand, according to [10], Location Safety is defined as “the problem of preventing an adversary from location (and thus destroying) nodes based on their location information revealed explicitly in communications.” Even though the aspects mentioned above prevent the LBS from knowing a user’s whereabouts, those do not avoid that the LBS knows what query was submitted by a user. Finally, Query Privacy (ℓ -diversity) attempts that the adversary cannot identify the user’s real query from others $\ell-1$ different queries.

Nevertheless, the existing solutions for the three above mentioned aspects have significant limitations. First, these works ignore the relationship between location privacy, location safety, and query privacy. Existing works mainly focus on one of these aspects independently. Second, these techniques also build up cloaking regions for each user individually and even more so when privacy protection is not relevant for the current location of the user. All these features turn the anonymity system into a bottleneck, especially in the case of continuous LBS services. Third, these works also do not consider a proper balance between location privacy/safety protection and processing cost at the LBS. Fourth, those solutions aimed to protect query privacy assume that the types of queries to provide ℓ -diversity are known beforehand, and they remain constant over time, which is not generally valid.

In this work, we propose to tackle the three above mentioned privacy aspects at the same time. To face the bottleneck issue, we will consider two features: batch processing and restricted space. Specifically, we propose a batching algorithm to build efficiently cloaking regions considering multiple mobile users. These users have different privacy requirements (i.e., location privacy, location safety, and query privacy requirements are not the same for every user, and all are required at the same time). However, cloaking regions are not built for those users located within a restricted area [13]. Finally, to address the type-of query assumption, we plan to classify the geographic queries according to their semantic proximity. Additionally, we present strategies to limit the usage of ℓ -diversity.

The remainder of this paper is organized as follows: In Section 2, some basic definitions, the adversary model, along with the problem description, are provided. In Section 3, we describe some implementation details. Related works are presented in Section 4. Finally, in Section 5, we conclude this work in progress relating to how we plan to evaluate the performance of our proposed solutions.

2 Basic Concepts

To provide location privacy for a user, we offer the traditional scheme based on k -*anonymity* (i.e., the real user's location cannot be distinguished from other $k-1$ places submitted to the LBS [20]). On the other hand, the location safety implies "to identify an area whose safety level is below to some threshold (θ)," where safety level corresponds to "the ratio of its area and the number of nodes inside it" [10].

We assume that there are two types of adversaries, *passive* and *active*. A *passive adversary* is any user that can monitor and eavesdrop on the wireless traffic or compromise any other user to obtain its private data. An *active adversary* is any user that can compromise the LBS server. In this definition, we also consider that the LBS is itself a potential adversary. Besides, the adversaries can perform the following attacks: inference attack, colluding attack, and accessibility attack.

According to [13], a space S is restricted to a user when he submits a service request at some time t , and this action reveals his presence in S at time t . Based on this concept, these authors propose a technique called *Restricted Space Cloaking* to limit the number of cloaking regions to be built for a mobile user.

We define the geographical ℓ -*diversity* notion. This notion involves seeking popular different queries, which are related to specific spaces that have distinguishing characteristics, but these are in different locations. To achieve this goal, we will use concepts related to places, which will have associated ontologies. These ontologies will allow us to give a semantic meaning.

3 Our Proposed Scheme

We assume there exists a set of mobile users moving within a defined network area. They trust in an anonymity server, which receives the exact locations of the users along with their queries and their privacy requirements. This anonymizer should build in batch the cloaking regions (i.e., for each user) considering the need for location privacy, location safety, and query privacy. It is essential to point out that the ℓ -diversity will be mainly formed by queries of real users, decreasing the number of query dummies. Additionally, the anonymizer should update the cloaking regions each time a users' movement takes place. Once this step has been carried out, this information is submitted to the LBS to process the queries. Subsequently, the query answers are sent to the users. We plan to use *Restricted Space Cloaking* to limit the number of cloaking regions to be built, and we plan to propose a similar criterion to limit the usage of ℓ -diversity.

Finally, we hope to have enough empirical results to obtain conclusions about the three aspects involved in our approach. Several experimental environments will be evaluated using simulations.

4 Related Work

Several approaches deal with location cloaking techniques. In [7], the authors provide a method to improve location privacy in LBSs. This approach involves the symmetric encryption as well as the k-anonymity technique taking advantage at the same time caching mechanism. This caching mechanism is based on repetitive queries. Empirical results compare four schemes, considering the average computation time and communication cost per query on the anonymizer and the LBS server. In the context of Vehicle Sensing Systems (VSS), in [4], the researchers present an efficient location privacy-preserving range query scheme for secure VSS communications. This approach is based on a protocol to hide sensitive information.

On the other hand, works such as [3] expose a privacy-preserving kNN (nearest neighbor) query scheme in the context of road networks. This scheme utilizes several cryptographic primitives such as Paillier cryptosystem [21], condensed RSA digital signature along with Voronoi diagram. This scheme holds the privacy of spatial data and kNN queries and confirms the authenticity of each query result.

Finally, other works address the location safety issue. In [1], the authors present the notion of differential private k-anonymity (DPkA) for query privacy in LBS. This notion integrates the concepts of k-anonymity and differential privacy. The authors provide a scheme to achieve the 0-DPkA; when 0-DPkA is not reachable, they propose an algorithm to solve this issue. In [15], the efficient construction of location cloaking areas for many users is presented. The building of location cloaking areas is carried out in a batch, considering the privacy and safety requirements of multiple users at the same time. This work presents two batching techniques.

To sum up, at present, there is not any work that unifies the three mentioned aspects of location privacy, location safety, and query privacy. Therefore, we believe that our approach is novel.

5 Conclusions

In this paper, we have presented an approach based on a batching algorithm, which addresses three aspects: Location Privacy, Location Safety, and Query Privacy. It should be noted that these three aspects have not been tackled at the same time. In this way, we hope to be able to evaluate several metrics considering the attack of active and passive adversaries. To this end, a wide range of experiments will be carried out to obtain empirical results from this approach.

6 Acknowledgments

This work supported by the Universidad del Bío-Bío of Chile under grant DIUBB 184615 1/I and the Group of Smart Industries and Complex Systems (gISCOM) under grant DIUBB 195212 GI/EF.

References

1. Wang, J., Cai, Z., Li, Y., Yang, D., Li, J., Gao, H.: Protecting query privacy with differentially private k-anonymity in location-based services. *Personal Ubiquitous Comput.* 22(3), 453–469 (Jun 2018).
2. Hu, P., Wang, Y., Li, Q., Wang, Y., Li, Y., Zhao, R., Li, H.: Efficient location privacy-preserving range query scheme for vehicle sensing systems. *Journal of Systems Architecture*106, 101714 (2020).
3. Yang, S., Tang, S., Zhang, X.: Privacy-preserving k nearest neighbor query with authentication on road networks. *Journal of Parallel and Distributed Computing*134, 25–36 (2019).
4. Hu, P., Wang, Y., Li, Q., Wang, Y., Li, Y., Zhao, R., Li, H.: Efficient location privacy-preserving range query scheme for vehicle sensing systems. *Journal of Systems Architecture*106, 101714(2020).
5. Peddinti, S.T., Dsouza, A., Saxena, N.: Cover locations: availing location-based services without revealing the location. In: *The 11th Privacy Enhancing Technologies Symposium, PETS*, (2011).
6. Nosouhi, M.R., Pham, V.V.H., Yu, S., Xiang, Y., Warren, M.: A hybrid location privacy protection scheme in big data environment. In: *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 1–6 (Dec 2017).
7. Zhang, S., Choo, K.K.R., Liu, Q., Wang, G.: Enhancing privacy through uniform grid and caching in location-based services. *Future Generation Computer Systems* 86, 881–892 (2018).
8. Niu, B., Li, Q., Zhu, X., Cao, G., Li, H.: Achieving k-anonymity in privacy-aware location-based services. In: *Proceedings of the IEEE International Conference on Computer Communications, INFOCOM*, (2014).
9. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: *ACM MobiSys, Proceedings of the first international conference on Mobile systems, applications, and services*, 31–42 (2003).
10. Xu, T., Cai, Y.: Location cloaking for safety protection of ad hoc networks. In: *Proceedings of the IEEE International Conference on Computer Communications, INFOCOM*, (2009).
11. Xu, T., Cai, Y.: Location safety protection in ad hoc networks. In: *Ad Hoc Networks* 7(8), 1551-1562 (2009).
12. Niu, B., Zhu, X., Li, W., Li, H., Wang, Y., Lu, Z.: A Personalized Two-Tier Cloaking Scheme for Privacy-Aware Location-Based Services. In: *International Conference on Computing, Networking and Communications and Information Security, ICNC*, (2015).
13. Yang, G., Cai, Y.: Full Location Privacy Protection Through Restricted Space Cloaking. In: *Journal of Information Processing* 25, 756–765 (2017).

14. Niu, B., Gao, S., Li, F., Li, H., Lu, Z.: Protection of Location Privacy in Continuous LBSs against Adversaries with Background Information. In: The International Conference on Computing, Networking, and Communications and Information Security, ICNC, 1-6 (2016).
15. Galdames, P., Gutierrez-Soto, C., Curiel, A.: Batching Location Cloaking Techniques for Location Privacy and Safety Protection. In: Mobile Information Systems 2019 11 pages,(2019).
16. Tobar, G., Galdames, P., Gutierrez-Soto, C., Rodriguez-Moreno P.: A Batching Location Cloaking Algorithm for Location Privacy. In: Proceedings of the Workshop on Collaborative Technologies and Data Science in Smart City Applications, CODASSCA, 26-36 (2018).
17. Pingley, A., Zhang, X., Fu, X., Choi, H.-A., Subramanian, S., Zhao, W.: Protection of query privacy for continuous location-based services. In: Proceedings of IEEE International Conference on Computer Communications, INFOCOM (2011).
18. Niu, B., Zhu, X., Lei, X., Zhang, W., Li, H.: Eps: Encounter-based privacy-preserving scheme for location-based services. In: Proceedings of IEEE International Conference on Global Communications, GLOBECOM, (2013).
19. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkatasubramanian, M.: L-Diversity: privacy beyond k-anonymity. In: Proceedings of IEEE International Conference on Data Engineering, ICDE, (2006).
20. Sweeney, L.: K-Anonymity: a model for protecting privacy. In: International Journal of Uncertain Fuzziness Knowledge-Based Systems 10(5), 557-570 (2002).
21. Paillier, P.: Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, EUROCRYPT, Springer, 223–238 (1999).

Protecting query privacy through semantic caching in location-based services

Fernando Vera-Catricura¹, Patricio Galdames^{1,2},
Claudio Gutierrez-Soto^{1,2}, and Arturo Curiel³

¹ Universidad del Bío-Bío, Concepción 4051385, Chile
fernando.vera1501@alumnos.ubiobio.cl

² Group of Smart Industries and Complex Systems (gISCOM)
Universidad del Bío-Bío, Concepción 4051385, Chile
{pgaldames,cogut.er}@ubiobio.cl

³ Universidad Veracruzana, Veracruz, México
acuriel@conacyt.ms

Abstract. We plan to address the problem of processing Location-Based Queries (LBQ) in a MANET to preserve query privacy as much as possible. Our idea is that mobile users will first ask themselves to solve a query before any user decide to submit its query to an untrusted LBS. Our first goal is to define a collaborative caching strategy to be run by the mobile users themselves that exploit the geographic and semantic similarities among the LBQs to perform efficient processing of LCQs. Our second goal to protect a user's query privacy when any user submits an LBQ to the LBS; we are planning to develop a distributed algorithm to provide l -diversity only when this protection is useful. Existing caching techniques for MANET do not exploit semantic and geographic similarities among the LBQs, and they assume mobile users have access to some storage infrastructure to maintain global information to compute l -diversity.

Keywords: Query Privacy · Semantic Cache · Query Similarity · MANET.

1 Introduction

Location-Based Services (LBS) are becoming popular due to the significant development of social networks and smartphones. Users can easily access through Google Player or Apple Store, services that allow them either to know the path with the least congestion to reach a specific destination, if there is a hospital in their surroundings, or to receive an alert when a family member is nearby. All these examples are called *Location-Based Queries* (LBQ) because a user demands a service from a provider (LBS) whose answer depends on the exact location of this user. However, the periodic release of our positions and our queries to the LBS can lead this server to conclude more details about our lifestyle and identity. If the LBS acts unethically or illegally, it may sell our information to others without our consent so that our privacy may be put in danger. Therefore, substantial research has been developed to provide users with techniques that protect their privacy.

Most research articles assume that users want to protect either location privacy or/and query privacy. One of the most promising ideas to protect location privacy in the context of LBS applies the concept of *k-anonymity*. This technique aims to build for every user. This cloaking area includes the user's exact location and also other chosen places where this user can also be with high probability [1, 3, 5, 4]. Other techniques aim to protect a user's query privacy [9, 6, 7, 11, 10, 2, 13] by building a set of different ℓ -1 LBQs, called *ℓ -diversity*. These queries are also released from the user's whereabouts. In this way, our adversary, the LBS, cannot distinguish which query from the ℓ possible alternatives is the real one.

In this article, our goal is to show what the challenges are to perform efficient processing of LBQs when query privacy is the primary concern. To tackle this problem, only at the LBS is challenging since a user's query has been replaced by ℓ queries. Therefore, the LBS must process many queries just to answer only one and may become a bottleneck when many users request service. Moreover, the LBS wastes unnecessarily limited computing resources since only one query is the real one. To face these issues, some researchers have proposed the use of a trusted third party called the *anonymizer*. This server can receive many LBQs and can cluster them by their proximity and, in this way, to build an *ℓ -diversity* query set from real queries. However, since the anonymizer works as a proxy, it can both become a bottleneck and attractive target to be compromised. To deal with these latter issues, other researchers [9, 11, 10, 12] have proposed a decentralized infrastructure, i.e., a MANET (a Mobile Adhoc Network) to compute proper query privacy protection and also to process LBQs.

However, these decentralized solutions show essential limitations. First, they consider that any information required to obtain *ℓ -diversity* for a given query is globally known by a trusted third party, which is not always accurate in ad-hoc networks. Second, they also assume that this information remains constant over time, which is not generally valid. Third, a few solutions suppose that users maintain a local cache, and they use it to solve LBQ locally or asking their neighbors before submitting any LBQ to the LBS. These caching solutions only work with exact semantic queries. Finally, these techniques intend to continuously provide an *ℓ -diversity* set to every user protecting its query privacy. However, there exist scenarios in which we think that it is not necessary, or it may become useless. Consider a massive public event, like a music concert, where it is expected that many users will be submitting LBQs related to this event. Since this is a public spectacle, we can consider these LBQs are general knowledge.

In this work in progress, we assume that users can communicate by themselves throughout an ad-hoc network, and they are also able to transmit throughout a cellular network infrastructure to reach an untrusted LBS. Based on these assumptions, we aim first to develop a fully distributed scheme to provide *ℓ -diversity* and to be run only by the users themselves. Second, to protect a user's query privacy and LBS resources, we aim to propose a collaborative caching scheme supported by mobile users themselves [16].

Our idea is that the users must first ask their caches before asking the LBS, and the management of this cache must take advantage of the semantic and geographic similarity among the LBQs. Third, to mitigate the possibility the LBS becomes a bottleneck, we consider proposing some criteria to decide when ℓ -diversity is computed or not to protect query privacy.

The rest of this paper is organized as follows. In Section 2, we offer some basic definitions and the adversary model. We also describe some implementation challenges and give some solution ideas. Finally, in Section 3, we conclude this work in progress relating to how we plan to evaluate the performance of our proposed solutions.

2 Our proposed scheme

2.1 Basic concepts

To provide query privacy for a user, we aim to offer ℓ -diversity, i.e., the real user's query cannot be distinguished from other selected $\ell-1$ queries [8]. To compute ℓ -diversity, users must keep updated a *query frequency table* (QFT), which registers the number of query types submitted from a specific location. We plan to propose a semantic similarity metric that allows us to find examples of popular and distinguishing queries in a particular context. We assume that these types of queries can be updated periodically from a provider, and these can be used to compute a QFT.

We assume that there are two types of adversaries. A *passive adversary* is any user that can monitor and eavesdrop on the wireless traffic or compromise any other user to obtain its private data. An *active adversary* is any user that can compromise the LBS server. In this definition, we consider the LBS itself as a potential adversary as well. Besides, we think these adversaries can perform the following attacks: *inference attacks* and *accessibility attacks*.

2.2 System overview

We split the network domain into a set of disjointed grid cells (or subdomains). The size of each cell is set to $r/\sqrt{2} \times r/\sqrt{2}$, where r is a node transmission radius. Thus, when a mobile user broadcasts a message, it covers the entire cell where this user resides. The network partitioning is made known to all mobile users, and each mobile user caches the queries that are relevant to its *home cell*. We assume each user can cache some data locally on their mobile phone or laptop. We plan to develop a caching strategy to decide whether or not a query (and its answer) are stored in this cache.

We say a cell is a node's home cell if the node is currently inside the cell, and each cell has its QFT. Since users cannot rely on a central server or storage point, users moving within their home cells must maintain their corresponding QFTs. When a user walks into a new cell, it sends a 1-hop broadcast to obtain the cell's QFT, and this new cell becomes its home cell. Any user moving within its home cell can broadcast back the cell's QFT. When nobody is within this new cell (every broadcast has a timeout), we are planning to follow a similar approach as proposed by [15]

When a node leaves its current cell, it sends a 1-hop broadcast message asking if it needs to retain this QFT until someone requests it. When a user needs to create an LBQ, first, it checks whether it finds a valid query answer in its local cache. In case it does not have an answer or data has become stale, then it updates its QFT and submits its query to their neighbors located within its home cell. Then, every user in this cell updates its cell's QFT. If any neighbor cannot answer the LBQ, the user computes ℓ -diversity by collecting several QFTs from several geo-casting [14] submitted to nearby cells. Then the user submits its LBQ protected with ℓ -diversity to the untrusted LBS.

3 Conclusions

In this paper, we plan to develop a distributed algorithm to process Location-Based Queries (LBQ) in a MANET. Our idea is to exploit the geographic and semantic similarities among the LBQs and build a distributed cache supported only by the mobile users themselves. If a user does not solve its query in the MANET, then it sends its query to the untrusted LBS. To protect a user's query privacy, we are planning to develop a distributed algorithm to provide ℓ -diversity. Thus, the user computes $\ell-1$ different dummy queries that are also submitted with the real one to the untrusted LBS. However, we think there are application scenarios where ℓ diversity is useless, and we are planning to work on some criteria to detect those scenarios. In this way, we plan to prevent overloading an LBS with unnecessary dummy queries.

It should be noted that existing caching techniques do not consider semantic and geographic similarities, and they also assume mobile users have a storage infrastructure to maintain global information (like the entire QFT). After we set solutions considering these previous aspects, we plan to evaluate several metrics considering the attack of active and passive adversaries. To this end, a wide range of experiments will be carried out to obtain empirical results from this approach.

Acknowledgments

This work supported by the Universidad del Bío-Bío of Chile under grant DIUBB 184615 1/I and the Group of Smart Industries and Complex Systems (gISCOM) under grant DIUBB 195212 GI/EF.

References

1. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: ACM MobiSys, Proceedings of the first international conference on Mobile systems, applications and services, 31–42 (2003).
2. Niu, B., Zhu, X., Li, W., Li, H., Wang, Y., Lu, Z.: A Personalized Two-Tier Cloaking Scheme for Privacy-Aware Location-Based Services. In: International Conference on Computing, Networking, and Communications and Information Security, ICNC, (2015).
3. Niu, B., Gao, S., Li, F., Li H., Lu, Z.: Protection of Location Privacy in Continuous LBSs against Adversaries with Background Information. In: The International Conference on Computing, Networking, and Communications and Information Security, ICNC, (2016).
4. Galdames, P., Gutierrez-Soto, C., Curiel, A.: Batching Location Cloaking Techniques for Location Privacy and Safety Protection. In: Mobile Information Systems, Vol. 2019, 11 pages, (2019).
5. Tobar, G., Galdames, P., Gutierrez-Soto, C., Rodriguez-Moreno, P.: A Batching Location Cloaking Algorithm for Location Privacy. In: Proceedings of the Workshop on Collaborative Technologies and Data Science in Smart City Applications, CODASSCA, pp. 26-36, (2018).
6. Pingley, A., Zhang, X., Fu, X., Choi, H.-A., Subramanian, S., Zhao, W.: Protection of query privacy for continuous location-based services. In: Proceedings of IEEE International Conference on Computer Communications, INFOCOM, (2011).
7. Niu, B., Zhu, X., Lei, X., Zhang, W., Li, H.: Eps: Encounter-based privacy-preserving scheme for location-based services. In: Proceedings of IEEE International Conference on Global Communications, GLOBECOM, (2013).
8. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramanian, M.: L-Diversity: privacy beyond k-anonymity. In: Proceedings of IEEE International Conference on Data Engineering, ICDE, (2006).
9. Amini, S., Lindqvist, J., Hong, J., Lin, J., Toch, E., Sadeh, N.: Cache: Caching location-enhanced content to improve user privacy. In: Proceedings. of ACM MobiSys (2011).
10. Shokri, R., Theodorakopoulos, G., Papadimitratos, P., Kazemi, E., Hubaux, J.-P.: Hiding in the mobile crowd: Location privacy through collaboration. In: IEEE Transaction on Dependable and Secure Computing, vol. 11, no. 3, pp. 266–279, May (2014).
11. Zhu, X., Chi, H., Niu, B., Zhang, W., Li, Z., Li, H.: Mobicache: When k-anonymity meets cache. In: Proceedings of IEEE GLOBECOM (2013).
12. Niu, B., Li, Q., Zhu, X., Cao, G., Li, H.: Enhancing Privacy through Caching in Location-Based Services. In: Proceedings of IEEE INFOCOM (2015).
13. Chen, Y., Wei, Y.: Location Privacy Protection Scheme Based on Location Services. In: Proceedings of the 2019 the 9th International Conference on Communication and Network Security (ICCNS), pp. 30–33, November (2019).
14. Navas, J. C., Imielinski, T.: GeoCast – Geographic Addressing and Routing. In: Proceedings of the 4th Annual Int'l Conf. on Mobile Computing and Networking (MOBICOM), Budapest, Hungary, pp. 66–76, (1997).
15. Galdames, P., Kim, K., Cai, Y.: A Generic Platform for Efficient Processing of Spatial Monitoring Queries in Mobile Peer-to-Peer Networks. In: Proceedings of the Eleventh International Conference on Mobile Data Management MDM, (2010).
16. Hu, P., Wang, Y., Li, Q., Wang, Y., Li, Y., Zhao, R., Li, H.: Efficient location privacy-preserving range query scheme for vehicle sensing systems. *Journal of Systems Architecture* 106, 101714 (2020).

Automatic image classification supported by expert knowledge

Sergio Peñafiel¹, Belisario Panay¹, Nelson Baloian¹, José A. Pino¹, and Wolfram Luther²

¹ Department of Computer Science, Universidad de Chile, Santiago, Chile
{spanafie, bpanay, nbaloian, jpino}@dcc.uchile.cl

² Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Germany
wolfram.luther@uni-due.de

Keywords: Image classification · Expert systems · Dempster-Shafer Theory

Extended Abstract

In recent years, several algorithms have been proposed to solve image classification problems mainly based on convolutional neural networks (CNN), which have reached high levels of accuracy. For example, Resnet [2] and LeNet [3] are models for object detection that have achieved accuracy over 98% on the Imagenet dataset [1].

However, many of these techniques have drawbacks. The black-box behavior of these networks is one of the most important weakness. This refers to the fact that if we train a very accurate neural network for a certain problem using a particular dataset, the user cannot know how the model decides. Thus it is impossible to ensure that networks learned correctly how to classify or they are just over-fitted to the dataset samples. Several articles illustrate these problems, for example Moosavi-Dezfooli et al. [4] show that these networks are unstable to adversarial perturbations. These perturbations are noise to the original image which might be imperceptible for humans, but cause the model to change their classification. Su et al. [6] present another example of these perturbations, in this case instead of adding a noise to the whole image, they just change one pixel of the image, making the model to misclassify losing over 16% of accuracy in Imagenet.

Many of the above problems of CNNs appear because we let these networks to change their thousands of parameters without restrictions. These unrestricted optimizations may lead into inexplicable models and require many data samples to converge. Our hypothesis is that we can develop a model with both fixed restrictions and “learnable” parameters. The constraints of this intermediate model should encode meaningful aspects to consider in the input when performing a prediction/classification. Also, it allows experts to include their knowledge about the classification problem. Finally, having fewer parameters makes the model require fewer data for training.

In this work we propose a two-phase method to perform expert assisted image classification. In the first phase, we will introduce the constraints of the problem to generate a meaningful feature vector for each image. In the second phase, we will perform interpretable classification using the same techniques for optimization as the neural network approaches such as gradient descent and backpropagation.

Peñafiel et al. proposed a new model for tabular classification based on the Dempster-Shafer Theory (DST) and Gradient Descent, we call this model DSGD [5]. The proposed method is rule-based, a rule is defined as a statement that can be verified with the data, and a mass assignment function (MAF) which is the knowledge encode element of DST. Rules can be defined either by expert knowledge or automatically based on data statistics. Then, the model can find the optimum values for the MAFs that produce the lowest error in prediction using Gradient Descent as optimizer. The main advantages of this method are: interpretability, it allows to include expert knowledge and it can handle missing information.

This method was tested on several datasets and it shows that the model is able to reach comparable accuracy to traditional classification methods like Support Vector Machines or K-Nearest Neighbors, without losing the ability to explain the decision it performs.

This proposed model is limited to work with tabular data, the goal of this work is to present alternatives about extending it to handle image classification. To do this, a straightforward strategy is to convert the image into a meaningful feature vector, and use this vector as the input for a tabular classification problem. Feature vectors should encode information about properties of the image. In our case, since we are interested in expert knowledge, we can define the vector values as how certain hypotheses about the image are.

Depending on the problem, the definition of these feature vectors variate. For example, for MNIST dataset [7] which have images of handwritten digits, we are interested in classifying according to the shape of the drawings. Then in the first phase, we can force the model to apply convolutional kernels or gradient histograms to the images to return a value that shows whether a certain shape is presented in the image. Figure 1 shows an example of the classification for this problem.

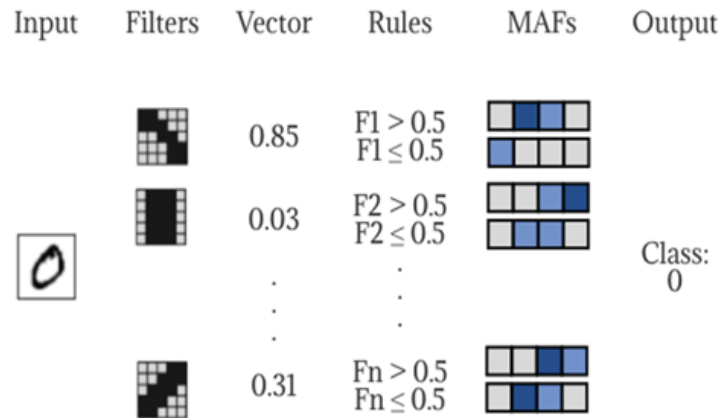


Figure 1 Example of digit classification using our proposed method. The model applies custom interpretable filters to the image to build a meaningful feature vector. These vectors are converted into rules and then we use them in DSGD model which assigns different MAFs to these rules. The MAF values are optimized using Gradient Descent to provide the best accuracy in prediction.

Another more complex example is the classification of scenes. In this problem, we need to determine which scene is shown in an image. Examples of scenes are bedrooms, parks, cities, etc. Here, we can define the expert knowledge as looking if certain items appear in the image. For example, in a bedroom is more likely to contain a bed, bedside,

furniture, etc.; in a park it is more likely to find trees, vegetation, and benches. Then we can develop an algorithm to detect those items in the image and create our feature vector according to the probability of these items come out in the image.

In order to evaluate the performance of the model, we can use standard classification metrics such as accuracy, precision and recall. We also propose an experiment to verify expert support classification requires fewer data to train compared to a traditional convolutional neural network and then achieving better accuracy when data is limited. We propose to compare the accuracy obtained by our model and a traditional convolutional neural network, but changing the number of samples provided for training. We hypothesize that our model achieves better performance when data is limited.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large- scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient- based learning. In: Shape, contour and grouping in computer vision, pp. 319–345. Springer (1999)
4. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
5. Peñafiel, S., Baloian, N., Sanson, H., Pino, J.A.: Applying Dempster-Shafer theory for developing a flexible, accurate and interpretable classifier. *Expert Systems with Applications* p. 113262 (2020)
6. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 23(5), 828–841 (2019)
7. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics* 22(3), 418–435 (1992)

Small is Beautiful – Temporal Accelerators for Embedded FPGAs

Christopher Cichiwskyj and Gregor Schiele

University of Duisburg-Essen, Bismarckstr. 90, 47057 Duisburg, Germany
{christopher.cichiwskyj,gregor.schiele}@uni-due.de

Abstract. When the complexity of a problem rises, its solution needs more hardware resources. A usual way to solve this is to use larger processors and add more memory. When using Field Programmable Gate-Arrays (FPGAs), which can instantiate arbitrary circuit designs, a larger, more costly and power hungry chip is used. In this extended abstract we propose a different approach, namely to split the problem into a graph of interdependent smaller tasks and to reconfigure a small FPGA during runtime to execute each of these tasks efficiently sequentially. This can result in cheaper and more energy efficient systems that can execute very complex problems locally.

Keywords: IoT · Embedded · FPGA · Reconfigurable Hardware

1 Introduction

The Internet of Things (IoT) consists of billions of cheap, low power devices that are embedded in everyday objects. They have very limited processing power and can execute only basic tasks. More complex tasks are usually offloaded to cloud services. This can lead to high latency as well as privacy and reliability risks. To mitigate this, researchers are looking into ways to make embedded IoT devices more powerful, allowing them to execute complex tasks locally.

To this end, recent years have seen a trend to augment IoT devices with embedded Field Programmable Gate Arrays (FPGA) [3, 5], that allow to instantiate arbitrary hardware circuits at runtime. FPGAs can be used to execute tailor-made accelerators to efficiently perform complex calculations “in hardware”, while having the capability to change the circuit when required. One example for such a platform is the Elastic Node [1, 2, 4]. It combines a low-power 8-bit Microchip AVR AT90USB1287 micro-controller unit (MCU), with a Xilinx Spartan 7 XC7S50 FPGA to create a low cost, low power asymmetric multiprocessor system.

An embedded FPGA contains a comparatively small amount of resources for circuit instantiation. This limits the size of accelerators and thus restricts the kind of application that can be supported. If a more complex accelerator is needed, the obvious solution is to use a larger FPGA with more resources. However, this increases the system cost as well as the power consumption. In this extended abstract, we propose to continue using a small FPGA and to take

advantage of the reconfiguration capabilities of the FPGA. Instead of designing an accelerator as a single, monolithic entity, we propose to divide the accelerator into smaller, modular parts. The FPGA then executes the parts sequentially by reconfiguring itself for each part, passing intermediate results between the parts. We call this concept a *Temporal Accelerator*.

2 Temporal Accelerators

Supporting Temporal Accelerators requires a number of steps, as shown in Figure 1. We discuss these steps in more detail in the following.

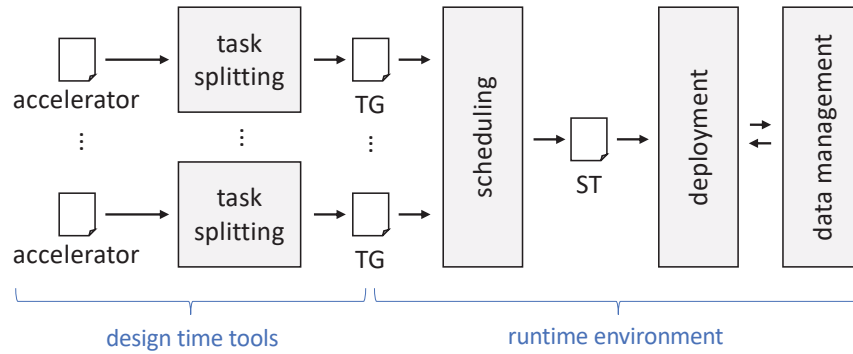


Fig. 1. System overview

Design The first step is to design a Temporal Accelerator. This can be done by a developer or by a task splitting algorithm based on an existing monolithic accelerator. A Temporal Accelerator consists of a set of interdependent *subtasks* (STs) that are arranged in a *Task Graph* (TG). Each ST realises a function without side effects that consumes input data and produces output data. STs can be calculated independently and are implemented as self-contained bit files that can be instantiated on an FPGA individually. Edges in the TG represent data dependencies between STs such that if a ST produces output data that is consumed by another ST, both STs are connected with an edge in the TG.

Scheduling At runtime, multiple TGs are handed to a scheduling algorithm that determines the next ST to be executed on the FPGA. Our current implementation uses a naive, priority-aware round robin scheduler that takes into account data dependencies in a TG. It is able to handle multiple concurrent TGs to support dynamic application scenarios with no knowledge about future incoming events. As an example, while a TG is executed, the IoT device may receive some important sensor measurement, signalling a critical change in its physical context. To react immediately, the system has to start processing this data with another temporal accelerator right away before finishing the old one.

Deployment Once the scheduler has chosen the next ST, the deployment manager loads the corresponding bit file and reconfigures the FPGA accordingly. Note that switching between different ST bit files to execute a TG may introduce a lot of FPGA reconfigurations. This can induce a large overhead with respect to performance and energy consumption. Reducing the number of required reconfigurations using more advanced scheduling algorithms is an active research topic. If no further ST has to be executed, the FPGA is deactivated. The deployment manager also coordinates with the data management to ensure that the ST receives the correct input data and that outputs are stored and forwarded correctly.

Data Management The result of a ST is required by its successors in the TG as their input data. Depending on the shape of the TG, a successor is not necessarily the next ST in the execution order, leading to the requirement to buffer intermediate results. Ideally, this would be done directly on the FPGA, such that the next instantiated ST can access the data directly. However, SRAM-based FPGAs are unable to keep any state during reconfiguration as the contents of their block memory, registers and flip-flops are lost. Therefore, we must buffer results outside of the FPGA. Partial reconfiguration could potentially solve this, but current embedded FPGAs don't support this feature. Due to this, we require a data management strategy that can handle intermediate results efficiently outside of the FPGA across multiple ST executions. This is another topic of active research.

3 Energy efficiency of a Temporal Accelerator

Intuitively the introduction of several reconfigurations should add a significant energy overhead to the accelerator execution, making it potentially unviable. However as can be seen in Fig. 2 the energy cost to reconfigure larger FPGAs surpasses the cost of smaller FPGAs significantly, e.g. the XC7S100 uses ~ 7 times more energy using the same settings than a XC7S25.

Using even larger FPGAs increases this gap, due to the longer time required to read in larger bit files. In combination with the much lower static power usage of smaller FPGAs, this could lead to situations, in which a temporal accelerator may be as energy efficient as or even more efficient than a normal accelerator. However this does not yet include the overhead introduced by the data exchange. This is currently being investigated

4 Conclusion

Temporal Accelerators are an opportunity to design high performing yet cheap embedded IoT devices. Intuitively, they can reduce the cost of devices, since a smaller, cheaper FPGA can be used. On the other hand it introduces overhead

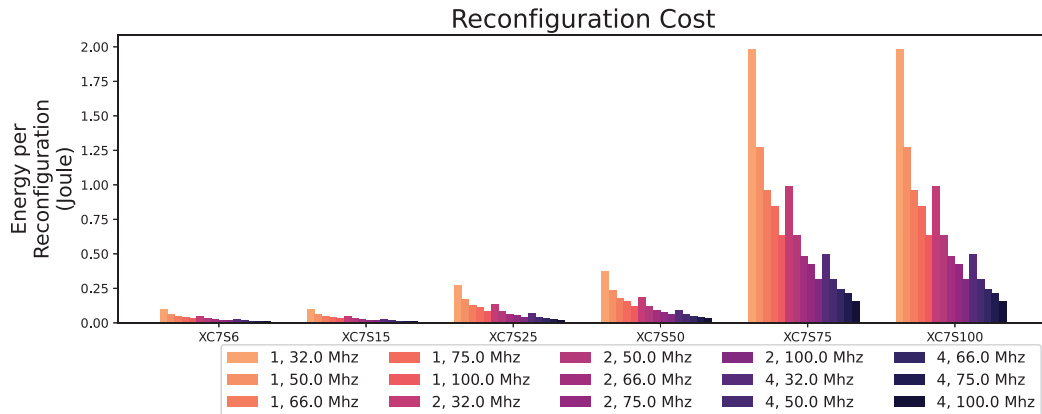


Fig. 2. Cost per reconfiguration for chips of the Xilinx Series 7 Spartan family

into execution, e.g. forwarding intermediate data and performing several reconfigurations, thus increasing the runtime. However, we expect in certain scenarios that using Temporal Accelerators on smaller chips can be more energy efficient than deploying larger chips. We plan to examine this in more detail in future work. We are also currently developing new scheduling algorithms to reduce the number of reconfigurations as well as new memory management algorithms for a more efficient forwarding of intermediate data between STs.

References

1. Burger, A., Cichiwskyj, C., Schiele, G.: Elastic Nodes for the Internet of Things: A Middleware-Based Approach. In: Proceedings - 2017 IEEE International Conference on Autonomic Computing, ICAC 2017. pp. 73–74. IEEE (jul 2017). <https://doi.org/10.1109/ICAC.2017.27>
2. Burger, A., Schiele, G.: Demo Abstract: Deep Learning on an Elastic Node for the Internet of Things. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 424–426. IEEE (mar 2018). <https://doi.org/10.1109/PERCOMW.2018.8480160>
3. Intel Corporation: Internet of Things - Accelerating the IoT with Intel FPGAs and SoCs (2019), <https://www.intel.de/content/www/de/de/internet-of-things/products/programmable/overview.html>
4. Schiele, G., Burger, A., Cichiwskyj, C.: The elastic node: An experimentation platform for hardware accelerator research in the internet of things. In: 2019 IEEE International Conference on Autonomic Computing (ICAC) (2019)
5. Soliman, S., Jaela, M.A., Abotaleb, A.M., Hassan, Y., Abdelghany, M.A., Abdel-Hamid, A.T., Salama, K.N., Mostafa, H.: FPGA implementation of dynamically reconfigurable iot security module using algorithm hopping. Integration - the VLSI Journal **68**, 108–121 (Sep 2019)

Online Collaborative Refinement to Increase the Quality of Students' Posed Questions

Ari Nugraha¹[0000-0002-5793-3157], Izhar Almizan Wahono¹, and Tomoo Inoue¹

¹ University of Tsukuba, Tsukuba, Japan
ari.nugraha@slis.tsukuba.ac.jp,
s1826099@s.tsukuba.ac.jp, inoue@slis.tsukuba.ac.jp

Abstract. Collaborative learning enables students to develop higher-order thinking skills and achieve richer knowledge generation. In this study we combined online collaborative learning with the student question posing activity to increase students' posed question quality guided by a question quality level taxonomy. Utilizing off-the-shelf online tools from the Internet, we proposed an online learning method for students to collaboratively pose question and refine their questions based on the learning material they have watched before.

Keywords: question posing; question generation; question refinement; question improvement; high quality question; collaborative question refinement.

1 Introduction

In current world situation where everything is become interconnected with the Internet, many new possibilities of learning methods exists to enable students learn everywhere and at any time, even from their home. Online collaborative learning in synchronous or asynchronous becoming new normal, either it is conducted in full online setting or blended setting. One of the effective methods for learning is question posing. By posing questions, students are actively enhance their understanding and comprehension when they construct relations between their prior knowledge and the learning material. However previous study showed that the questions posed by student were in low-level quality[1] which are sign of low cognitive level. Combined the flexibility of online learning, collaborative learning and question posing and refinement, in this study we proposed an online synchronous learning method to increase the quality of student generated questions to higher level based on the question level taxonomy.

2 Related Works

2.1 Online Collaborative Learning

In collaborative learning, learners are sharing and transmitting knowledge amongst them as they work towards common learning goals. Different from traditional one-way learning method from teacher to student, learners in collaborative learning are active in their process of knowledge acquisition as they exchange ideas or opinions in discussions, and information seeking with their peers. Knowledge acquired from these processes is shared among peers, not owned by one learner[2]. The Internet has opened another possibility to expand collaborative learning space outside of classroom by enabling electronic distance learning. Computer and the Internet play an important role in mediating interaction among participants in the process of meaning making through a joint activity as stated by Koschmann in his definition of Computer Supported Collaborative Learning (CSCL)[3]. The high usage of social media nowadays also impacts the way of online collaboration conducted, as study showed positive effect of collaborative learning through social media[4].

2.2 Question Generation and Quality Refinement

Posing or generating questions by students promotes a higher level of thinking to students as they tried to pose questions in which the answer can be found in the learning material[5]. Several studies on student-generated questions exist. One of the studies is PeerWise[6], a tool that allows students to create multiple-choice questions (MCQs) and answer those created by their peers. Using this tool, students were asked to focus on the learning outcomes by creating questions that align with these outcomes. By creating questions, the students could improve their understanding by writing an explanation of the answer to their question. When posing questions, students allocated their attention and cognition to generate meaning by finding any important information in the learning material and connecting the information with their knowledge (active processing), which results in increased comprehension[7].

Related to the collaborative learning, previous study comparing between guided peer-questioning groups and discussion groups showed that students questioning asked more critical thinking questions, gave more explanations, and demonstrated higher achievement than students from discussion groups which in the end enhanced the interaction between peer in the classroom[8].

One problem related to the question posing activity by the students is that the questions posed by student were in low-level quality[1] which are sign of low cognitive level. To increase the quality of student's posed questions, refinement activity can be utilized. Several studies exist related to the question refinement. Yu et al.[9] stated that by scaffolding in form of comments or feedbacks were perceived as not only helping the question-authors to refine their work, but also helped students to detect and correct incomplete ideas or misconceptions and improve their questions. Another method for refinement by Yeckehzaare et al.[10] are using students collaborative strategy in which the students can modify each other's questions and claim ownership

of the modified question by adding justification to why the previous question needed to be refined.

2.3 Question Quality Taxonomy

Table 1. Question Rubric Level for Measuring Student Generated Question

Level	Name
1	<i>Factual Information.</i> In this level, questions are simple in form and request a simple answer, such as a single fact. Example of this level of question are: What is the name of the United States Capital? What is the largest lake in the world?
2	<i>Simple Description.</i> In this level, questions are request general information that de-notes a link between concepts. The question can be simple, yet the answer may contain multiple facts and generalizations.
3	<i>Complex Explanation.</i> In this level, Questions request for an elaborated explanation about a specific aspect of concept with accompanying evidence.
4	<i>Pattern of Relationships.</i> In this level, questions display science knowledge coherently expressed to probe the interrelationship of concepts, they are a request for principled understanding with evidence for complex interactions among multiple concepts and possibly across concepts.

Several taxonomies have been constructed to categorize the quality of student generated questions. Question taxonomy from Gallagher & Aschner[11] divided question into four types : 1) **Memory questions**, focus on identifying, naming, defining, designating, and responding with yes or no; 2) **Convergent thinking questions**, focus on explaining, stating relationships, comparing, and contrasting. 3) **Divergent thinking questions**, focus on predicting, hypothesizing, inferring, and reconstructing. and 4) **Evaluative thinking questions**, focus on valuing, defending, judging, and justifying choices. The Question Rubric Level from Guthrie and Taboada[12] comprises of four levels of question quality as can be seen on **Table 1**.

Another well-known taxonomy of question is the Revised Bloom's Taxonomy[2] which divided into two parts, Low Order Thinking and High Order Thinking. In this study our proposed method uses the Question Rubric Level. Compared to the Bloom's taxonomy which previous study showed that some students had difficulty to work with[5], and the original authors of this taxonomy developed it as a rubric to describe students' growth and to guide instruction for making high quality questions. Based on our pilot study, most of our participants do not have difficulty in understanding our guidance explanation on how to create high quality questions based the question taxonomy with the Question Rubric and they can transferred what they have learned in improving their peers question.

3 Online Collaborative Question Refinement Method Design

In this section we will describe in detail about our online collaborative question refinement method and the tools which we are using in our evaluation study.

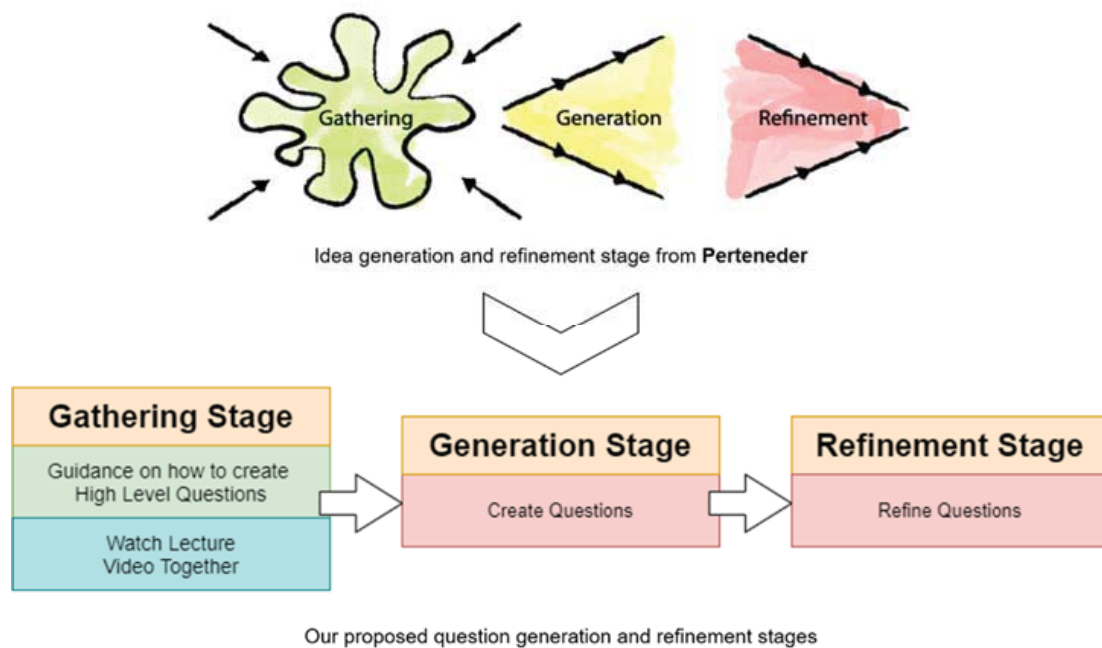


Fig. 1. Our proposed online collaborative question refinement stages adapted from the idea refinement stages

3.1 Refinement Strategy

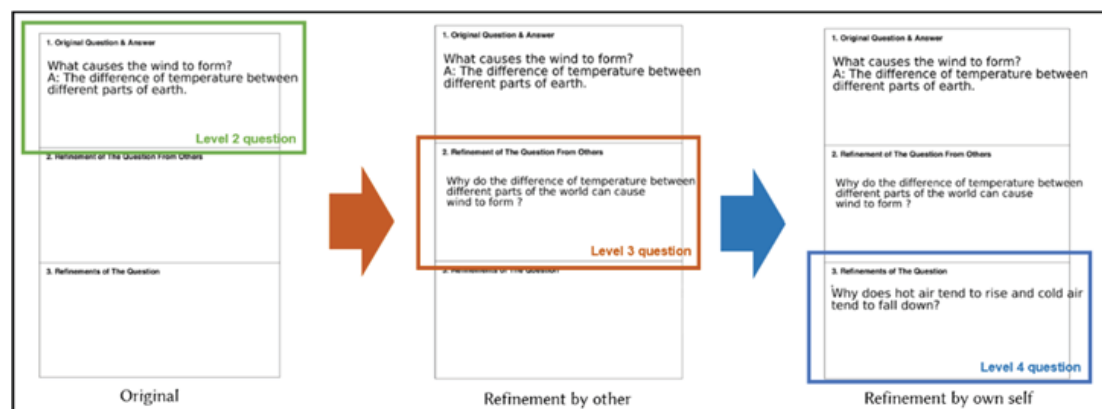


Fig. 2. Example of question refinement flow taken from our pilot study. Here the original question poser generate question on level 2 question. The other student then refined the question to the level 3 question and after the question was given to the original poser, she can improved again her question to the level 4.

Our method in question refinement strategy is based on idea refinement method by Perteneder[13]. We defined three stages of question generation and refinement for this study. The first step is the gathering stage where in our study, the students will be given a guidance and training on how to create high quality questions and improve existing questions into higher level. The second stage is the generation stage where students create questions as many as they can in limited time based on the learning

content in the video they have watched earlier. The third stage is the refinement stage where students actively refining other students' questions and refining their own questions after they got the improvement feedback from the others (see **Fig. 1**). We break-down refinement activity into several sub-activity: refining others question and refining own questions as can be seen on **Fig. 2**.

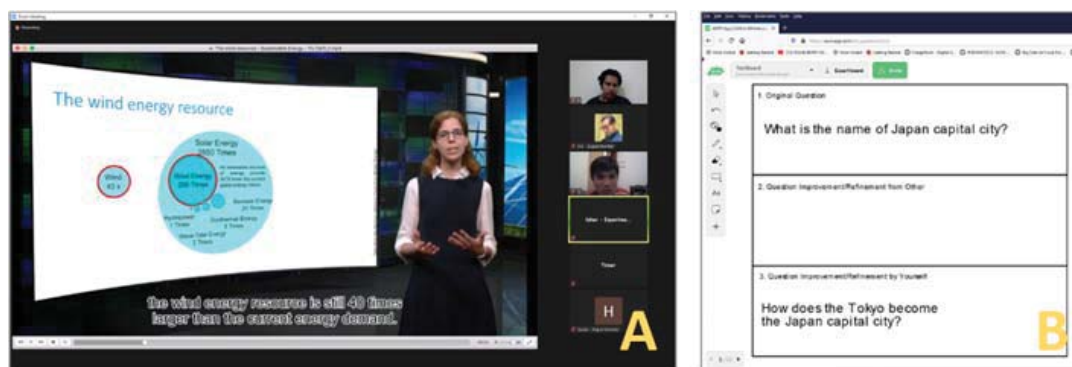


Fig. 3. A) Video conferencing system will be used for our online collaborative question refinement working space, B) Online board to create and refine questions using our Question Posing & Refinement form with three boxes.

3.2 Collaborative Refinement Tools

We designed our collaborative question refinement activity to be conducted in full online setting without any need for physical interaction, so all the interactions between participants and instructors will be online. For this, an online communication tools such as video conferencing and a shared working space that replicates papers are needed to generate and refine question in free format while the students are not in the same place. To achieve this, we use off-the-shelf tools which available freely in the Internet to support the online collaboration. For the synchronous video conference system in this study, we choose Zoom as it is fairly easy to use and provide decent video and audio quality (see **Fig. 3A**). Another reason we choose Zoom is because it also provides a collaboration tools such as screensharing and chat box.

As a place for writing and refining questions, we choose The Web Whiteboard Application or AWW App which is an online browser-based board tools to write text and draw object. In the online board we created Question Posing & Refinement form for students to write their questions into. Our form consists of three boxes, the box number 1 will be used to pose the original question. The box number 2 will be used as a place for question refinement based on the question in the box number 1 from another student. The box number 3 will be used as place for question refinement by the original question poser after they saw the refinement from other students (see **Fig. 3B**).

4 Initial Study

To investigate the effective form of collaborative online question generation and refinement, we planned to conduct a user study using between-subject design with three conditions that compare the effect between collaborative (two student and three student) and individual activities on the question refinement quality. The conditions are:

Individual Question Refinement. Participants will create and then refine their own question alone. We add refinement step in this group so we can observe differences between refining individually without any feedback from other people and refining after seeing other's question refinement in collaborative treatment.

Pair-student Question Refinement. Two participants will create questions and then refining each other's questions, and then refining their own question after being refined by their partner.

Triad-student Question Refinement. Three participants will create questions and then refining their first partner's questions, refining their second partner's questions, and refining their own question after being refined by their peers.

We have conducted small pilots with 18 participants divided into 6 individual groups, 3 pair groups, and 2 triad groups using our proposed method. Based on the pilots, participants were able to refine other student questions into higher quality in the refinement stages. As we can see on the example in **Fig. 2**, particularly in first refinement stage, other student was able to improve the original question in level 2 to level 3. We allow students in our pilots to create questions with or without answer, and our pilot participant here created the original question with an answer. The answer she provided in the original question then became clue for another student to refine the question into higher level.

5 Conclusion

In this study, we proposed fully online collaborative learning for question generation and refinement using off-the-shelf online tools. Our study aims to investigate how students can learn collaboratively but asynchronously by sequential activities of creating and refining questions on a learning material. From a small pilot study, we got the prospect that students can increase the quality of their questions based on Question Rubric Level.

References

1. Logtenberg, A., van Boxtel, C., van Hout-Wolters, B.: Stimulating situational interest and student questioning through three types of historical introductory texts. *Eur J Psychol Educ.* 26, 179–198 (2011). <https://doi.org/10.1007/s10212-010-0041-6>

2. Brindley, J., Blaschke, L.M., Walti, C.: Creating Effective Collaborative Learning Groups in an Online Environment. *IRRODL*. 10, (2009). <https://doi.org/10.19173/irrodl.v10i3.675>
3. Koschmann, T.: Dewey's contribution to the foundations of CSCL research. In: Proceedings of the Conference on Computer Support for Collaborative Learning Foundations for a CSCL Community - CSCL '02. p. 17. Association for Computational Linguistics, Boulder, Colorado (2002)
4. Al-Rahmi, W.M., Zeki, A.M.: A model of using social media for collaborative learning to enhance learners' performance on learning. *Journal of King Saud University - Computer and Information Sciences*. 29, 526–535 (2017). <https://doi.org/10.1016/j.jksuci.2016.09.002>
5. Papinczak, T., Peterson, R., Babri, A.S., Ward, K., Kippers, V., Wilkinson, D.: Using student-generated questions for student-centred assessment. *Assessment & Evaluation in Higher Education*. 37, 439–452 (2012). <https://doi.org/10.1080/02602938.2010.538666>
6. Denny, P., Luxton-Reilly, A., Simon, B.: Quality of student contributed questions using PeerWise. 9
7. King, A., Rosenshine, B.: Effects of Guided Cooperative Questioning on Children's Knowledge Construction. *The Journal of Experimental Education*. 61, 127–148 (1993). <https://doi.org/10.1080/00220973.1993.9943857>
8. King, A.: Enhancing Peer Interaction and Learning in the Classroom Through Reciprocal Questioning. 24
9. Yu, F.-Y.: Multiple peer-assessment modes to augment online student question-generation processes. *Computers & Education*. 56, 484–494 (2011). <https://doi.org/10.1016/j.compedu.2010.08.025>
10. Yeckehzaare, I., Barghi, T., Resnick, P.: QMaps: Engaging Students in Voluntary Question Generation and Linking. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–14. ACM, Honolulu HI USA (2020)
11. Gallagher, J., Aschner, M.J.: A Preliminary Report on Analysis of Classroom Interaction. *Merrill-Palmer Quarterly of Behavior and Development*. 9, 183–194 (1963)
12. Guthrie, J.T., Wigfield, A., Perencevich, K.C. eds: *Motivating reading comprehension: concept-oriented reading instruction*. L. Erlbaum Associates, Mahwah, N.J (2004)
13. Perteneder, F., Grossauer, C., Seifried, T., Walney, J., Brosz, J., Tang, A., Carpendale, S., Haller, M.: *Idea Playground: When Brainstorming is Not Enough*. 7

Promotion of continuous use of a self-guided mental healthcare system by using a chatbot

Takeshi Kamita¹, Atsuko Matsumoto², Tatsuya Ito¹ and Tomoo Inoue³

¹ Graduate School of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Japan

s1730527@u.tsukuba.ac.jp

² Graduate School of Comprehensive Human Science, University of Tsukuba, Tsukuba, Japan

s1130368@u.tsukuba.ac.jp

³ Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan

inoue@slis.tsukuba.ac.jp

Abstract. Although the stress check registration has been introduced in Japan as a measure to employee mental disorders in companies, there are few employees who wish to have an interview with an industrial physician even if they receive a high stress judgment. Thus, promotion of self-care is required at the same time. The authors developed the digital content from the SAT counseling method, a VR self-guided mental health care system that used the content, and found its stress reduction effect. Then, we developed a chatbot system on a smartphone that used the content, and also found its stress reduction effect. In this study, we conducted a two-week longitudinal study of the chatbot system, and found promotion effect for continuous use as well as the stress reduction effect.

Keywords: SAT counseling method, chatbot, mental healthcare.

1 Introduction

It is obligatory to conduct stress checks for business establishments with 50 or more employees in Japan. However, the number of employees who actually want to have an interview with an industrial physician is limited even if they are judged to be highly stressed [1]. Therefore, the Ministry of Health, Labor and Welfare also promote self-care. The authors have so far developed a self-guided mental health care system based on the SAT counseling method [2] with VR, and found that it had brought the stress reduction effect [3] [4]. However, at present, VR devices are not so popular and practical that they are used by company employees in-house. There also remains a motivation issue to encourage continuous use of the system, as the effect is fixed by the frequent stimulation of repeatedly watched images in the SAT method [5].

To address these issues, a self-guided mental healthcare system with a chatbot on smartphones, which users use easily and on a daily basis, was developed. Even if the user does not actively participate, he / she can use it more easily by following the guidance of the chatbot. In the previous study [6], we found that one time use of the chatbot system brought higher stress reduction effect and user's willingness to use compared to the non-chatbot system.

In this study, we conducted a two-week comparative survey between the chatbot system (the CB course) and a system composed of web pages without a chatbot (the WEB course) to investigate the chatbot effect through the period by evaluating both the continuity with the access histories and the stress reduction effect with the psychological scales.

2 Related work

2.1 Use of smartphones for mental healthcare

With the increasing demand for self-guided mental healthcare, research is being conducted on information systems that allow existing psychotherapies to be carried out on their own. Cognitive behavioral therapy is one of the most popular psychotherapies. Researches have been conducted to use this therapy as a complementary tool for treatment and counseling [7] or as a self-guided tool by making its digital content. Some have been commercialized as smartphone applications [8]. The cognitive bias adjustment method [9], which aims at alleviating a specific bias exhibited by a person with depression or strong anxiety in the process of thinking, is also being digitized [10]. “Mood Mint” [11], a commercial smartphone app that uses this method, increases the awareness of positive information by repeating the training of quickly tapping on the screen of smiling faces that are mixed with multiple negative face images, and reduces the degree of attention to events with negative cognition. However, since the stress reduction effect is not felt with each use, giving point incentives as in the token economy [12] is used to encourage continued use.

Mindfulness using meditation techniques is also actively studied and used in psychological practice [13]. Research and development for digital content is also proceeding [14], and the smartphone application “Headspace” [15] is commercially available. This app provides programs for each purpose such as coping with anxiety and coping with depression, and assists in the progress of meditation. It is necessary to listen to and carry out the audio lecture of 10 minutes 10 to 30 times in one program, which requires high motivation for the user.

2.2 Use of smartphones for mental healthcare

A chatbot is a program that automatically conduct conversations through text or voice, which have evolved since the development of ELIZA [16] in 1966. The development environment of a chatbot was opened on social networking service platforms such as Facebook [17] and LINE [18] in 2016, and it became possible to provide a chatbot as their message functions. A chatbot have been developed to support interpersonal skills as a training component of a depression treatment program [19]. Chatbots specializing in reducing and coping with stress problems are also being studied. Gaffney et al. have developed a chatbot-based self-help program MYLO based on Perceptual Control Theory. As a result of comparing the effectiveness of MLYO with ELIZA, MYLO and ELIZA were associated with relief of pain, depression, anxiety and stress. MYLO led to greater problem solving and was considered more effective [20].

In a two-week study using “Woebot”, a system that encourages users to learn by providing knowledge of cognitive-behavioral therapy through a chatbot, participants

continuously received the self-help content based on cognitive-behavioral therapy via chatbot and as a result showed that anxiety improved significantly [21]. However, in the case of Woebot, programmed answers and scenarios are intended for use by people with a strong depression tendency, and the users are assumed to have a certain level of motivation.

In this study, we assume users with various mental health issues and levels of motivation. Therefore, we have developed a self-guided mental healthcare system which can be easily applied to various issues, bring an immediate stress reduction effect, and does not require user's higher motivation to use.

3 Digital content of the SAT method

3.1 SAT method

The SAT method is counseling therapy method in the form of an interview developed by Munakata, and is composed of multiple techniques [2]. There are temperamental coaching method and health coaching method to clarify the clients' problems and characteristics and motivate them, and SAT image therapy for solving stress problems, which consists of emotional stabilization therapy and behavior modification therapy. The self-guided mental healthcare app has been developed adopting the emotional stabilization therapy.

The emotional stabilization therapy is a technique that can be used to alleviate and solve daily stress problems such as current stress problems, past problems, and relief of physical symptoms, and can be carried out by self if trained. First, the client recalls a stress scene, and the disliked image prompts the perception of a feeling of discomfort caused by the reaction of the body such as the stomach tingling. In response to this discomfort, the counselor presents a list of landscape images printed on a paper that make us imagine the gentle light (Fig. 1), let the client select one, remind of the image that the discomfort part is wrapped and healed by the light, resulting discomfort reduction (optical image method) [5]. In addition, the image of the smile is selected by the client, and the image is reminded of the sense of security and safety that the person in the image is on the client's side and protected. Then, by promoting awareness of the client's own commitments, captivity, and beliefs about the problems that cause stress, worries, moods, pains, and the image of what one should be, release them, enhance self-affirmation, and solve problems (surrogate face representation method [5]).



Fig. 1. Printed image list used in SAT method

3.2 Digital-SAT method

In the emotional stabilization therapy, the process that the counselor asks questions and the client responds is repeated. The counselor leads the operation by calling out or asking the eyes to be closed as needed from the dialogue with the client, facial expressions, and body movements during the operation. We have developed the Digital- SAT method as a method that allows this therapy to be self-guided using a Head Mounted Display (HMD) or a smartphone without the guidance of a counselor [3][6] [22].

The structure and procedure of the Digital-SAT method were defined as follows: (1) know your own mental state (Assessment part), (2) reduce stress (Solution part), and (3) (depending on the personal mental characteristics clarified in (1) and (2)) learn to improve mental resistance (Learning part). The parts other than the learning part are the targets of this research.

In the assessment part, the user's stress state and characteristics are measured, and the psychological check test used in SAT method (Table 1) is carried out for the purpose of clarifying the changes and effects before and after using the application, and the results are browsed by the user.

Table 1. Psychological check test

Scale	Contents	Total score range (SAT method criterion)
State-trait anxiety inventory (STAI)	The tendency to become anxious, not state anxiety that varies over time, but a vague degree of anxiety that reflects an individual's past experience. [23]	20-80(20-31 lower/32-34mid/35-41higher/42-80 much higher)
Self-rating depression (SDS)	The depressive symptoms in mood, appetite, and sleep. [24][25]	20-80(20-35 none/ 36-48 lower/49-68 higher/ 69-80 painful)
Self-esteem	The degree to which a person has a good or positive image for self. Higher self-esteem is more likely to be able to cope with stress and less likely to cause anxiety or depression. [26]	0-10(0-6 lower/7,8 mid/9,10 higher)
Emotional support network from family	The degree of perception that seems to be supported emotionally and psychologically from the family.	0-10(0-5lower/6-7mid/8-10higher)
Emotional support network from peers	The degree of perception that seems to be supported emotionally and psychologically from outside the family (e.g., colleagues, friends, superiors, etc.)	0-10(0-5lower/6-7mid/8-10higher)
Health counseling needs	Whether or not the response to stress manifested in the mind, body, or behavior, and to the extent.	0-20(0-6 lower/7-10 mid/11-20 higher)
Self-repression behavioral trait	The behavioral characteristics that suppress one's own feelings and thoughts.	0-20(0-6 lower/7-10 average/11-14 slightly higher/ 15-20higher)

Problem solving behavioral trait	The behavior that seeks to respond positively, effectively, and realistically to immediate challenges and issues.	0-20(0-6lower/7-10slightly lower/ 11-14slightly higher/15-20higher)
Emotional Interpersonal dependency inventory	The degree of emotional dependence and expectancy for others.	0-15(0-3lower/4 slightly lower/5-8 slightly higher/9-15higher)
Difficulty in recognizing emotions	The tendency to avoid feeling of one's own feelings, either subjectively or involuntarily. Higher scores tend to accumulate stress and become chronic with physical symptoms even if they are not aware of them.	0-20(0-5lower/6-8higher/9-20 much higher)
Self-pity	The degree to which they are sympathetic with their own treatment and decide not to abandon themselves alone.	0-20(0-5 lower/ 6-8slightly higher/ 9-20higher)
Self-dissociation	The degree to which a person dissociates from oneself who is troubled by a serious problem and is calmly observing oneself.	0-20(0-3 lower/ 4-7 Slightly higher/ 8-20 higher)
Self-denial	A lack of interest or motivation in trying to heighten oneself, such as being happy. Higher scores tend to give up and feel guilty.	0-20(0-2 lower/3-4 mid/5-20 higher)
PTSS (Post-traumatic stress syndrome)	Having experienced or observed serious crises of oneself or others and tends to flashback to tension and fear of releasing noradrenaline when encountered in certain circumstances.	0-10(0-1 lower/ 2-3mid/ 4-10 higher)

In the Digital-SAT method, the process is advanced by turning the page on the screen, so individual steps of emotional stabilization therapy are disassembled, and each explanation is simplified to form the solution part (Table 2). First, by recalling the stress he/she is currently having (Q1) and by converting an unpleasant image into a color or shape (Q3, Q4), the perception of physical discomfort is promoted (Q5), and the stress level is clarified by imagining it as a number(Q6).

Next, the physical discomfort is relieved with a light image (Q7), and a substitute facial representation image is used to foster a sense of security and safety (Q8), and a decrease in stress level is confirmed (Q9). If he/she can have an image of a positive personality (Q.10), he/she will be encouraged to recognize that his/her perception of stress problems will differ (Q11, Q12). Finally, confirm how much the stress problem has become felt and end (Q13).

Table 2 Processes of digital-SAT method

No.	Question item
Q1	What are you feeling stressed right now? Think of it
Q2	How much stress is it? (5-point Likert scale)
Q3	What color is the stress?
Q4	If you compare that color to a shape?
Q5	Where do you feel discomfort in your body?
Q6	What is your current stress level? (Answer from 0% to 100%)
Q7	What light seems to heal this discomfort?
Q8	Please choose the comfortable face that came into your eyes.
Q9	How do you feel with this companion? Will you be healed?
Q10	What kind of personality are you likely to have when the stress disappears?

Q11	How do you deal with stress with this personality?
Q12	Does that work?
Q13	How much stress did you feel?

3.3 Implementation of CB course

We implemented a chatbot application as the LINE application, the most popular social networking service (SNS) in Japan. In this study, as the functions of the assessment part, "Mental meter" (measurement of characteristic anxiety scale) and "My data" (radar chart display of results of the check test and the mental meter) were added (Fig. 4) to the conventional "check test" (Fig. 3). In addition, "My Golden Rule" (repeated browsing of the optical image and surrogate face representation image selected in Quick Care) was added to the conventional "Quick care" as the function of the solution part, and "Login" was added as a common function. Each function is accessed from the top menu on LINE (Fig. 5).

The user selects and registers a "friend account" dedicated to the chatbot application on LINE. Each menu can be used by the user selection, or can be used by selecting a link in the automatic delivery message such as "How are you today?" from the chatbot that guides each menu (Fig. 5). The check test is answered on the screen in Fig. 3. When he/she selects "quick care", which is the main stress care program, the questions in Table 2 are presented in order, and the process proceeds while answering the questions and selecting images (Fig. 6).



Fig. 3. The screen of check test



Fig. 4. The result of check test

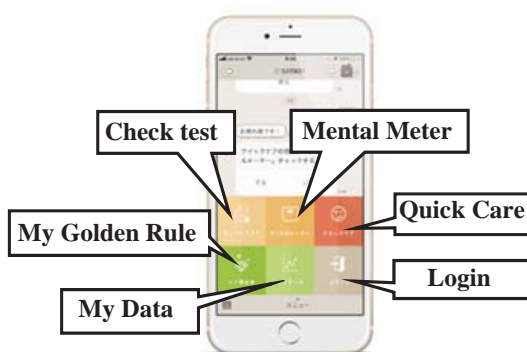


Fig. 5. Top menu



Fig. 6. The screen of viewing light image

4 Study of the chatbot system

We evaluated the number of access to the system and the stress reduction effect by using the system by a longitudinal study for 2 weeks, and examined the effectiveness of using the chatbot in the self-guided mental healthcare system. As a control group, we used a WEB course that was created using only web pages without using a chat-bot. In order to control the experimental conditions, both groups used smartphones to use the system. This research was approved by the Ethics Review Committee of the authors' institution (No. 30-105).

4.1 Study method

We conducted a study to 30 participants, who were graduate students and working people selected by the snowball sampling. The participants were randomly assigned into two groups of an experimental group that experienced the CB course and a control group that experienced the WEB course (experimental group: N=21, control group: N=9). The WEB course was implemented so that the question / selection scene and the image browsing scene proceeded in the same way as the CB course. The difference from the CB course was that the layout of the initial screen (Fig. 7) and the quick care function was in the form of answering the displayed questions (Fig. 8) instead of in an interactive form with the chatbot.

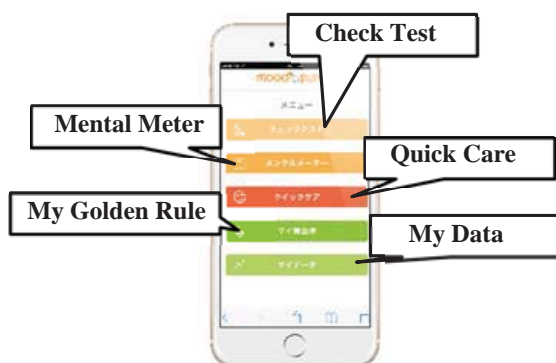


Fig. 7. Top menu of the WEB course



Fig. 8. The screen of viewing light image for the WEB course

The procedure was shown on Table 3. On the first day, participants were asked to gather in a meeting room for each group and an orientation was conducted. From the second day, although we told the participants that the course should be experienced at least once a day, but we left it up to them to decide whether or not they actually continue to use the course. Every psychological check test was mandatory. From the day after the start of the study, participants were informed daily via a chatbot in the experimental group and by email in the control group to encourage system use (table 4).

4.2 Data analysis

Regarding the continued use of the system, the number of accesses that each participant accessed to each function during the study period was recorded and evaluated. The av-

erage number of accesses for all participants was calculated for each function, and the Mann-Whitney U test was performed on the difference between the experimental group and the control group (5% level).

Table 3. Procedure

Group	Experimental Group	Control Group
# of participants	21	9
Group overview	Conduct CB course at least once a day Chatbot notification	Conduct WEB course at least once a day Email notification
Procedure		
Day1 Orientation	<ol style="list-style-type: none"> 1. Overview description (5 min.) 2. Consent form (2 min.) 3. Logging in to the app 4. Psychological check test (10 min.) 5. Principle and how-to use the course (30 min.) 6. Course experience (10 min.) 7. Psychological check test (10 min.) 8. Personal feedback of the test result (5 min.) 	
Day2-Day13	Conduct CB course Chatbot notification (Weekdays)	Conduct WEB course Email notification (Weekdays)
Day14	Psychological check test	

Table 4. Notification content

No.	Notification Content
1	Hello! I look forward to working with you. How are you feeling today?
2	Today, let's take care of things that should be solved by the Check Test. (From the Check Test question)
3	When you meet a person who feels stress, or when you feel nervous, take Quick Care, or if you don't have time, just look at My Golden Rule.
4	Welcome. Today, let's check the state of stress with Mental Meter.
5	Hello! How are you today?
6	Welcome. Let's eliminate what may be causing stress today. Please select from below (From the Check Test question)
7	Welcome. How are you today?
8	Check your stress condition with Mental Meter.
9	One more thing. Let's eliminate the cause of stress. Please select from below (From the Check Test question)

The stress reduction effect was evaluated using the score change of the psychological check test consisting of 165 questions in 14 categories used in the SAT method. For the results of the check test, Friedman's test (5% level) was performed on the score change of the first day before and after each course use, and 7 days and 14 days after the start of the study. As a post hoc test, Wilcoxon's signed rank test with Bonferroni correction (5% level) was performed. IBM SPSS Statics ver.25 was used for statistical processing in this study.

4.3 Result

Number of Accesses. Table 5 shows the average number of accesses for each function in each group (experimental group N = 21, control group N = 9). Regarding the functions of the assessment part, the average number of accesses to the check test was 5.14

in the experimental group and 3.00 in the control group, showing a significant difference ($p = 0.004$). The number of accesses to the mental meter was 10.76 in the experimental group and 1.44 in the control group, showing a significant difference ($p = 0.000$). The number of accesses to the My Data was 11.05 in the experimental group and 3.44 in the control group, showing a significant difference ($p = 0.006$). In the entire assessment part, the experimental group accessed 26.95 times, whereas the control group 7.89 times, showing a significant difference ($p = 0.000$).

Table 5. Number of accesses

	Average # of Access		p
	Experimental	Control	
Total	42.57	15.78	0.000*
Assessment part subtotal	26.95	7.89	0.000*
Check Test	5.14	3.00	0.004*
Mental Meter	10.76	1.44	0.000*
My Data	11.05	3.44	0.006*
Solution part subtotal	15.62	7.89	0.014*
Quick Care	9.43	4.89	0.010*
My Golden Rule	6.19	3.00	0.121

Mann-Whitney's U test: * $p < 0.05$

In the solution part, the number of access to quick care was 9.43 in the experimental group and 4.89 in the control group, showing a significant difference ($p = 0.010$). In the My Golden Rule, the number of accesses in the experimental group was 6.19 on average, while that in the control group was 3.00. Although the average number of accesses in the experimental group was higher than that in the control group, no significant difference was found ($p = 0.121$). In the solution part as a whole, a significant difference was found between the groups, 15.62 times in the experimental group and 7.89 times in the control group ($p = 0.014$). Finally, the total number of accesses was 42.57 in the experimental group and 15.78 in the control group, showing a significant difference ($p = 0.000$).

Then, Fig. 9 shows the average number of daily accesses to the system, where accesses to multiple pages was counted as one access to the system. Both the experimental group and the control group decreased sharply after the second day, but the experimental group continued to average more than once, while the control group continued to fall below the average of 1.0 overall, the experimental group had a higher average daily access count than the control group.

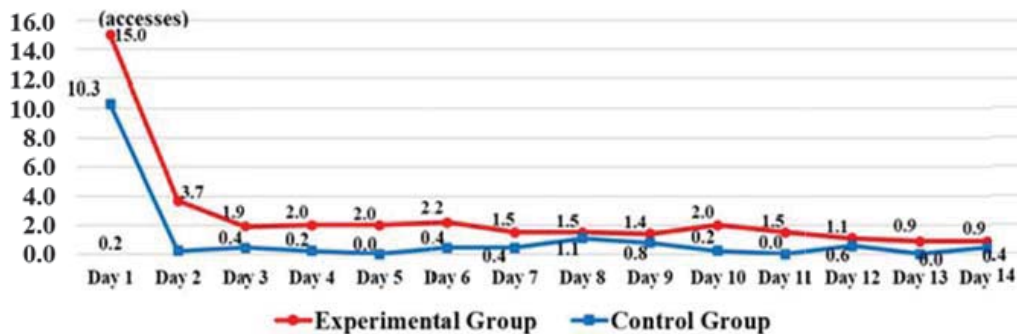


Fig. 9. Change in average number of accesses per day

Changes in the psychological check test scores. Fig. 10, Fig. 11, Fig. 12 and Fig. 13 show the score time series transitions which have significant changes through the study period. In the experimental group, there were significant changes on the score of the self-trait anxiety inventory (STAI), the self-rating depression (SDS), the self-esteem and the emotional support network from peers. In the control group, there were no significant changes on the score of any scales.

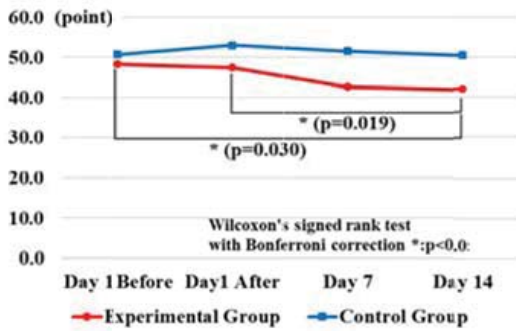


Fig. 10. STAI score time series

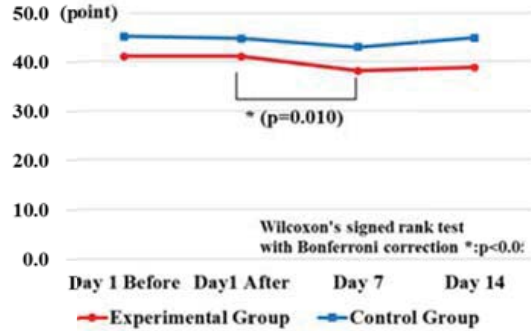


Fig. 11. SDS score time transition

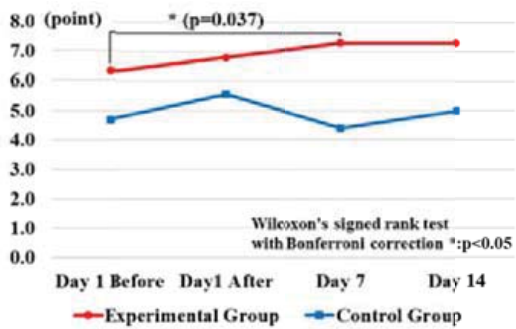


Fig. 12. Self-esteem score time series transition

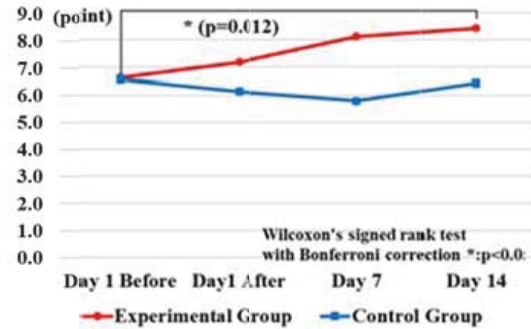


Fig. 13. Emotional support network from peers score

4.4 Discussion

From the changes in the scores of the psychological check test, significant improvements were found on the scales of the self-esteem, the STAI, the SDS, and the emotional support network from peers. According to the interpretation of the SAT method, the changes in these scales are as follows. Improving the sense of self-esteem means that one's self-image is positively grasped, the expectation that he/ she can overcome the difficulty is increased, anxiety and depression are reduced, and stress is reduced. In addition, the emotional support network from peers is a scale that measures the degree that he/ she feels there is a person who evaluates and understands him/ her in his/ her peers. The improvement of the score means that self-esteem is increased and the stress state of mind and body is reduced. It can be evaluated that the changes in the scale scores this time was also qualitatively valid. On the other hand, there was no significant difference in the WEB course in any of the scales.

Regarding the number of accesses, it was confirmed that the CB course had a significantly higher number of accesses. The average daily accesses also showed the differ-

ences between the CB course to continue the access at least once a day approximately and the control group less than once a day. In the previous survey [6], we found the possibility that the CB course more motivated users to continue use than WEB course, and this study also showed the results to support the possibility.

Based on the above discussion, it is suggested that the chat bot guidance of the mental healthcare process and notification brought more accesses to the system and stress reductions as a result. In addition, it is very interested in the significant improvement on the emotional support network from peers score. This suggests that dialogues asking or responding to questions from a chatbot may have provided a sense of support from others, and triggered to raise the self-esteem, and then brought an effect on the stress reduction as mentioned above. As stated at the beginning, it was desirable to continue to use the course in self-care, as well as the SAT method, which ensures that the effect is established by repeated frequent stimulation. The user's motivation of continuous use was a challenge. In this study, it is found that the use of chatbot could be effective in solving this issue.

5 Conclusion

In this study, we compared the self-guided mental healthcare system using a chatbot, which course was developed according to the Digital SAT method, and the system without a chatbot for two weeks of continuous use. The stress reduction effect and the degree of continuous use were evaluated. This study shows that the use of a chatbot may be an effective means for solving the issue of how to motivate continuous use, which has been presented in our previous study.

References

1. Ministry of Health, Labour and Welfare. 2017. Implementation of the stress check system. (in Japanese). <https://www.mhlw.go.jp/file/04-Houdouhappyou-11303000-Roudoujunkyokuanzeneseibu-Roudoueseika/0000172336.pdf> last accessed 2019/1/13.
2. Munakata, T.: SAT therapy. Kanekoshobo, Japan (2006).
3. Kamita, T., Matsumoto, A., Munakata, T., Inoue, T.: Realization of self-guided mental healthcare through the digital content based on the counseling technique SAT method, IPSJ Transactions on Digital Content, vol.6, no.2, pp.32-41 (2018).
4. Matsumoto, A., Kamita, T., Munakata, T., Komazawa, M., Itao, K., Inoue, T.: Stress Reduction Effect in Female Managers of a Self-Guided Mental Healthcare VR Content for Smartphone Based on the SAT Counseling Technique: A Psychological Scale and Heart Rate Variability Analysis. Applied human informatics, 209, 1(1):18-37 (2019).
5. Munakata, T.: The applicability of the simple edition of SAT method in promoting universal health, Journal of Health Counseling, 17, pp. 1-12 (2011).
6. Kamita, T., Ito, T., Matsumoto, A., Munakata, T., Inoue, T.: A Chatbot System for Mental Healthcare Based on SAT Counseling Method. Mobile Information Systems. Hindawi, Volume 2019, Article ID 9517321, 11p (2019).

7. Batterham, P.J., Calear, A.L., Farrer, L., McCallum, S.M., Cheng, V.W.S.: Fitmindkit: Randomized controlled trial of an automatically tailored online program for mood, anxiety, substance use and suicidality. *Internet Interventions*, Vol.12, pp. 91- 99 (2018).
8. Torous, J., Levin, M.E., Ahern, D.K., Oser, M.L.: Cognitive Behavioral Mobile Applications: Clinical Studies, Marketplace Overview, and Research Agenda. *Cognitive and Behavioral Practice*, Vol.24, Issue 2, pp.215-225, May (2017).
9. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G. M., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A. T.: Student Life: assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp.3-14 (2014).
10. Steel, C., Wykes, T., Ruddle, A., Smith, G., Shah, D. M., Holmes, E. A.: Can We Harness Computerized Cognitive Bias Modification to Treat Anxiety in Schizophrenia? A First Step Highlighting the Role of Mental Imagery. *Psychiatry Research*, Vol. 178, No. 3, pp. 451-455 (2010).
11. Mood Mint, <http://www.biasmodification.com/>, last accessed 2020/5/3.
12. Dickerson, F.B., Tenhula, W.N., Green-Paden, L. D.: The token economy for schizophrenia: review of the literature and recommendations for future research. *Schizophrenia Research*, Vol.75, pp.405-416 (2005).
13. Kabat-Zinn, J.: An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: Theoretical considerations and preliminary results. *General Hospital Psychiatry*, Vol. 4, No. 1, pp. 33-47(1982).
14. Bennike, I. H., Wieghorst, A.: Online-based Mindfulness Training Reduces Behavioral Markers of Mind Wandering. *Journal of Cognitive Enhancement*, Vol.1, issue 2, pp.172-181 (2017).
15. "Headspace". <https://www.headspace.com/>, last accessed 2020/5/3.
16. Weizenbaum, J. ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, vol.9, no.1, pp.36-45 (1966).
17. Facebook, <https://www.facebook.com/>, last accessed 2020/5/3.
18. LINE, <https://line.me/ja/> last accessed 2020/5/3.
19. Elmasri, D. and Maeder. A.: A conversational agent for an online mental health intervention. *Brain Informatics and Health*, pp. 243-251 (2016).
20. Gaffney, H., Mansell, W., Edwards, R., Wright, J.: Manage your life online (MYLO): A pilot trial of a conversational computer-based intervention for problem solving in a student sample. *Behavioral and cognitive psychotherapy*, 42(6), pp. 731-746 (2014).
21. Fitzpatrick, K.K., Kara, K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, vol. 4, no. 2, p. e19 (2017)
22. Kamita, T., Ito, T., Matsumoto, A., Munakata, T., Inoue, T.: A WEB Course Based on the SAT Counseling Method that Reduces Anxiety by Continuous Use. *International Journal of Informatics Society (IJIS)*, vol.11, no.2, pp.75-84(2019).
23. Spielberger, C.D: STAI manual. Palo Alto. Calif. Consulting Psychologist Press (1970).
24. Zung, W.K.K.: A self-rating depression scale. *Archives of general psychiatry*, vol.12, pp.63-70 (1965).
25. Fukuda, K., Kobayashi, S.: SDS Manual. Sankyobo, Kyoto (1973).
26. Munakata, T: Health and disease from the view point of behavioral science. *Medical Friend Co. Ltd. Tokyo*, pp.25-29, pp.128-129 (1996).

Validation and Verification Assessment of Genetic Counseling and Testing

Ekaterina Auer¹, Wolfram Luther²

¹ Faculty of Technology, University of Applied Sciences Wismar, Germany
ekaterina.auer@hs-wismar.de

² Computer Science and Applied Cognitive Science,
University of Duisburg-Essen, Germany
wolfram.luther@uni-due.de

Keywords: Verification and validation assessment · genetic testing · family history assessment model · BRCA1/2 mutations · binary interval decision tree

Extended Abstract

Verification and validation (V&V) assessment [17] plays an important role in a variety of applications, in particular, in safety critical ones. In our previous work, we employed it in the context of visual analytics [3] and biomechanics [4], with a focus on its reliability. In this contribution, we consider BRCA (tumor suppressor genes) related cancer and BRCA1/2 mutation probability prediction tools [1, 26] as a case study. Having introduced the audience to the topic of reliable V&V assessment, we point out existing standards as well as quality criteria and metrics for V&V of genetic counseling and clinical molecular testing [20] with a special emphasis on risk assessment, sense making and decision making under various forms of uncertainty.

In the context of our case study, reliability of test interpretations and counseling conclusions is especially important since they have a direct influence on humans and their decisions. Although most kinds of breast, ovarian, prostate, and pancreatic cancers are sporadic, a minority are caused by germline mutations in breast cancer susceptibility BRCA1/2 genes. These mutations reduce their tumor-suppressor qualities essential for the repair of DNA double-strand breaks by homologous recombination (HR). However, the HR pathway for DNA repair is disrupted by pathogenic mutations not only in BRCA1/2, but also in other involved genes [27]. This and other factors mentioned below are responsible for the high degree of uncertainty present in genetic counseling and testing.

Claus tables [10], BRCAPRO [12], BOADICEA [9], and Penn II risk model [25] are four important mathematical models for computing the probability of breast or ovarian cancer based on Mendelian genetics and the Bayes theorem [5]. Typical genetic counseling tools used for identifying candidates for whom genetic testing is necessary are the family history assessment tool FHAT [15], the referral screening tool RST [6], the Manchester scoring system (MSS) [12] (validated in [19]), and the family history (FH) screen FHS-7 [2]. Influential studies which aim at validating various approaches for predicting BRCA1/2 mutations are the following. In [24], Parmigiani et al. quantify the accuracy of seven publicly available models (including BRCAPRO, Penn II and FHAT) employing them for 3342 persons to predict the status of a mutation carrier. The study is based on three population-specific sample groups of participants from research and eight samples from genetic counseling clinics. As validation criteria, it relies on sensitivity and specificity of predictions as well as on how well a model discriminates between individuals testing positive for a BRCA 1/2 mutation and those testing negative using statistical methods as a metric.

In [18], James et al. apply six models to 257 FH. In [23], Panchal et al. evaluate the performances of seven currently used risk models (again including BRCAPRO, Penn II and FHAT as well as MSS and BOADICEA) with 300 patients from a large familial program relying as criteria on high sensitivity of predictions, possibility of simple data collection and entry, and availability of BRCA score reporting. Quite recently, Himes [16] assesses five currently widespread screening tools for genetics referral, namely, FHS-7, Pedigree Assessment Tool [28], MSS, RST, and Ontario Family History Assessment Tool (Ontario-FHAT) [15]. Based on ancestries of 85 women, metrics for the criteria of sensitivity, specificity, positive and negative predictive value are calculated to describe each tool's ability to identify patients with high/increased risk as defined by Claus tables. Finally, the meta-study in [22] reviews 103 medical studies and 110 research articles (with 92712 individual patients) to assess them w.r.t. research questions regarding methodology, scientific rigor, study parameters, relevance, quality criteria and metrics, performance, accuracy, limitations as well as adverse effects and benefits for the patients.

From these studies and meta-studies, it is evident that there is significant uncertainty both in data and in the processes/models. Aside from the usual, in this case less important, sources of uncertainty due to the used numerical methods and the modeling error, the major uncertainty factors in this context are the age of cancer onset in a patient's relative, degree of kinship of the patient and the affected as well as number of instances and kinds of cancer in the family tree.

To evaluate and compare the genetic counseling and probability prediction tools for BRCA1/2 mutation, we take a critical look at recent meta-studies and evidence reports. To take into account uncertainty in patient data in combination with the advantages of established models and to ensure validation, we present a new multi-criteria categorical counseling test. It combines the binary decision structure [29] of RST with the features of FHAT and MSS and uses an accumulated interval risk function. We propose to use interval scores and lower limits for probabilities of pathogenic variants since crisp scores do not reflect the available information correctly. Formerly, the result of the test combined the presence of different constellations of cancers in the family history; now, a referral vector U with components storing decisions in the form of an interval score and the corresponding probability where possible is produced to assign participants to low, moderate and high risk categories. Our validation for the scores is based on prevalence and frequency of mutations in BRCA1/2 correlated with various combinations of personal and family histories of cancer given by tables reported in [14]. In this way, patients can be assigned a risk category characterizing the likelihood of a BRCA1/2 mutation [7] more accurately.

To summarize, it is our opinion that requirements for genetic risk evaluation in the counseling and testing process of individuals and families affected by breast and ovarian cancer caused by pathogenic mutations should address the following key points:

1. Each involved discipline should recognize the need for standardized evaluation and V&V of processes and their models as well as for fusion of conclusions from their outcome. For this, appropriate quality criteria and their metrics as well as standardized procedures need to be used.
2. The collected heterogeneous data should be standardized with a focus on the test participants' numbers, origin and age. Additionally, the data should be assessed and if necessary completed to make the application of the specified quality criteria and metrics possible. Although the number of subjects in the studies has increased considerably over the years, so has the number of examined parameters, which the test results strongly depend on, with the additional data coming from women of varying socioeconomic and ethnic groups.

Currently, comparability and calibration in meta-studies is not always possible in a satisfactory manner due to the heterogeneity and varying availability and quality of the data, tests and conducted studies. Aside from Nelson [21, 22], most cited meta-studies are based on a relatively small samples that are not representative of the age groups, regions of origin and type of pathologies as described in the ClinGen [11] clinical validity framework.

3. Uncertainty in the data should be represented by suitable data types and propagated through the models and processes. Only after that can the results be judiciously evaluated according to standardized criteria and the recommendations of the experts feasibly merged into proposals for risk management strategies that are understandable and not harmful for the patients [26]. Finally, effects of these strategies need to be assessed for subsequent referral and tests to be adapted accordingly.

References

1. Amir, E. et al.: Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models. *JNCI* 102 (10) Oxford University Press (2010) 680-691
2. Ashton-Prolla, P. et al.: Development and validation of a simple questionnaire for the identification of hereditary breast cancer in primary care. *BMC Cancer* 283(9) (2009) 9 p.
3. Auer, E. Result Verification and Uncertainty Management in Engineering Applications. Dr. Hut, 2014. Habilitation Monograph
4. Auer, E., Luther, W., Weyers, B.: Reliable Visual Analytics, a Prerequisite for Outcome Assessment of Engineering Systems. Special Issue of the 11th Summer Workshop on Interval Methods. *Acta Cybernetica* 24(3) (2020) 287-314
5. BayesMendel Lab Harvard University – BRCAPRO
<https://projects.iq.harvard.edu/bayesmendel/brcapro>
6. Bellcross, C. A. et al.: Evaluation of a breast/ovarian cancer genetics referral screening tool in a mammography population. *Genetics in Medicine* 11 (2009) 783-789
7. Bellcross, C. A., Peipins, L. A. et al.: Characteristics associated with genetic counseling referral and BRCA1/2 testing among women in a large integrated health system. *Genet. Med.* 17 (2015) 43–50.
8. Berry, D. A. et al.: Probability of Carrying a Mutation of Breast–Ovarian Cancer Gene BRCA1 Based on Family History. *J. Natl. Cancer Inst.* 89 (1997) 227-237
9. BOADICEA: Centre for Cancer Genetic Epidemiology, Cambridge University
<https://ccge.medschl.cam.ac.uk/boadicea/>
10. Claus, E. B., Risch, N. et al.: Autosomal dominant inheritance of early onset breast cancer: implications for risk prediction. *Cancer* 73(3) (1994) 643–651
11. ClinGen. Gene-disease validity. <https://www.clinicalgenome.org/curation-activities/gene-disease-validity/>
12. Evans, D. G. R. et al.: A new scoring-system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO. *J. Med. Genet.* 41 (2004) 474–480
13. Frank, T. S., Manley, S. A. et al: Sequence analysis of BRCA1 and BRCA2: Correlation of mutations with family history and ovarian cancer risk. *J. Clin. Oncol.* 16 (1998) 2417-2425
14. Frank, T. S. et al.: Clinical Characteristics of Individuals With Germline Mutations in BRCA1 and BRCA2: Analysis of 10,000 Individuals. *J. Clin. Oncol.* 20(6) (2002) 1480-1490
15. Gilpin, C. A. et al: A preliminary validation of a family history assessment form to select women at risk for breast or ovarian cancer for referral to a genetics center. *Clin. Genet.* 58 (2000) 299–308
16. Himes, D.: Breast cancer risk assessment: Evaluation of screening tools for genetics referral. *J Amer. Assoc. of Nurse Practitioners* 31 (10) (2019) 562-572

17. IEEE standard for system, software, and hardware verification and validation. IEEE Std 1012-2016 (Sept 2017) 1-260
18. James, P. A. et al.: Optimal Selection of Individuals for BRCA Mutation Testing: A Comparison of Available Methods. *J. Clin. Oncol.* 24(4) (2006) 707-715
19. Karst, K. et al.: Validation of the Manchester scoring system for predicting BRCA1/2 mutations in 9,390 families suspected of having hereditary breast and ovarian cancer. *Int. J. Cancer:* 135 (2014) 2352–2361
20. Mattocks, C. J. et al.: A standardized framework for the validation and verification of clinical molecular genetic tests: *European Journal of Human Genetics* 18 (2010) 1276–1288
21. Nelson, H. D. et al.: Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer: Systematic Review to update the U.S. Preventive Services Task Force Recommendation. Rockville, MD: Agency for Healthcare Research and Quality (2013)
22. Nelson, H. D. et al.: Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer in Women - Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *Clinical Review & Education* (2019) 666-685
23. Panchal, S. M. et al.: Selecting a BRCA risk assessment model for use in a familial cancer clinic. *BMC Medical Genetics* 9 (2008) 116-124
24. Parmigiani, G. et al.: Validity of Models for Predicting BRCA1 and BRCA2 Mutations. *Ann. Intern. Med.* 147(7) (2007) 441–450
25. Penn II Risk Assessment Model: <https://pennmodel2.pmacs.upenn.edu/penn2/>
26. Peshkin, B. N. et al.: Genetic testing and management of individuals at risk of hereditary breast and ovarian cancer syndromes. Wolters Kluwer, www.UpToDate.com (2020)
27. Stoppa-Lyonnet, L.: The biological effects and clinical implications of BRCA mutations: where do we go from here? *European Journal of Human Genetics* 24 (2016) S3–S9
28. Teller, P., Hoskins, K. F., Zwaagstra, A., et al.: Validation of the pedigree assessment tool (PAT) in families with BRCA1 and BRCA2 mutations. *Ann Surg Oncol.* 17(1) (2010) 240–246.
29. Zhang, Q., Varshney, P. K.: Towards the fusion of distributed binary decision tree classifiers (1999). <https://www.researchgate.net/publication/228989654>

Presentation Slides

1 Overview

- Introduction
- Verification & validation assessment for process outcome
- V&V of clinical molecular genetic tests
- Meta-study and evidence report
- Risk assessment, genetic counseling and testing for BRCA-related cancer
- BRCA1/2 mutation probability tools: Claus Tables, BRCAPRO and Penn II
- From BRCA1/2 mutation probability to scoring systems
- Genetic counseling tools – comparison and mergers
- A categorical counseling test – the extended referral screening tool (ERST)
- Conclusions and further work

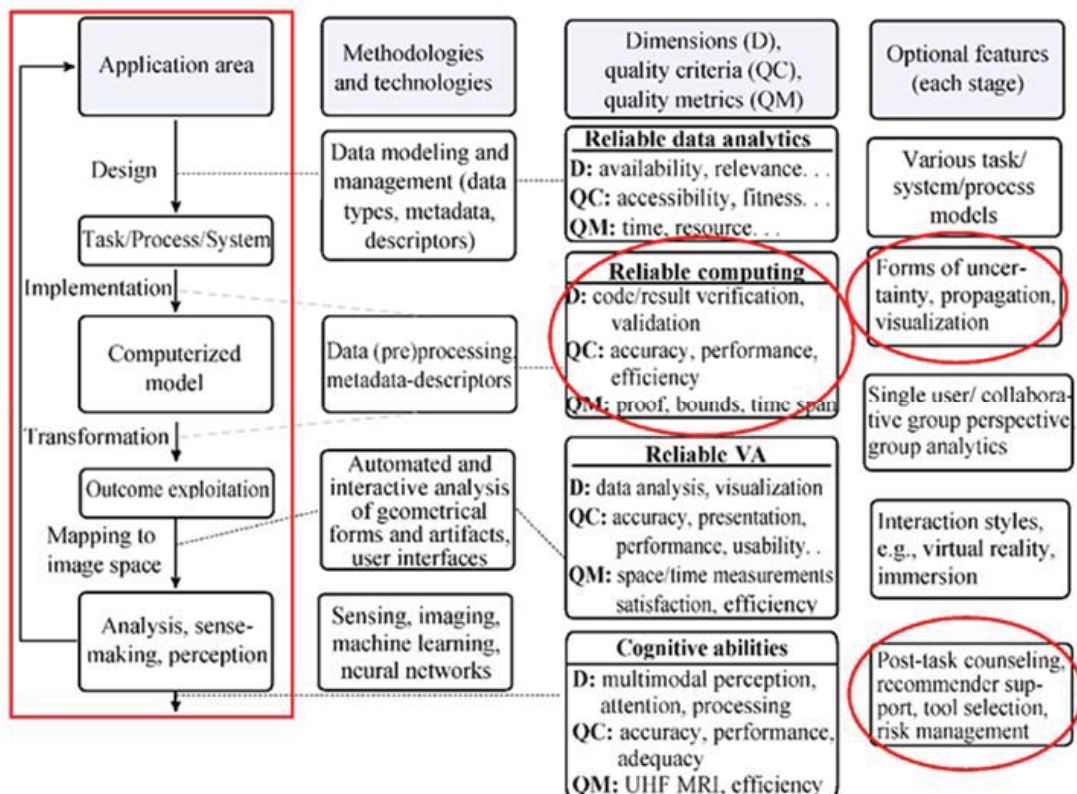
This talk deals with the verification and validation assessment of genetic tests and the accurate prediction of the probability of a BRCA1/2 mutation. Models for counseling patients based on the family history are described, with which help a secure risk assessment can be carried out, quality parameters and their measures introduced and the performance of the widely used tests compared in meta-studies. For this purpose, a new categorical test using binary interval decision trees (BDT) and variants is presented. It derives various risk classes from a person's family history and compares them with the cumulative and annual risks of breast and ovarian cancers for carriers and non-carriers of pathogenic BRCA1/2 mutations.

2 Verification & Validation Assessment for Process Outcome

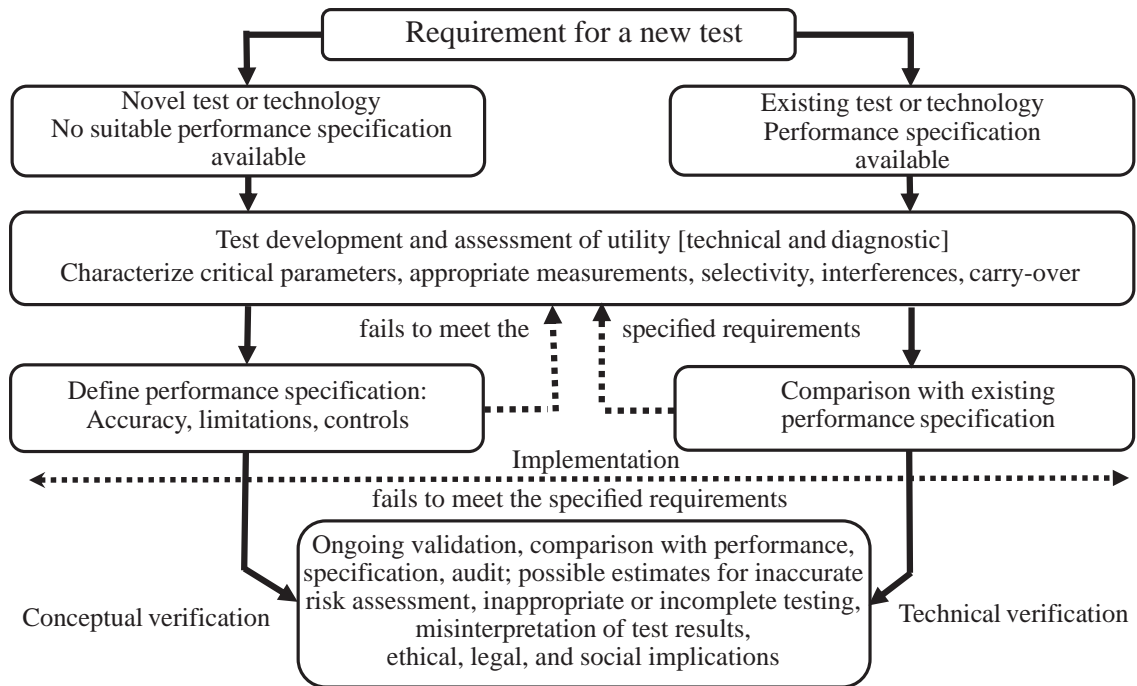
Complex real world processes use/produce huge, heterogeneous data and incorporate such various components as users; models, algorithms, implementations; outcome analytics, knowledge creation, risk assessment, and decision making. To obtain reliable results for any component or for the whole process, it is necessary to:

- Define assessment goals and apply a verification and validation (V&V) assessment to improve reliability.
- Implement a comprehensive quality management system to fulfill the requirements for internationally recognized standards.
- Introduce quality criteria depending on the task: accuracy, performance, efficiency, safety, usability...
- Develop measurements/metrics to establish the similarity degree (whether ground truth requirements are fulfilled for specific intended use).
- Assess (fused) data from different sources associated to variables and parameters including epistemic and aleatoric uncertainty appearing due to a lack of information/ knowledge, variability or ambiguity in visual perception and speech.
- Detect imperfections in the system design, process modeling and implementation using a validation cycle that highlights various sources of inaccurate and poorly calibrated risk predictions.

A Scheme for a Verification & Validation Approach to Assess an Environment from its Modeling to its Outcome Analysis Stage



3 V&V of Clinical Molecular Genetic Tests [20]



V&V assessment depends on the availability of a suitable performance specification

Definitions of Terms and Types of Tests

P/NPV: Positive/Negative Prediction Value **TP/TN/FP/FN:** True/False Positive/Negative
PPV = TP/(TP+FP) **NPV** = TN/(TN+FN)

Sensitivity = TP/(TP+FN) **Specificity** = TN/(TN+FP) **Accuracy** = (TP +TN)/Total

Positive likelihood ratio = sensitivity /(1-specificity): Chance of having the disease if the test is positive
Negative likelihood ratio = specificity /(1 - sensitivity)

ROC: Receiver-Operator Curve – A plot of true positive rate/sensitivity (y) vs. false positive rate for a range of test results (cdf of detection vs. false alarm probability). Describes predictive performance of models as an ascending curve.

AUC: Area Under Curve (i.e. ROC) – Demarcates between individuals at high and at low risk; AUC=0.5 means a test is not conclusive.

Types of Tests (Outcome)	Example	Quality Criteria		Validation
Quantitative (Numerical value)	Claus tables (Probability of BRCA mutation)	Accuracy; Precision	Performance; Repeatability	Specifications; Metrics
Categorical (A number out of a range)	Frank tables (Groups of patients)	Truth of the result	Efficiency Selectivity	Model and risk evaluation
Qualitative (Binary or n-ary)	Counseling genetic testing(Yes/No)	Sensitivity; Specificity; Accuracy	Group selection Consistency Limitations	Uncertainty modeling; ROC

4 BRCA-Related Cancer: Risk Assessment, Genetic Counseling and Testing

- BRCA1/2 mutations occur in 1 in 300–500 individuals in the general population and account for 5%–10% of breast and 15% of ovarian cancer.
- Specific BRCA1/2 mutations, known as founder mutations, are clustered among certain ethnic groups, for example, Ashkenazi Jews.
- Breast cancer **risk** increases to 45%–65% by age 70 years for persons having pathogenic mutations in either the BRCA1 or the BRCA2 gene; ovarian, fallopian tube, or peritoneal cancer risk increases to 39% for mutations in BRCA1 and 10%–17% in BRCA2.
- Population-based, case-control studies conducted by an authority are used to provide **age-specific risk estimates** of breast cancer for women with a family history of breast cancer.
- Data repositories include a number of patients with histologically confirmed breast cancer by age categories and control subjects matched to patients by geographic region and age intervals.
- A typical data set also includes family histories of breast cancer in first and second degree relatives of both patients and control subjects.
- **Genetic counseling** involves identifying and advising individuals at risk for inherited cancer susceptibility and is recommended before and after BRCA1/2 mutation testing.
- **Conclusion and Relevance:** “Among women without recently diagnosed BRCA1/2-related cancer, the benefits and harms of risk assessment, genetic counseling, and genetic testing to reduce cancer incidence and mortality have not been directly evaluated by current research.” (*cf. Nelson, H. D. et al. [22, p. 666]*)

5 A Meta-study and Evidence Report: Questions and Answers [21, 22]

1. In women with unknown BRCA1/2 mutation status, does risk assessment, genetic counseling, and genetic testing result in reduced incidence of BRCA1/2-related cancer and cause-specific and all-cause mortality? (Covered in 0 articles)
 - 2a) What is the accuracy and optimal ages/intervals of family risk assessment for BRCA1/2 related cancer performed by non-specialists in genetics in a clinical setting? (14)
 - 2b) What are benefits of pretest genetic counseling in determining eligibility for genetic testing for BRCA1/2-related cancer? (30)
 - 2c) What are optimal testing approaches to determine the presence of pathogenic BRCA1/2 mutations in women at increased risk for BRCA1/2 related cancer? (1)
 - 2d) What are post-test counseling approaches to interpret results and eligibility for interventions to reduce risks for BRCA1/2 related cancer? (0)
3. What are adverse effects for the risk assessment (0), pretest genetic counseling (30), genetic testing (22), post-test counseling (0) for BRCA1/2-related cancer?
4. Do interventions reduce the incidence and mortality in women at increased risk? (15)
5. What are adverse effects of interventions to reduce risk for BRCA1/2-related cancer? (28)

Type, scope, data sources, main outcomes, measures and results of the studies considered in [22] (overall of 103 studies with 110 articles and $n = 92\,712$ probands)

- 14 studies ($n = 43\,813$) showed that 8 considered **risk assessment tools** guided referrals to genetic counseling with moderate to high accuracy (AUC under ROC: 0.68-0.96).
- 28 studies ($n = 8\,060$) showed that **genetic counseling** reduced breast cancer worry, anxiety, and depression; increased understanding of risk; decreased intention for testing;
- 20 studies ($n = 4\,322$) showed that breast cancer worry and anxiety were higher after **testing** for women with positive results and lower for others; understanding of risk was higher after testing;
- In 8 randomized clinical trials ($n = 54\,651$), tamoxifen (RR relative risk 0.69 [95%CI, 0.59-0.84]; 4 trials), raloxifene (RR 0.44 [95%CI, 0.24-0.80]; 2 trials), and aromatase inhibitors (RR 0.45 [95%CI, 0.26-0.70]; 2 trials) were associated with lower risks of invasive breast cancer compared with placebo; **mastectomy** was associated with 90% to 100% reduction in breast cancer incidence;
- (6 studies; $n = 2\,546$) and 81% to 100% reduction in breast cancer mortality (1 study; $n = 639$); **oophorectomy** was associated with 69% to 100% reduction in ovarian cancer (2 studies; $n = 2\,108$); complications were common with mastectomy.

Our Observations

- Until now, approaches to validation of BRCA1/2 studies' results were focused on aleatory uncertainty only; epistemic uncertainty was considered indirectly through comparisons with other studies and meta-studies.
- Since different studies use different criteria for test persons, ages or degrees of kinship that often cannot be mapped to each other, it is necessary to work with sets while comparing such approaches as, for example, FHAT and MSS.
- Our suggested approach ERST uses interval arithmetic in combinations with decision trees to compute interval bounds for risk scores (rs) given by counseling tools and BRCA1/2 mutation probabilities (mp) taking care of missing or uncertain information.
- This approach can become problematic if the risk is overestimated since this can impair patients in their decision process so that strategies with no or minimal overestimation are necessary.

6 BRCA1/2 Mutation Probability Models/Tools

The following risk probability tables are provided in [10]:

- Claus and Frank Tables
- Predicted cumulative probability of breast cancer for a woman who has
 - ... one first/second-degree relative ... two first-degree relatives
 - ... mother and maternal/paternal aunt
 affected with breast cancer, by age of onset of the affected relative.
 - Predicted cumulative probability of breast cancer for a woman who has
 - ... one maternal and one paternal second-degree relative
 - ... two second-degree relatives (both maternal or both paternal)
 affected with breast cancer, by age of onset of the affected relatives.
 - Mutations in BRCA1/2 correlated with age of diagnosis, personal and family history of cancer in (non)-Ashkenazi individuals are found in [14]

Based on Mendelian genetics and Bayes' theorem, this mathematical model provides the probability for a woman with a family history of breast and/or ovarian cancer to carry a mutation of BRCA1. The model takes into account

- BRCAPRO [5,12]
- Relationships of all affected and unaffected 1st and 2nd degree relatives;
 - Current ages at diagnosis;
 - BRCA1 mutation frequencies in the general population;
 - Age-specific incidence rates of breast and ovarian cancers in carriers and non-carriers of mutations.

This empirical risk model for individuals and families is validated using families with multiple cases of breast and/or ovarian cancer.

- Penn II [25]
- Suppose that a family pedigree had two or more individuals affected with either breast and/or ovarian cancer; one of them must be a 1st, 2nd or 3rd degree relative of the other;
 - The model provides a prior probability for individual and family of BRCA1 or BRCA2 pathogenic variant; <https://pennmodel2.pmacs.upenn.edu/penn2/>

7 From BRCA1/2 Mutation Probability to Scoring Systems (FHAT)

Family History Assessment Tool (FHAT) was developed by Gilpin et al. [15]:

184 model patients with breast or ovarian cancer (BC/OC) ready for BRCA1/2 testing answered a questionnaire on BC; OC; bilateral BC; BC and OC in the same person; male BC; colon and prostate cancer in proband's (PR) ancestors 1st, 2nd, and 3rd degree relatives; birth year and history of cancer, age of diagnosis. Individual scores (1-10) are combined; the final score ≥ 10 indicates referral (doubling the lifetime risk for BC, 22%).

Diagnosis – The proband has the role of a child in FH (criteria from [14])	Modeled mutation probability in BRCA1 (left) and BRCA2 (right)		FHAT
Any Relative with BC (ARBC) < 50y	10.1	14.5	4-10
Any Relative with OC (AROC)	22.9	12.5	5-13
ARBC<50y and the proband with BC (PrBC) < 40y	28.2	11.6	11-19
PrBC < 40y and ARBC < 50y and AROC	50.9	7.9	16-32
(Pr Bilateral BC or OC) and ARBC<50y and AROC	65.0	5.7	15-35
(Pr Bilat BC<40y or OC) and ARBC<50y and AROC	86.7	2.2	19-35

8 Genetic Counseling Tools – Comparison and Mergers

Ontario Family History Assessment Tool

FHAT Gilpin et al. [15]

Manchester Scoring System MSS

2004 [12]

1st dr: mother (m),
parent (p), sibling (s)
2nd dr: grandparent,
grandchild, aunt,

				2 nd , 3 rd / 2 nd degree relative (dr) +		
				5/4 < 30y		
				3 30-39y		
				2 40-49y		
				1 50-59y		
					+	∴ BRCA1/2
Breast	2	p4 s3	1/1	<40y	6	3/0
Bilateral/multifocal	+3		x2	40-60y	4	3/0
Male	+2		5/5 5/8	>60y	2	
Ovarian	3	m7 s4				5/5
B&O	5	m10 s7				BRC1 tested
Colon	+1		+1 If breast, ovarian, colon or			0
Prostate	+1		+1 if prostate cancer and < 50y			0/1
Pancreas	+0/1	<60y	Only the lineage providing the highest score was counted			0/1

Example: A family where the mother (*m*) was diagnosed with breast cancer at age 46 years and a maternal aunt (*a*) was diagnosed with ovarian cancer at age 54 years.

Description: (BC ≤ 50y) (BC *m* 46y: first degree relative/all degree relative 40-49y) and OC (OC *a* 54y second degree relative 40-60y/all degree relative < 60y).

Recommendation: All mergers recommend screening with **FHAT** score 6+7=13 and **MSS** score 3+8=11 for BRCA1 and 3+5=8 for BRCA2 if BRCA1 is tested. (22% is a conceptual limit for recommending tests).

Scoring:	FHAT <i>m</i>	MSS <i>m</i>	FHAT <i>a</i>	MSS <i>a</i>
	[6,10]	[3,6]	[5,9]	[5,8]
Score in example	6	3	7	BRCA1 8/ BRCA2 5
Min cut		6		5
Max cut		6		8
Mean value		4.5		7.5
Mean value BRCA2		4.5		6

For **RST**, two or more checks are needed in the table with the following entries: 2 BC ≥ 50y on the same side of the family; BC ≤ 50y; OC; male breast cancer; Ashkenazi Jewish ancestry. Since BC ≤ 50 and OC age onset information is not fully used in RST, the interval should include the worst case when FHAT and MSS are applied

Probabilities for genetic mutations are given by Penn II and Frank tables; Frank tables are more pessimistic.

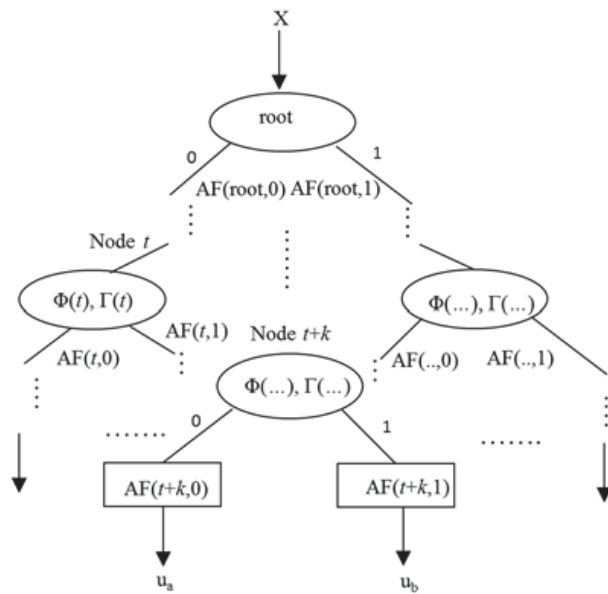
Penn II: Individual(left) and family (right) mutation prediction result

Risk of BRCA1 Mutation 10 [4] % 20 [8] % ([Non-]Ashkenazi Jewish in FH)

Risk of BRCA2 Mutation 5 [3] % 9 [5] %

Frank (2002) [14]: 65/211 (30.8%) [58/354 (16.4%)] (prevalence of mutations in BRCA1 and BRCA2 correlated with personal and family history of cancer in 2 233 [4 716] [Non]-Ashkenazi individuals)

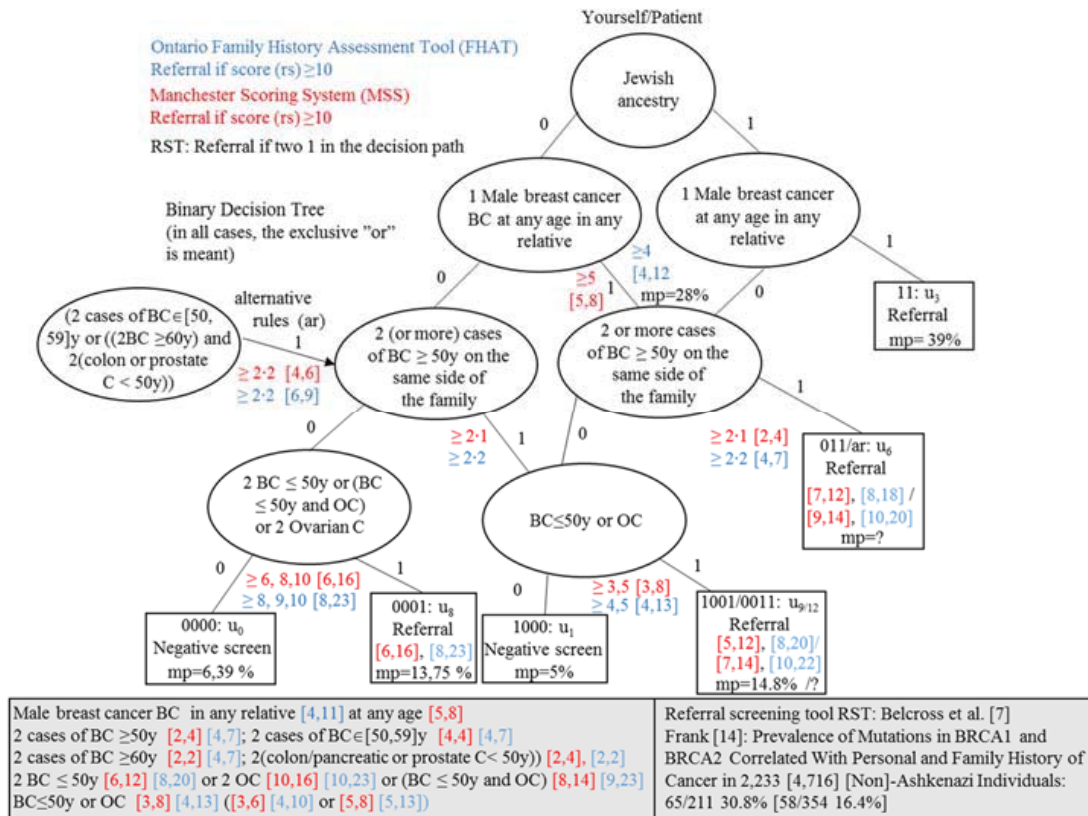
9 Multi-Criteria Binary Decision Trees–Categorical RST (ERST) [29]



X denotes an input data vector, U a decision/referral vector with indices built from a sequence of 0/1 (no/ yes) decisions from root to leaf. $\Phi(t)$ refers to the set of interval valued features used by the BDT applied to node t , $\Gamma(t)$ denotes the decision rule as function of the features/conditions with values $\{0,1\}$ and accumulated interval risk function $AF(t,.)$. The risk function can be comprised of the scores (rs) or cumulative probabilities (mp) If the decision rules of an inner node are replaced or completed [29], $AF(t,.)$ needs to be adapted accordingly.

The resulting vector $U=(u_0, \dots, u_{n-1})$ has components u_a storing decisions in the form (interval rs, (mp)%) where the index a is the decimal representation of the binary decision path to the leaf, read from left to right.

10 ERST: Merging RST decision rules with FHAT/MSS risk assessments and BRCA1/2 mutation probabilities



11 Conclusions and Further Work

To summarize, it is our opinion that requirements for genetic risk evaluation in the counseling and testing process of individuals and families affected by breast and ovarian cancer caused by pathogenic mutations should address the following key points:

- Each involved discipline should recognize the need for standardized evaluation and V&V of processes and their models as well as for fusion of conclusions from their outcome. For this, appropriate quality criteria and their metrics as well as standardized procedures need to be used.
- The collected heterogeneous data should be standardized with a focus on the test participants' numbers, origin and age. Additionally, the data should be assessed and if necessary completed to make the application of the specified quality criteria and metrics possible. Although the number of subjects in the studies has increased considerably over the years, so has the number of examined parameters, which the test results strongly depend on, with the additional data coming from women of varying socioeconomic and ethnic groups. Currently, comparability and calibration in meta-studies is not always possible in a satisfactory manner due to the heterogeneity and varying availability and quality of the data, tests and conducted studies. Aside from Nelson [21, 22], most cited meta-studies are based on a relatively small samples that are not representative of the age groups, regions of origin and type of pathologies as described in the ClinGen [11] clinical validity framework.
- (Bounded) uncertainty in the data should be represented by suitable data types and propagated through the models and processes. Only after that the results can be judiciously evaluated according to standardized criteria and the recommendations of the experts feasibly merged into proposals for risk management strategies that are understandable and not harmful for the patients [26].
- Unfortunately, no articles were included in Nelson's meta-studies that dealt with the consequences of risk assessment and post-test counseling. Young women with BRCA1/2 mutations and their families face conflictive healthcare decisions regarding family formation and risk management. They must decide whether and when to prioritize risk reducing interventions or to pursue family formation goals. Peshkin [26] addresses these questions and gives an up-to-date overview on genetic testing and management of individuals at risk of hereditary breast and ovarian cancer syndromes documented with 115 references highlighting the state of the art.
- Finally, effects of the above mentioned strategies need to be assessed and, subsequently, referral and tests adapted accordingly.

Explaining and visualizing recurrent neural network decisions

David Qaramyan¹ and Hrant Khachatryan^{1,2}

¹ Yerevan State University, Yerevan, 0025, Armenia

² YerevaNN, Yerevan, 0025, Armenia

Abstract. Neural Networks can approximate any complex function, so they work very well in many disciplines such as computer vision, natural language processing, etc. Despite their performance, it is unclear, how neural networks incorporate with input features and make decisions. In this work an attempt was made to shed light on the work of neural networks or, in other words, to visualize their work. The study recruited a recurrent neural network, which tries to predict the probability of death for a given clinical data of the patient in the resuscitation department. In this paper we adapted two methods developed for image recognition and natural language processing for neural networks based on clinical data and visualize the work of these networks.

Keywords: Recurrent neural networks · Gradient based visualization · Layer wise relevance propagation · MIMIC-III · Clinical Time Series

1 Visualization methods

To explain and understand the decision-making process of neural networks, we will consider two types of visualization techniques, gradient based visualization and layer-wise relevance propagation.

Gradient based visualization. This technique computes a class saliency map (M), specific to a given input features (I_0) and class (S_c). This achieves by computing the gradient of the class score (S_c) with respect to the inputs (I). Signal is back propagated from output layers to each intermediate layer and finally to the original input features [2, 3, 5].

$$M \simeq |\nabla_I(S_c)|_{I_0} \quad (1)$$

Layer-wise relevance propagation. Layer-wise Relevance Propagation (LRP) [4] operates by propagating the prediction score ($f(x)$) backward in the network, using a set of purposely designed propagation rules. It defines relevance score ($R_d^{(l)}$) for each neuron in net and propagation rules which must satisfy conservation law according to which the relevance score is preserved by back passing from one layer to another.

$$f(x) = \dots = \sum_d R_d^{(l+1)} = \sum_d R_d^{(l)} = \dots = \sum_d R_d^{(1)} \quad (2)$$

In contrast to feed-forward nets, visualization of recurrent neural nets (RNN) needs specific propagation rules which will allow to handle multiplicative connections as they arise in recurrent network architectures such as LSTMs and GRUs. We refer to the [6] in which all propagation rules are described.

2 Visualized model

Work [1] addresses four key tasks based on MIMIC-III data [7]. However, in this article we will focus on only one problem: in-hospital mortality (predicting mortality probability in resuscitation department). The following is a brief description of Channel-wise LSTM, on which the visualization methods were applied. Unlike standard LSTM, which works directly with sequential input data, Channel-wise LSTM independently process different group axes of input data with bidirectional LSTMs, then all Bi-LSTM outputs are concatenated together and fed to another LSTM network as input.

3 Visualization results

Consider the patient clinical measurements for which the network made the correct prediction (see Fig. 1) (True Positive - the patient actually died). Corresponding heatmpas are depicted in Fig. 2 and Fig. 3.

In both illustrations, great attention was placed on descriptions such as Glasgow coma scale eye opening, Glasgow coma scale motor response, Glasgow coma scale total, Glasgow coma scale verbal response (see black rectangle) which characterize the state of human consciousness and directly related to mortality. In the last few hours, the above-mentioned descriptions have changed dramatically (see black rectangle), based on which the Channel-wise LSTM network was able to make the right decision.

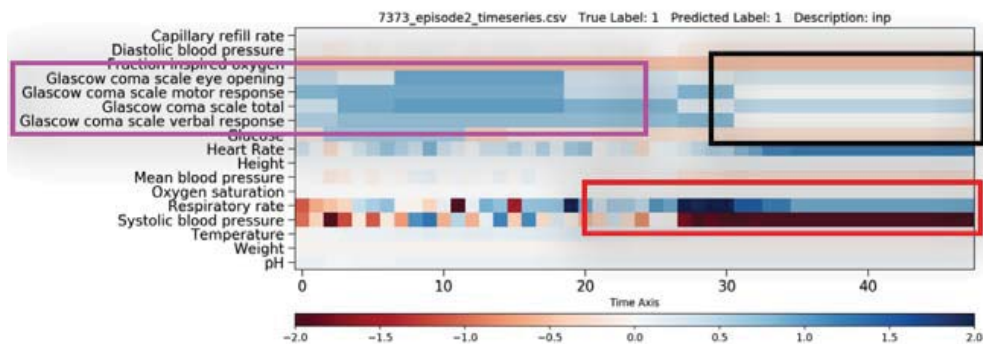


Fig. 1. Visualization of input features. The horizontal axis describes the time: from 0 to 47. Each row represents some health condition measurements which are indicated on the left.

In contrast to the gradient visualization, which mainly looks at the measurements of the last hours (see black and red rectangles), the visualization obtained

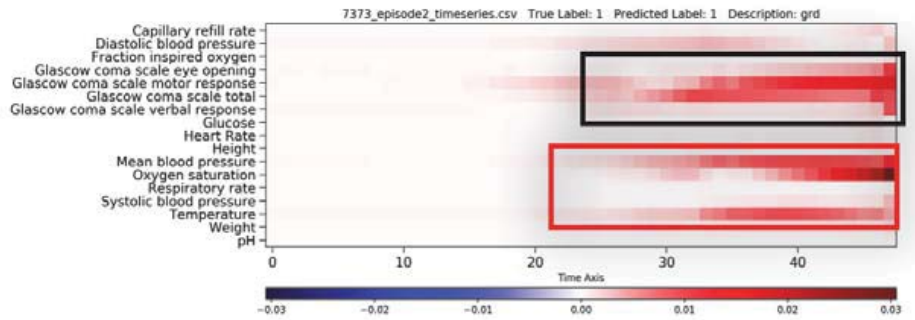


Fig. 2. Gradient-based visualization.

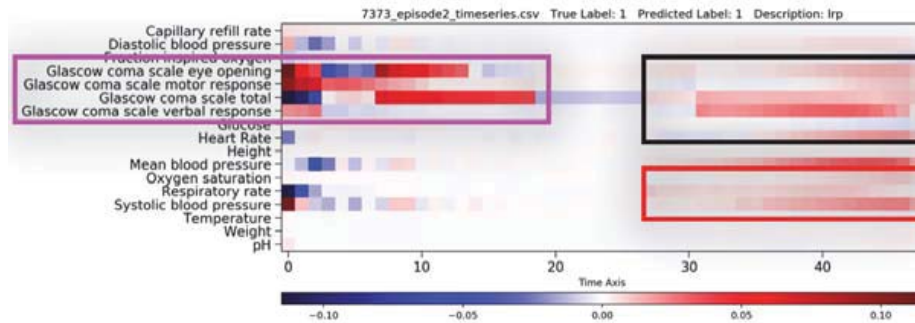


Fig. 3. LRP visualization.

by LRP draws attention to both the initial and final measurements of the above-mentioned descriptions (see black and pink rectangles) which may be caused to fact that gradient has much stronger signal in last timestamps rather than in first ones.

References

1. Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. “Multitask Learning and Benchmarking with Clinical Time Series Data”. arXiv:1703.07771
2. D. Erhan, Y. Bengio, A. Courville, and P. Vincent. “Visualizing higher-layer features of a deep network”. Technical Report 1341, University of Montreal, Jun 2009.
3. Karen Simonyan, Andrea Vedaldi, Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR, 2014
4. Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, Wojciech Samek. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”.
5. Li, Jiwei and Chen, Xinlei and Hovy, Eduard and Jurafsky, Dan .”Visualizing and understanding neural models in NLP”, Proceedings of NAACL-HLT 2016, San Diego, California, June 12-17, 2016. c 2016 Association for Computational Linguistics
6. L. Arras, G. Montavon, K. Müller, W. Samek. “Explaining recurrent neural network predictions in sentiment analysis”, Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, 8 September, 2017 (2017), pp. 159-168
7. Johnson, A. E. et al. “Mimic-III, a freely accessible critical care database. Sci. Data 3 (2016)”

Revisiting the Promotion Effectiveness Measurement in Retail

Nelson Baloian¹, Jonathan Frez², Cristóbal Fuenzalida¹, Belisario Panay¹, Sergio Peñafiel¹, José A. Pino¹, Horacio Sanson³

¹ Department of Computer Science, University of Chile, Santiago, Chile
{nbaloian, crfuenza, bpanay, spenafie, jpino}@dcc.uchile.cl

² Diego Portales University, Santiago, Chile
jonathan.frez@inf.udp.cl

³ Allm Inc. Tokyo, Japan, horacio@allm.net

Abstract. Retail store managers need to measure the effect of promotions have in terms of the number of people visiting a store and sales in order to make decisions for an efficient use of resources. The state of the art approach to this problem is the authors propose to measure the effects of a promotion is to develop a predictor for the number of people entering the store, in conditions of absence of any promotion campaigns and then compare the output against actual data and measure the difference. This approach shows some drawbacks because stores have permanently some kind of promotion running, which makes the training of a model with the required data almost impossible. This work proposes to use a different methodology to the one mentioned above which is based on first identifying periods of time with abnormal behavior and then examining back to check the promotions being applied during that time. For this purpose, we first developed an accurate predictor and then we compute a function which will compare the difference of the predicted value with the real value, against the same indicator but for the whole set of retail stores which are considered to be of the same market type. This work presents the results obtained and examples of how can this be used to compare a store behavior compared to the rest of the market in order to search for singularities.

1 Introduction

Measuring the effect of promotions in terms of the number of people visiting a store and/or the increase in the number and final amount of sales is an important task in the retail management activity. Managers need to know that information in order to make decisions about the kind of promotion to apply to make efficient use of resources. This use not only concerns the actual cost of applying a promotion but also the previous needed preparations with goods and sales force. Consequently, this problem has been studied extensively in the market literature by many authors in the past, as early as the 80's [1] and nowadays [2], [3], [4].

In one of the most recent publications on the subject [5], the authors propose to measure the effects of a promotion on a certain store on the number of people visiting that store by the following method; first, they propose to develop an accurate predictor

for the foot traffic, i.e., the number of people entering the store, in conditions of absence of any promotion campaign and then compare the output against real data for the number of people who entered the store for the period of time when the promotion was held.

However, stores present an increasing trend to have permanently some kind of promotion running on any season. In fact, already in 2003 Steenkamp et al. [6] report that 24% of all purchases in Dutch supermarkets take place under some form of promotional support. Comparable numbers are observed in the United Kingdom and Spain, while in the United States, this number approaches 40%. This makes the previous approach difficult to apply or almost impracticable, because there is no data available on the number of people entering a store under a condition of absence of any promotion. Additionally, some promotions are targeted not to increase the number of people visiting the store but to increase the amount each person buys, or the number of articles each person buys (e.g., get 2 for the price of 1). Moreover, there might be actions taken by other entities, from which the store in question might have little information, which may also influence the number of people entering the store. This makes impossible to develop a model which can be trained with the required data, in order to compare the output to the actual numbers.

This work proposes to use a different methodology to the one mentioned above which is based on first identifying periods of time with abnormal behavior in the number of people entering the store (called foot traffic), the percentage of people actually buying (called conversion) and the mean total value for each purchase (called mean ticket value) and then examining back to check the promotions being applied during that time. For this purpose, we first developed an accurate predictor for the foot traffic, the conversion rate and the ticket value, and then applying a function which will compare the difference of the predicted value with the real value, against the same indicator but for the whole set of retail stores which are considered to be of the same market type.

2 Available Data and Data Embedding

The data for the number of people entering stores is obtained using cameras which are placed at the entrances of retail stores. The camera registers the number of customers' entrances and exits every hour in a relational database using a single table. Each entry is associated with a timestamp and a unique ID-number of the store. Additionally, we added to this table the data for the number of sales and the amount of all sales for a certain hour which are easily obtained from the records of the checkout machines of the store.

The algorithm used to make the predictions solves a regression problem, i.e., it tries to predict a continuous value (in this case, these are the number of entries, number of tickets and total amount of sales) from an input vector. This algorithm is a supervised learning model based on a modified version of the theory of evidence also called the Dempster-Shafer theory [7].

When making a prediction of an input vector, the algorithm uses the information from the vectors that were previously provided to it, usually the training data set, as evidence. The algorithm uses a strategy of closest K-Neighbors, where the K vectors are obtained with the shortest distance at which you want to predict. Then, the similarity of the vector to be predicted with the neighboring K obtained through the use of a

weighted Euclidean distance is calculated; this is a distance in which each element of the vector has a weight that represents its importance in the similarity calculation. The obtained similarity value represents the probability that the input vector is one of the values in the evidence set. Finally, with this probability, an expected value of the input vector is calculated, giving a value for the prediction. In addition to this value, the model is capable of predicting a confidence interval for the prediction, i.e., an upper and a lower limit for the prediction.

One of the advantages of this algorithm over other supervised training models is its interpretive power. Many algorithms being used in current state of the art are “black boxes” in which end users cannot obtain relevant information on how a prediction is made. Since this algorithm uses a K-Neighbors strategy and dimension weights for the calculation of similarity, we can easily obtain an explanation behind a prediction, which is valuable for further analysis. The used embedding considers the following characteristics for a sample:

- Day of the month
- Time of the day
- Weekday
- Number of tickets in the time block one week before
- It is a holiday
- It is a special date
- It is the week before a special date

In the case of the variables: Time of the day, Weekday, and Day of the month, codifying them as increasing numbers is not optimal because it does not reflect the cycle correctly. For example, if the day of the week Sunday is coded as 0 and Saturday is coded as 6, the distance (numerical) between these two days is too high, given the fact they are two contiguous days. To avoid this problem, it has been proposed to use circular codifications for these cyclical variables. This implies that we increase the dimensionality of the vector as long as these days are at a distance equivalent to any other pair of contiguous days.

To test the model performance, we propose to use five accuracy metrics for regression problems. The metrics we use as indicators for the precision of the predictions are:

- Mean Square Error (MSE): Average of the square of the distance between the predictions and the actual values
- Mean Absolute Error (MAE): Average of the absolute value of the difference between the prediction and the real values
- Determination coefficient (R²): Proportion of variance of the real values that is determined by the predicted values. It is related to the correlation between actual and predicted values.
- Relative Mean Error (ERM): Percentage of absolute error of the predictions with respect to the mean of the data. It corresponds to the Average Absolute Error divided into the mean of the data. Also, the relative precision is used as 1 minus this value.
- Cumulative Percent Error (TSM): It corresponds to the sum of the predicted values in a prediction interval divided by the sum of the real values in the same period

For R2 and ERM metrics, a value of 1 or 100% corresponds to a perfect prediction. Concerning the rest of the indicators (MSE, MAE, and TSM): a value of 0 corresponds to a perfect prediction.

The available data used for this work was supplied by a company which offers business intelligence solutions to retail stores in Santiago, Chile, mainly located in large shopping malls. We got almost 4.5 years' data for 27 stores starting from beginning of 2016 until April 2020.

3 Predictions

The first prediction we made was the foot traffic per hour. Using the embedding described above and a training data set which contained records from 3 years before the predicted time slot which spans from March 8th 2019 to May 8th 2019 (three months in total), we generated for each store the necessary data to produce graphs like the one shown in Figure 1.

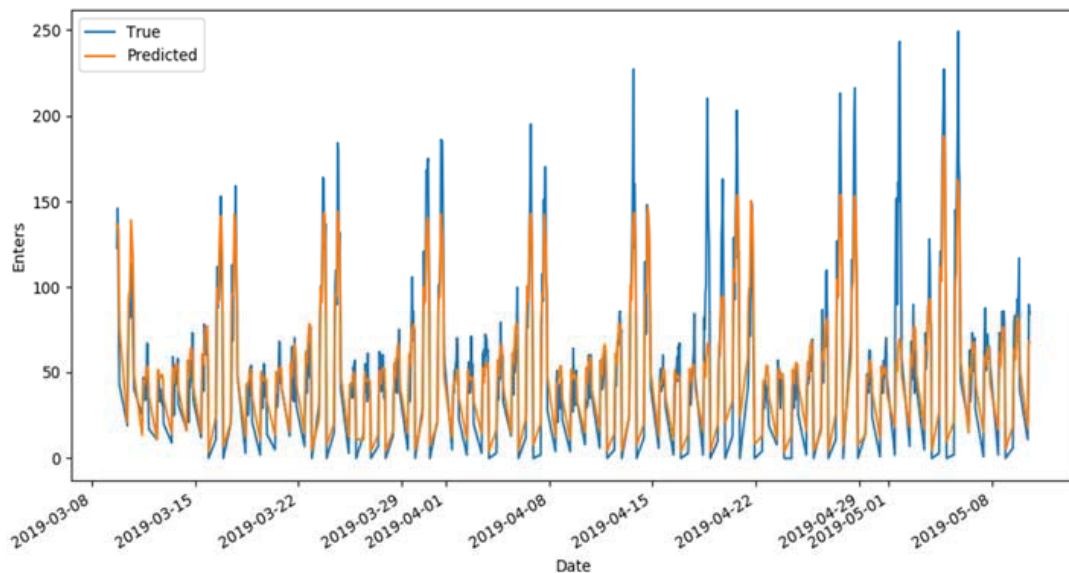


Fig. 1. Example of prediction result for a random store using 3 years of data for training compared with the actual numbers

As we can see in figure 1, the predicted curve closely resembles reality. In a few cases the algorithm cannot fully predict some input peaks, but even so the precision metrics are quite good, obtaining a relative precision of 78% (being 100% the maximum) and a coefficient of determination of 0.76 (being 1.0 the maximum). In addition to this particular case, a massive test was carried out with 27 stores belonging to a shopping mall in the city of Santiago, Chile, in order to obtain values for the general trend of the prediction for a larger number of stores.

Table 1. Mean values for the metrics obtained for the prediction applied to the foot traffic (entrances in one hour) for a sample of 27 stores

Indicator	Mean Value
R2	0.55
Relative Precision	70.87%
MSE	825.76
MAE	17.52
TSM	12.97%

From this table we can conclude that the model manages to predict the inputs with great precision so it can be used as a valid predictor of this variable.

The next variable to predict is the number of sale tickets, i.e., the number of sale transactions made in a certain period of time regardless of the amount of each sale. In a first approach we also did this prediction for an hour's time period. However, unlike the number of entries, there are many hours in which the number of tickets is 0, which artificially worsens the prediction since the model would seldom predict such a value, thus increasing the mean error considerably. We opted instead to group results for a one-day time slot.

Figure 2 depicts the predicted and real values grouped by day for a random store, which obtains an R2 of 0.60 and a relative error of 73%. Table 2 shows the mean results of the metrics for 27 stores.

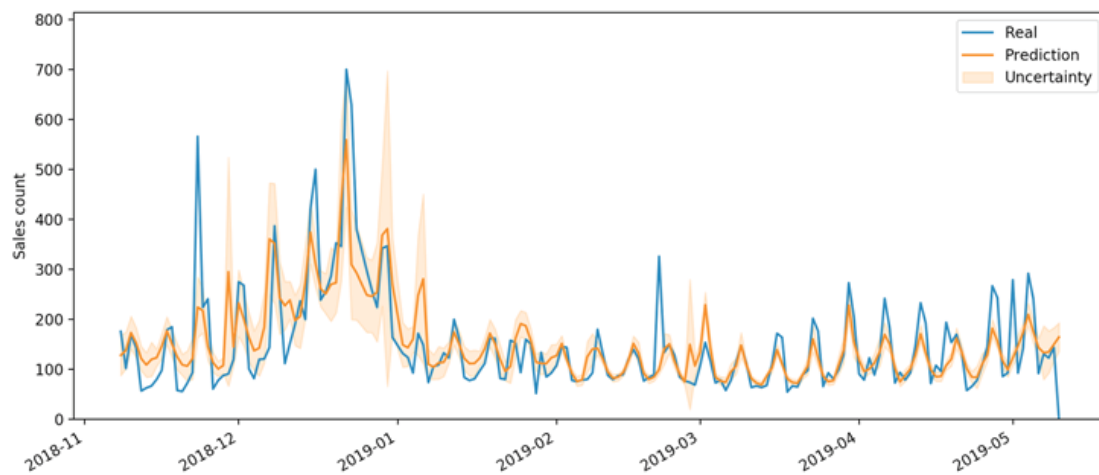


Fig. 2. Example of a prediction for a random store of the number of tickets using one day intervals compared with the actual numbers.

Table 2. Mean values for the metrics obtained for the prediction applied to the number of tickets using one day intervals for a sample of 27 stores

Indicator	Mean Value
R2	0.4064
Relative Precision	59.42%
MSE	12.98
MAE	2.577
TSM	7.00%

The last variable to predict corresponds to the amount of sales. This variable is defined as the total amount of sales generated by all customers in a certain period of time. This variable is also measured for a whole day for the same reasons considered in the previous variable. Figure 3 shows a graph comparing the predicted number for the daily amount of sales compared with the actual number, obtaining an R2 of 0.23 and a relative precision of 69%. Table 3 shows the average of the results of the indicators for the sample of 27 stores.

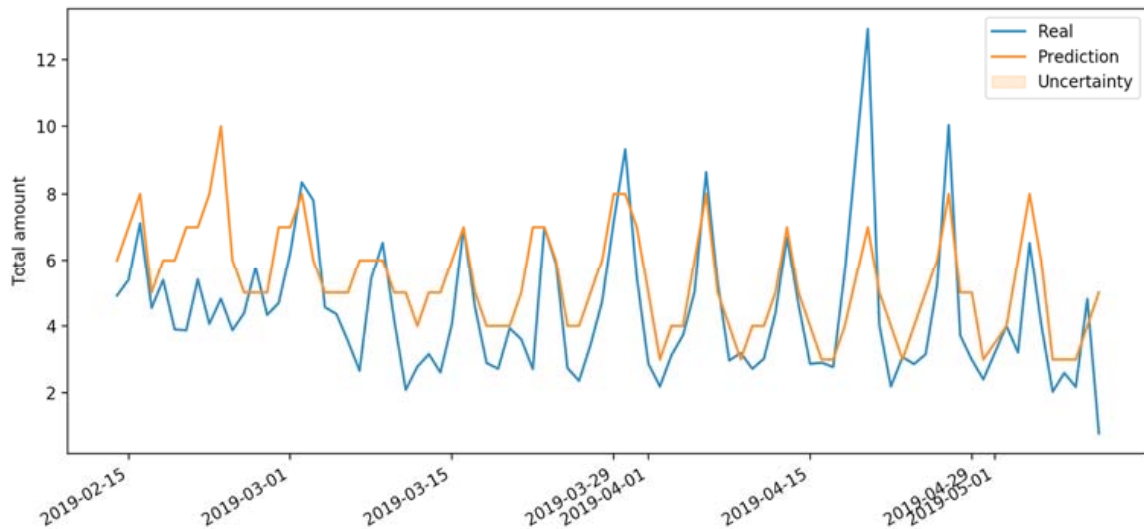


Fig. 3. Example of a prediction for daily total amount sales for a random store compared with the actual numbers.

Table 3. Mean values for the metrics obtained for the prediction applied to the amount of sales using one day intervals for a sample of 27 stores

Indicator	Mean Value
R2	0.2945
Relative Precision	65.89%
MSE	5.520
MAE	1.384
TSM	13.7%

From the analysis of these three prediction cases, we can see that the model performs better in the foot-traffic prediction which reaches the highest value of R2 and Relative Precision. For the case of the number of tickets and the amount of sales, both have a low error in the TSM metric which implies that analyzing the accumulated predictions (i.e. aggregating by a week) the error decreases and the prediction is more accurate.

4 Identifying Singularities

As we said, our approach to measure the effects of promotions in the sales will be identifying periods of time with abnormal behavior in the number of people entering the store (called foot traffic), the percentage of people actually buying (called conversion) and the mean total value for each buy (called mean ticket value).

In this document we explain our approach using the number of entries as example, but our proposal is to do the same with the number of sales and amount of sales per day. We do this computation for daily numbers of entries since promotions usually last for at least one whole day.

Calling *entersPred* and *entersReal* to the number of predicted and actual entries per hour, we compute for each retail store j ep_{ij} and er_{ij} which corresponds to the sum of entries from day $i-w$ until day i for the store j , (w is the size of the window which should be set according to the length of the typical promotion in force).

$$ep_{ij} = \sum_{i-w}^i entersPred_{ij}, er_{ij} = \sum_{i-w}^i entersReal_{ij}$$

After this, we compute the difference between the actual entries and the predicted ones, ($Delta_{ij}$),

$$Delta_{ij} = (er_{ij} - ep_{ij})$$

Then, we define $DeltaMarket_i$ as the mean value of all the differences computed for a set of stores for a single day i . The stores which are included in the set should be previously defined considering the stores that define the “market” against which we want to consider the behavior of a store. Typically, these are the stores that sell similar products, have a similar size, are in a same neighborhood, etc.

$$DeltaMarket_i = \frac{\sum_{j=1}^n Delta_{ij}}{n}$$

Finally, we compute for each day i for a store j the difference between $Delta_{ij}$ and $DeltaMarket_i$, being this the metric we use to measure if there is a deviation in the client’s behavior of this store against the market.

$$Rel_{ij} = \frac{(Delta_{ij} - DeltaMarket_i)}{DeltaMarket_i}$$

Now we will analyze some examples of the application of this method in order to illustrate how this method helps to find singularities.

Figure 4 depicts the Rel_{ij} values for three stores for a period of time starting on February 21 and ending on March 25. The 0 line represents the behavior of the whole market (in this case, the behavior of the 27 stores for which we had data). Store A corresponds to a bookstore whose high performance can be explained by the start of the school year which in Chile takes place in March. However, on March 20 bookstores were closed due to the outbreak of the coronavirus pandemia. B corresponds to a shoe-store, which shows a similar behavior to the market (close to baseline 0), but was not closed on March 20, so its performance increased compare to the rest of the market. Finally store C sells bedroom items, mostly bedding accessories like linen, cushions, etc. Its behavior during the summer season is lower than the market and falls sharply at the beginning of the closings of shopping centers.

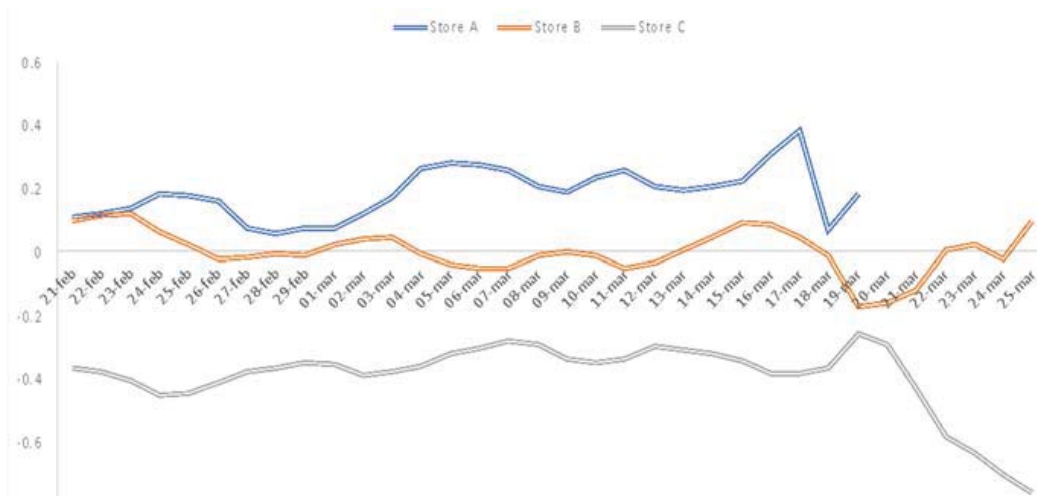


Fig. 4. Example of values of Rel_{ij} for three stores.

Figure 5 exemplifies another analysis which can be performed with this method. In this graph we add two lines, one above the zero line and another below, at a distance of the standard deviation of the market. Here the blue line represents the difference between the performances of the stores of that chain against the whole market. It shows that the performance was quite similar. However, when we compare the performance of the stores of that chain against the performance of only the rest of the shown stores (yellow line) we see that the store had a lower behavior compared to its competitors regarding the number of people who visited the stores.

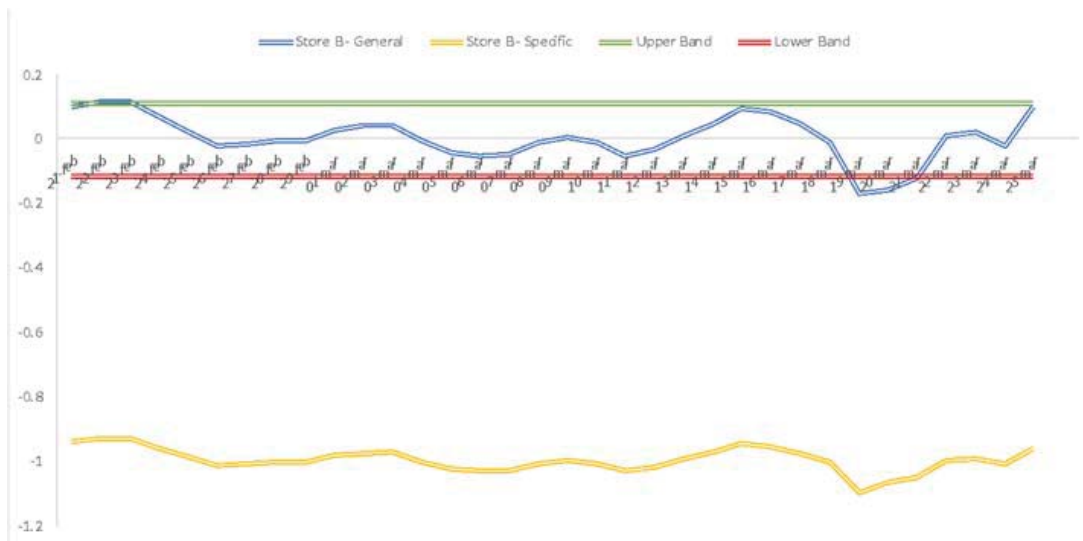


Fig. 5. Comparing the behavior of all stores belonging to a certain shoestore chain against the whole market (blue line) and the same chain against the stores which sell shoes (yellow line).

5 Conclusions

In this work we presented a novel approach to analyze the impact of promotions on the behavior of customers in retail stores regarding the number of people entering a store each hour, the number of daily sales and the total amount of sales in a day. A previous proposed approach suggests to develop a prediction model based on past data and compare the predicted number against the real ones, assuming that the model is trained with data generated on days without promotions. This approach is almost impossible to apply since stores usually apply promotions almost every day. So it is necessary to find another baseline to compare the predicted data to find instances of abnormal behavior. We propose to compare the differences of the predicted values and the real values for a whole market against the difference of the predicted and actual data of people of a store. For this purpose, we developed three predictors, one for the number of people entering a store, one for the number of sales and one for the daily total amount of sales. In order to train and test the model we used data of 27 stores in Santiago, Chile provided by a company which offers business intelligence services to retail stores. The predictions performance was quite good according to various metrics. Finally, we presented

a formula to calculate deviations of the predicted values from the actual ones compared with the rest of the market and examples of graphs which can help to identify them. Large deviations may trigger alerts to managers.

Acknowledgements

We would like to thank the support from the FollowUp Company (Chile) and the contributions from Isabel Cumplido and Fabián Ramírez.

References

1. Blattberg R. C., Levin, A.: Modelling the effectiveness and profitability of trade promotions. *Marketing Science* 6(2) (1987) pp. 124-46.
2. Goyal, P.: Measures to improve sales promotion effectiveness: The consumer perspective. *Pranjana: The Journal of Management Awareness* 22(1) (2019) pp.54-67.
3. Gedenk, K.: Retailer promotions. In *Handbook of Research on Retailing*. Edward Elgar Publishing, 2018.
4. Dekimpe, M. G., Hanssens, D. M., Nijs, V. R. and Steenkamp, J. B. E.: Measuring short-and long-run promotional effectiveness on scanner data using persistence modelling. *Applied Stochastic Models in Business and Industry*, 21(4-5) (2005) pp.409-416.
5. Epstein, L. D., Flores, A. A., Goodstein, R. C., and Milberg, S. J.: A new approach to measuring retail promotion effectiveness: A case of store traffic. *Journal of Business Research* 69 (10) (2016) pp. 4394-4402.
6. Gijsenberg, M. J., and Nijs, V. R.: Advertising spending patterns and competitor impact. *International Journal of Research in Marketing* 36.2 (2019) pp. 232-250.
7. Peñafiel, S., Baloian, N., Sanson, H., and Pino, J. A.: Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier. *Expert Systems with Applications* 148 (2020) 113262.

Enriching Word Vectors with Morphological Information

Martin Mirakyan¹ and Hrant Khachatryan²

¹ YerevaNN mirakyanmartin@gmail.com

² YerevaNN hrant@yerevann.com

Abstract. This paper presents an end-to-end approach for word representation learning which takes into account the morphology of the language. The system consists of three parts: semantic analysis of a sentence, morpheme extraction from each word, and word-vector learning. The novelty of our approach is the linguistically correct morphological word features and the end-to-end pipeline for learning word vectors. Our method achieves state of the art performance on morpheme segmentation, while outperforms most of the solutions for lemmatization, part of speech (POS) tagging, and morphological feature extraction. Finally, we evaluate our approach on the obtained word embeddings and demonstrate that linguistically correct word features can lead to better word representations especially for rare words.

Keywords: Morphology · Word embeddings · Word vectors

1 Introduction

Most of the modern NLP systems use word vectors as part of their pipelines for obtaining accurate results in machine translation, semantic analysis, question answering, etc. Those word vectors are expected to have close representations for similar words. In some approaches, similarity refers to semantic similarity (e.g. running is close to jogging, or sad is close to depressed, etc.). In other works, similarity should also represent the morphological aspect of the language (e.g., walking should be close to walked, simple should be close to simplistic, etc.). Current attempts to extract word embeddings like fastText [3] mostly rely on subword information instead of explicitly addressing the morphology. That works well under the assumption that there are only few vowel interchanges in the language and the language itself is not morphologically rich. For languages like Armenian, Russian, etc. it is vital to take into account the morphology of the language during the process as slight modifications in the words may lead to drastic changes in their meaning.

The attempt of addressing the morphology, however, has a lot of challenges as there are scarce resources for most of the languages: most of them lack good datasets for such training. In this paper, we present our attempt to improve word embedding learning for morphologically rich languages. Our experiments are mostly based on the Russian language, but the pipeline can be applied to any other language if an appropriate dataset is provided.

The codebase developed in the scope of this project is open sourced and available on GitHub ^{3 4 5 6}.

2 Related work

Traditional approaches to model word embeddings like word2vec [9] have been successful in learning word representations for languages with large corpora. As these approaches capture the contextual information from large texts, they successfully estimate representations for frequent words but have difficulties with rare words, or the ones they have not seen before.

One way of overcoming this problem is the method of representing a word by its subwords like characters, substring, or morphemes. Approaches that try to estimate the morphology of the language like fastText [3] address the issue of rare words by taking into account the subword information. Instead of treating each word as a separate entity, the initial word is split into several continuous substrings called n -grams (where n represents the number of continuous characters; for the vanilla fastText they take all the 3,4,5,6-grams of the input word), after which for every subword a vector representation is obtained, and then collected to represent the final word-embedding. In our work, we have used a slight modification of the fastText model to address the linguistically correct morphology of the language.

For morphologically rich languages it is important to capture the linguistically correct nuances of words while estimating their embeddings. There have been several approaches that address this issue; prop2vec [2], morph2vec [15]. The experiments for prop2vec were done on the Hebrew language with a focus on inflectional morphology rather than derivational as derivations (e.g. affected→ unaffected) often change the meaning of the word drastically [2]. A modification of fastText is used to capture the wordform, lemma, and morphological tags for each word and the training was done on the Hebrew UD corpus, which contains all the needed information. As the morphology of Hebrew is not concatenative, the experiments did not include morpheme segmentation of the word. Finally, the approach is tested on a new dataset for Hebrew semantic similarity [1] and show that this method outperforms the n -gram approach used in fastText. Morph2vec, on the other hand, uses unsupervised morpheme segmentation network and later minimizes the distance between combination of vectors for morphemes and the final word vector from word2vec. They demonstrate significant improvement in semantic word similarity evaluation for the Turkish language.

As the morphology for languages like Russian or Armenian is concatenative unlike Hebrew, we have trained a neural network to do morphological word segmentation

3 <https://github.com/MartinXPN/sentence2tags>

4 <https://github.com/MartinXPN/word2morph>

5 <https://github.com/MartinXPN/morph2vec>

6 <https://github.com/MartinXPN/word2morph2vec>

using char-by-char classification. The initial work for the Russian language was done by [11] where they employed conditional random fields (CRF) to extract morphemes from the input word. [14] used convolutional neural networks and showed a significant improvement over the CRF approach. The experiments were done on the electronic version of the Tikhonov dictionary.

As word embeddings need to be obtained for any word, not only lemmas, we first need to extract the lemma from the wordform. We have obtained this information from the UD datasets and the models trained on their corpora like in prop2vec. We have used the COMBO architecture [12] which uses both convolutional and recurrent layers in its model to obtain lemmas, POS-tags, morphological features, and dependency graph for each word in the input sentence. Their approach achieved 3rd/4th places in the 2018 CONLL-U competition.

3 Method

To learn word vectors we process the given text in three steps. First, we process sentences to get lemmas, morphological-tags, and POS-tags for each word. Second, we pass the obtained lemma to a morphological analyzer which obtains the morphemes of the word; each lemma is split into several continuous subwords which are one of {root, prefix, suffix, end, link}. And third, all the extracted features for the words are collected to be passed to a modification of the fastText algorithm.

3.1 sentence2tags: Extraction of lemmas and morph tags from a sentence

For the part of the system where we need to extract lemmas, morphological-tags, and POS-tags for each word, we've used one of the most popular systems for UD training called COMBO [12]. Our task is a subset of the things that are solved by the UD systems. So, we've tuned the system to be more specific for our problem.

Data Our experiments are based on the 2019 SynTagRus ⁷ UD dataset which contains several thousands of human annotated sentences, having all the information we need; lemma, morph-tags, POS-tag for each word.

Model We've used a slight modification of the COMBO [12] model which is publicly available on GitHub ⁸. The model is ranked 3rd/4th on the CoNLL 2018 competition ⁹ depending on the language. We've used only the features that are needed for our task by changing the loss weights for the outputs and selected only the final output layers for lemmatization, morphological tagging, and POS-tagging.

⁷ https://github.com/UniversalDependencies/UD_Russian-SynTagRus

⁸ <https://github.com/360er0/COMBO>

⁹ <http://universaldependencies.org/conll18/>

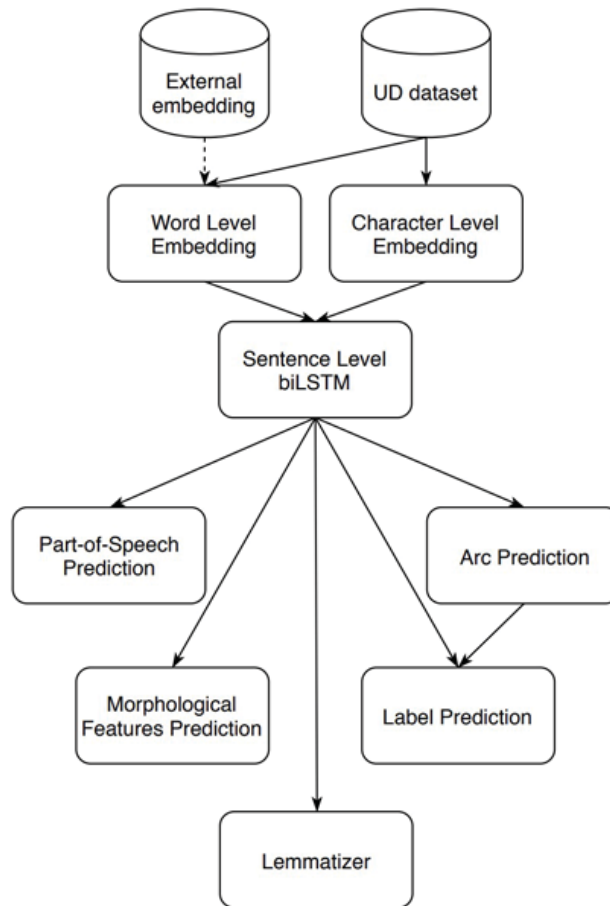


Fig. 1. The schema of the COMBO model

The loss weights for lemmatization, POS-tagging, and morphological tagging were set to 0.2, 0.2, and 0.8 respectively. The final loss of the model is the sum of these three losses.

Results We have not changed the default parameters for training and experimented with providing fastText external embeddings and learning those automatically during the training. We have noticed a slightly better performance when the embeddings were learnt automatically.

3.2 word2morph: Extraction of morphemes from the lemma

After having the lemmas, POS-tags, and morphological features extracted for each word in the sentence we then attempt to further enrich the information by extracting morphemes from the lemmas.

	lemma	pos	feat
CoNLL'18 COMBO	97.60	98.36	96.32
Ours - embeddings	97.03	97.68	94.39
Ours - no embeddings	97.87	98.47	96.50

Table 1. Accuracy of several models on Russian UD dataset (ru_syntagrus)

у
ч
и
т
е
л
ь
 B-ROOT E-ROOT S-SUFF B-SUFF M-SUFF M-SUFF E-SUFF

Fig. 2. Char-by-char classification example

Data We’ve used the data published by [14] available on GitHub¹⁰. The dataset for the Russian language consists of several thousands of lemma annotations with their morphological representations. There are 5 types of morphemes {ROOT, PREFIX, SUFFIX, END, LINK}.

radioactivity	radio/act/iv/ity
рудник	руд:ROOT/ник:SUFF
ножной	нож:ROOT/н:SUFF/ой:END

Table 2. Samples from the dataset

The data was initially split into train and test sets. We’ve further split it into a train, test, and validation sets by randomly dividing the train data into 20% and 80% sets validation and train correspondingly to be consistent with the data split used in [14].

The resulting dataset split was: 57k, 14k, 24k samples for training, validation, and test sets respectively.

One important thing to note here is that the data doesn’t contain transformations of characters from the source word to morpheme-segments, which opens space for char-by-char classification.

We treat the problem as a character-level classification problem where we need to guess a class for each character in a word. The model gets the word (list of characters) as an input and predicts a class for each character. Each symbol in the word can be part of {ROOT, PREFIX, SUFFIX, END, LINK} and also it can be at the {Begin, Middle, End, Single} of the segment. So, each character can belong to the cross product of these possibilities. For an example of such classification from [14] see Figure 2.

¹⁰ <https://github.com/AlexeySorokin/NeuralMorphemeSegmentation/tree/master/data>

Model We’ve experimented with both convolutional and recurrent (bidirectional GRU [4] [13]) networks and observed that both types show comparably similar results. We’ve done an extensive hyperparameter search on both hyperparameters and the architecture parameters with Bayesian tuning and Bandits [5].

Parameter	Range	Final choice
Learning rate	[0.001, 0.01]	0.0016
Learning rate decay	[0.01, 0.1]	0.0687
Batch size	[4, 128]	52
Model type	[CNN, RNN]	RNN
Char-embed. size	[4, 18]	13
Pre-output dense	[16, 256]	217
Dropout	[0, 0.6]	0.2319
Use CRF	[yes, no]	yes
CNN (each layer)		
Layers	[3, 4]	3
Kernel size	[3, 7]	5,5,5
Filters	[32, 384]	256,192,128
Dilations	[1, 5]	1,1,1
RNN (each layer)		
Layers	[2, 3]	2
Recurrent units	[16, 512]	438,427

Table 3. Hyperparameters for model selection

In CNNs, we apply dropout after every convolutional layer. In RNNs, we apply dropout after each BiGRU layer. The feature layers are then processed by a time-distributed fully connected layer which is the pre-output layer.

We use PReLU activation [6] for all the layers in our model. The weights are initialized with glotot-uniform. The batch size is also a hyperparameter and the training lasts for at most 100 epochs, while we prematurely stop if the word-level accuracy doesn’t improve for 10 epochs in a row. We use Adam [8] optimizer with its default parameters, except the clipnorm which is set to 5.

The learning rate is updated at the end of every epoch and decreased exponentially. For each epoch the value of the learning rate can be obtained with:

$$lr * e^{-epoch \cdot decay}$$

As the network outputs probabilities for each class for every character in the word, we post-process the probabilities to remove impossible combinations (i.e. two consecutive characters cannot be a start of a segment at the same time) and select the combination which maximizes the joint probability with a beam-search [16].

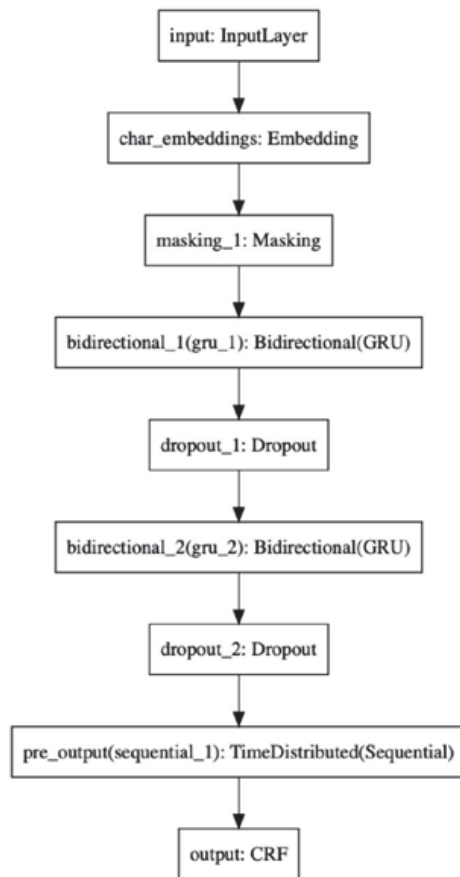


Fig. 3. Sample RNN model architecture

Results We have compared our results with those of [14] and [11] and demonstrate that our model achieves state of the art results on morpheme segmentation from words. Word accuracy indicates how many words had a flawless morpheme segmentation on the test set. We do not report results on character-level features like accuracy and F1 score as they are not comparable to the ones in other methods because they prepended start and end symbols to the words while we do not modify the inputs.

We have noticed that the CRF final layer plays a minor role in improving the final quality if the output probabilities are post-processed with a beam-search. While if no post-processing is done, using CRF improves the performance by almost 2% on word-level accuracy.

The RNN model with the final CRF layer has 88.8 word-level accuracy and outperforms the previous approaches.

To ensemble different models, we have used 3 models (2 CNNs and 1 RNN) trained without the final CRF layer. The final layer contains probabilities for each class for every character in the word, so, to ensemble the models we have calculated the average of the 3 models' output probabilities. The two CNN models have the same architecture and differ only in random initialization. The RNN model has the same architecture as the best single RNN model without the final CRF layer.

3.3 morph2vec: Learning word-vectors from a subset of WLTMN features

The last piece in the pipeline is responsible for learning word-vectors from extracted features.

	Word accuracy
Ruokolainen	65.29
Ruokolainen, Harris features	68.19
Sorokin	86.42
Sorokin memo	86.42
Sorokin ensemble+memo	88.62
Our model	88.8
Ensemble	89.6

Table 4. Comparison with existing methods

For each word we use several features:

- W - wordform (original word)
- L - lemma (lemma from sentence2tags)
- T - morphological tags (including POS-tag)
- M - morphemes (extracted by word2morph)
- N - n-grams (the same as in fastText)

Data We’ve preprocessed the Russian Wikipedia and extracted the plain text from it. Next, tokenized it with SpaCy [7] and saved the raw tokenized text file which is available on GitHub ¹¹.

After the text-processing step, we pass the tokenized text to sentence2tags to extract lemmas, morphological tags, and POS-tags for each word. Which is followed by morpheme segmentation of the extracted lemma for each lemma in the sentence. All of these features are then saved in the WLTMN format.

An example of a WLTMN formatted word:

```
w:развития l:развитие t:Animacy=Inan t:Case=Gen t:Gender=Neut t:Number=Sing
t:POS=NOUN m:раз:PREF m:ви:ROOT m:ти:SUFF m:e:END n:раз n:азв n:зви
n:вит n:ити n:тия n:разв n:азви n:звит n:вити n:ития n:разви n:азвит n:звити
n:вития n:развит n:азвити n:звития
```

Model We use a modification of fastText [3] called prop2vec [2]. Instead of the vanilla fastText where the model splits the word into n-grams and extracts vectors for each n-gram and then sums them up, we split the WLTMN-formatted word by some special character and use each feature as a separate subword. So instead of n-grams, we use special-character separated subwords.

We’ve trained fastText with the default parameters except for the minCount which indicates the minimum number of word occurrences to be included in the data. We set it to be 1 to include all the words in the corpus. The training corpus is the preprocessed Wikipedia corpus and the whole process lasts for about 4 hours on 16 core CPU.

Results We have evaluated the word embeddings on the human-judgement dataset published by Russe Evaluation [10]. The dataset includes 398 word-pairs with their simi-

¹¹ <https://github.com/MartinXPN/morph2vec/releases/download/v0.2.0/ru-wiki-text.zip>

ilarity judged by several human annotators (word1, word2, sim). As the dataset included only lemmas, we have expanded it to be able to evaluate the vectors both semantically and morphologically.

Having the Russian UD corpus, we have collected all the lemmas, wordforms, and their morphological tags. After that, given the (lemma1, lemma2, sim) from the original human-judgement dataset, we have added (wordform1, wordform2, sim) to the expanded dataset if the morphological tags of both lemmas match.

For each word-pair, we have only taken the first 10 possible expansions as most of the words had 0 possible expansions, and some had about 50. After the expansion, the dataset consists of 1500 unique word pairs with their similarity.

We have evaluated the models trained on the subset of WLTMN features by computing the Spearman’s correlation between the expanded human-judgement similarity dataset and the cosine similarity of vectors predicted by the model.

To show how performance may vary from the number of times the word occurs in the dataset, we have evaluated the word embeddings on several data splits: most rare 250, and all the word pairs. Where the word pairs are ordered by the rarest occurrence in the Russian wiki.

To show the robustness of the evaluation and the models, we have calculated the mean and standard deviation obtained from bootstrapping for 10,000 times.

Features	hj-rare 250	hj-all
M	0.513 +/- 0.050	0.586 +/- 0.017
N	0.445 +/- 0.057	0.602 +/- 0.017
TM	0.435 +/- 0.053	0.509 +/- 0.019
WLTM	0.535 +/- 0.048	0.586 +/- 0.017
WLTMN	0.515 +/- 0.052	0.576 +/- 0.017
WM	0.535 +/- 0.050	0.612 +/- 0.016
WN	0.495 +/- 0.052	0.623 +/- 0.016

Table 5. Mean and std obtained from bootstrapping the Spearman’s correlation between gold similarity and cosine distance of predicted vectors on 1500 word pairs

One thing to note here is that WN is the equivalent of the vanilla fastText algorithm as it takes into account only the wordform and the n-grams.

The superior performance of WLTM model on rare words suggests that taking into account linguistically correct features of the word rather than substring (n-grams) helps in modelling the word even when the model has not trained on it before. We suspect that using some other architecture for getting word embeddings instead of the modification of the fastText algorithm may boost the performance especially for the models like TM (morphological tags + morphemes).

The TM model can represent all the WLTM combinations as the lemma can be represented by the concatenation of morphemes, the wordform by the combination of lemma and the morphological tags, and the n-grams can be derived from the wordform

itself. So, in future, a model with less explicit parameters may outperform a model which needs much more explicit content like WLTM.

The performance of WN (wordform + n-gram; fastText) on all the evaluation dataset may suggest that the hyperparameters and the architecture is tuned to be specifically suitable for the extraction of vectors from n-grams. So, a better investigation of other architectures and hyperparameters may improve the results.

4 Conclusion

We described an end-to-end system for obtaining word vectors for morphologically rich languages and demonstrated the performance of every piece of the system. It is worth noting that this system is especially useful for low-resource languages as it is able to capture the meaning of rare words only from having information on their morphological and syntactical features.

References

- [1] Oded Avraham and Yoav Goldberg. “Improving Reliability of Word Similarity Evaluation by Redesigning Annotation Task and Performance Measure”. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, Jan. 2016, pp. 106–110. DOI: 10.18653/v1/W16-2519.
- [2] Oded Avraham and Yoav Goldberg. “The Interplay of Semantics and Morphology in Word Embeddings”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 422–426. URL: <https://www.aclweb.org/anthology/E17-2067>.
- [3] Piotr Bojanowski et al. “Enriching Word Vectors with Subword Information”. In: arXiv preprint arXiv:1607.04606 (2016).
- [4] Junyoung Chung et al. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. Cite arxiv:1412.3555 Comment: Presented in NIPS 2014 Deep Learning and Representation Learning Workshop 2014. URL: <http://arxiv.org/abs/1412.3555>.
- [5] Laura Gustafson. “Bayesian Tuning and Bandits: An Extensible, Open Source Library for AutoML”. M. Eng. Thesis. Cambridge, MA: Massachusetts Institute of Technology, May 2018. URL: https://dai.lids.mit.edu/wp-content/uploads/2018/05/Laura_MEng_Final.pdf.
- [6] Kaiming He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *CoRR* abs/1502.01852 (2015). arXiv:1502.01852. URL: <http://arxiv.org/abs/1502.01852>.
- [7] Matthew Honnibal and Ines Montani. “spaCy2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. In: *To appear* (2017).
- [8] Diederik Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations* (Dec. 2014).

- [9] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119. URL: <http://dl.acm.org/citation.cfm?id=2999792.2999959>.
- [10] Alexander Panchenko et al. “RUSSE’2018: A Shared Task on Word Sense Induction for the Russian Language”. In: *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”*. Moscow, Russia: RSUH, 2018, pp. 547–564. URL: <http://www.dialog-21.ru/media/4324/panchenkoa.pdf>.
- [11] Teemu Ruokolainen et al. “Painless Semi-Supervised Morphological Segmentation using Conditional Random Fields”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 84–89. DOI: 10.3115/v1/E14-4017. URL: <https://www.aclweb.org/anthology/E14-4017>.
- [12] Piotr Rybak and Alina Wróblewska. “Semi-Supervised Neural System for Tagging, Parsing and Lemmatization”. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 45–54. URL: <https://www.aclweb.org/anthology/K18-2004>.
- [13] M. Schuster and K.K. Paliwal. “Bidirectional Recurrent Neural Networks”. In: *Trans. Sig. Proc.* 45.11 (Nov. 1997), pp. 2673–2681. ISSN: 1053-587X. DOI:10.1109/78.650093. URL: <http://dx.doi.org/10.1109/78.650093>.
- [14] Alexey Sorokin and Anastasia Kravtsova. “Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language”. In: *Artificial Intelligence and Natural Language*. Ed. by Dmitry Ustalov et al. Cham: Springer International Publishing, 2018, pp. 3–10. ISBN: 978-3-030-01204-5.
- [15] Ahmet Üstün, Murathan Kurfalı, and Burcu Can. “Characters or Morphemes: How to Represent Words?” In: *Proceedings of The Third Workshop on Representation Learning for NLP*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 144–153. URL: <https://www.aclweb.org/anthology/W18-3019>.
- [16] Sam Wiseman and Alexander M. Rush. “Sequence-to-Sequence Learning as Beam-Search Optimization”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1296–1306. DOI: 10.18653/v1/D16-1137. URL: <https://www.aclweb.org/anthology/D16-113>

The role of alignment of multilingual contextualized embeddings in zero-shot cross-lingual transfer for event extraction

Karen Hambardzumyan¹, Hrant Khachatrian^{1,2}, and Jonathan May³

¹ YerevaNN

² Department of Informatics and Applied Mathematics, Yerevan State University
{mahnerak,hrant}@yerevann.com

³ Information Sciences Institute, University of Southern California
jonmay@isi.edu

Abstract. Contextualized word embeddings like BERT enabled significant advances in many natural language processing tasks. Recently, multilingual versions of such embeddings were trained on large text corpora of more than 100 languages. In this paper we investigate how well such embeddings perform in zero-shot cross lingual transfer for an event extraction task. In particular, we analyze the impact of the alignment of contextualized word embeddings using a parallel corpus on the performance of the downstream task.

Keywords: Multilingual contextual embeddings · Zero-shot transfer · Event extraction.

1 Introduction

Many recent advances in natural language processing were made possible due to the progress made in unsupervised language modeling [5, 7]. By training a language model on a huge amount of unstructured text and fine-tuning the model on labelled data of the downstream task it was possible to achieve state-of-the-art results on several natural language understanding tasks. However, the vast majority of works focus only on data-rich languages such as English, while low-resource languages with scarce labeled datasets are yet to see significant benefits from unsupervised pretraining.

Recently, a multilingual version of BERT [2], one of the most popular language models, was released under the name *mBERT*. It was trained on the concatenation of 104 language versions of Wikipedia. The hope is that if the embeddings of similar sentences in various languages are close in the shared space, then a classifier trained on sentences of one language will generalize to sentences in another language. This setup is known as zero-shot cross-lingual transfer, as no labeled sentences from the target language are involved in the training process.

mBERT showed surprisingly good cross-lingual transfer performance by obtaining state-of-the-art results on Cross-Lingual Natural Language Inference

(XNLI [1]) task. [6] suggested that the representation spaces learned in mBERT are different for each language, as the network has to encode the language in order to predict the masked word in the same language. Further analysis showed that it is possible to find translation pairs in the mBERT space with high accuracy by adding the difference of the average embeddings from two languages. The effectiveness of multilingual representations from mBERT for the cross-lingual setup in event trigger extraction task was analyzed in [4].

In this work we attempt to analyze whether bringing the representations of similar sentences in two languages even closer to each other can improve the zero-shot cross-lingual transfer for one particular NLP task, event extraction. Following terminology from [3], we call this process “alignment” of sentence representations in two languages. We leverage an external parallel corpus to perform the alignment.

2 Experimental Setup

The event trigger extraction is an information extraction task that requires to find the “triggers” - the words in the sentence that show an event. For example, in the sentence “The toughest *fight*, though, may lie ahead in the heart of the Iraqi capital.” the word “fight” is a trigger word for an event of type *Conflict.Attack*. This example comes from the ACE’05 dataset [8], which contains labelled event extraction data in three languages (English, Arabic and Chinese) with a fixed set of 33 event types. In this work, we consider only zero-shot transfer task, that is, our models are only trained on the English version of ACE’05 training data. Arabic and Chinese parts are used only for evaluation.

For simplicity, we investigate multi-class multi-label sentence-level classification setup: instead of finding the exact trigger words we attempt to predict the set of all trigger types available in the sentence. For more detailed analysis, we also translate both training and test sets of the English sentences into Arabic and German using Google Translate.

Inspired by [6], we attempt to map mBERT embeddings into a shared space such that similar sentences in various languages end up in similar locations in the space. We obtain a representation of a sentence by aggregating mBERT embeddings of its tokens using a simple attention-based mechanism. We perform the mapping using a single linear layer on top of the sentence representations. We use a multi-task training setup. For each batch, we optimize either the downstream task L_{ee} (standard cross-entropy loss), or the alignment loss between two languages L_{align} . Alignment loss is implemented as a triplet loss with semi-hard negative mining. For alignment, we used English-German and English-Arabic parallel data from OPUS ⁴.

We trained two types of models. First, we froze mBERT layers and only trained the pooler, mapper and the classifier. In this setup we used Adam optimizer with a learning rate 10^{-4} , with batch size 16. In the second set of experiments we also fine-tuned the whole mBERT model. For this setup we used batch

⁴ <http://opus.nlpl.eu/>

size 4, AdamW optimizer with a weight decay rate 0.01 along with a slanted triangular learning rate schedule. We performed hyperparameter search over the relative weight of the two terms in the loss function and the margin parameter in the triplet loss.

3 Results

	Model selection	Event Classification F-score				Alignment %	
		en	en → de	en → ar	ar	en-de	en-ar
No alignment	en	65.6	40.2	31.8	18.7	86.2	51.4
No fine-tuning	en → de	63.6	48.3			89.0	
	en → ar	59.4		43.0	23.4		35.0
	ar	59.0		39.3	30.9		34.8
No alignment	en	65.6	43.0	30.3	16.8	90.0	62.4
Fine-tuning	en → de	63.6	55.4			46.6	
	en → ar	62.8		51.0	33.6		62.6
	ar	61.1		45.1	36.8		14.8
Align on en-de	en	64.5	42.5			97.2	
No fine-tuning	en → de	62.6	52.5			98.8	
	en → de #2	63.4	51.9			91.6	
Align on en-de	en	60.2	6.0			86.2	
Fine-tuning	en → de	57.5	45.3	48.4	34.6	97.0	
Align on en-ar	en	63.6		38.7	23.3		90.2
No fine-tuning	en-ar	60.6		44.2	23.4		96.4
	en-ar #2	59.5		43.5	30.4		94.8
	ar	52.0		39.0	33.3		97.0
	ar #2	57.4		34.5	33.0		97.6
Align on en-ar	en	63.2		21.3	7.8		89.6
Fine-tuning	en → ar	59.2	46.0	49.5	24.8		91.2
	ar	60.5	46.0	42.3	35.9		91.2

Table 1. The results of our experiments. We report the scores for top one or top two models per each model selection, alignment and fine-tuning strategies.

Table 1 lists the results of all our experiments. We first note that the performance of our models strongly depends on the way we perform model selection. Although we always train only on English event extraction data, there are several ways to select the best performing model. We try the following criteria:

1. The best F1 score on the English dev set,
2. The best F1 score on the Arabic/German translation of the English dev set,

3. The best F1 score on the Arabic dev set.

The table shows that the fine-tuned models generally perform better than the frozen models. Also, it can be seen that, unsurprisingly, the cross-lingual transfer works better from English to German than from English to Arabic. Finally, the models with alignment significantly increase the alignment score, which is calculated as the percentage of the sentences for which the closest sentence of the other language in the batch is its translation. So, the representations of the same sentence in different languages get closer, but its impact on the event classification accuracy in the target language is much less significant.

References

1. Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2475–2485 (2018)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
3. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: International Conference on Learning Representations (2018), <https://openreview.net/forum?id=H196sainb>
4. M’hamdi, M., Freedman, M., May, J.: Contextualized cross-lingual event trigger extraction with minimal resources. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). pp. 656–665 (2019)
5. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
6. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4996–5001. Association for Computational Linguistics, Florence, Italy (Jul 2019). <https://doi.org/10.18653/v1/P19-1493>, <https://www.aclweb.org/anthology/P19-1493>
7. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. URL <https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language-understanding-paper.pdf> (2018)
8. Walker, Christopher and, S.S., Medero, J., Maeda, K.: Ace 2005 multilingual training corpus ldc2006t06. Web Download. Philadelphia: Linguistic Data Consortium, 2006.

On the Tradeoff Between Accuracy and Fairness in Representation Learning

Tigran Galstyan¹ and Hrant Khachatrian²

²YerevaNN Research Lab hrant@yerevann.com

¹Department of Informatics and Applied Mathematics,
Yerevan State University tigran@yerevann.com

Abstract. In many applications of machine learning, it is desirable to have models which not only have good accuracy on the prediction task but are also “fair” with respect to some protected variable. One approach to achieve fairness is to learn an invariant representation of the data with respect to that variable and then learn the predictor on top of the representation. Recently, an information-theoretic approach called DSF (Discovery and Separation of Features) was introduced, which demonstrated strong results in cases where the label and the protected variable are independent. In this paper we extend the model to work in cases when the protected variable is correlated with the label. We perform experiments on a small image classification dataset and show that our model enables significantly better tradeoffs between accuracy and fairness.

Keywords: Machine Learning · Representation Learning · Information Theory.

1 Introduction

Learning “fair” or “invariant” representations of data with respect to some protected variable is an important task in machine learning. In supervised learning, the goal is to find a function which can predict an unknown variable y from the given variable x . In some applications, it is also necessary to guarantee the function does not depend on a specified variable c , which might be correlated with x and/or y . For example, the training data might have biases, which can result in functions that discriminate against certain groups of people [4]. One general approach to obtain functions with such guarantees is to learn a “fair” representation of data which attempts to simultaneously satisfy two constraints: it should not contain information about the protected variable (e.g. race of the person), but it should also keep enough information about x so that it is possible to predict the desired variable [5].

We use upper case letters such as X for random variables, lower case letters such as x for realizations. We assume the data comes from an unknown $P(X, Y, C)$ joint distribution. For simplicity, we assume the support for random variables C and Y (denoted by Ω_C and Ω_Y , respectively) are finite. For $(x, y, c) \sim P(X, Y, C)$, the goal is to learn a representation of x , $z = f_z(x)$, such that it is possible to learn a predictor $\hat{y} = f_y(z)$ with high accuracy, while z has no information about c .

The recently proposed DSF model [3] attempts to achieve the goal by solving the following constrained optimization problem:

$$\begin{aligned} \max \quad & \alpha I(Z_p : Y) + I(X : \{Z_p, Z_n\}) \\ \text{s.t.} \quad & I(Z_p : X) \leq I_c \text{ and } I(Z_p : Z_n) = 0 \end{aligned} \quad (1)$$

Here $I(\cdot : \cdot)$ is Shannon mutual information between two random variables; Z_p and Z_n are representations of X . This model gets great results on several benchmarks: it learns a representation which can be used to predict Y with a perfect accuracy, while it has no information about C . Here we show that if $p(C|Y = y)$ is not uniform (unlike the examples discussed in the DSF model), then a representation that allows perfect classification of Y will contain at least some information about C .

Proposition 1. *If there exists a classifier f with $\text{acc}(Y|f(X)) = 1$, then there exists a classifier f_C s.t. $\text{acc}(C|f_C(X)) = \sum_{y \in \Omega_Y} p(y) \max_{c \in \Omega_C} p(c|y)$,*

Where $\text{acc}(Y|f(X))$ is the accuracy of a classifier f .

2 Our model

We attempt to extend DSF model to work in the case when Y and C are not independent. In particular, we consider the case when the marginal distribution $p(C)$ is uniform, but $p(C|Y = y)$ is not necessarily uniform for each $y \in \Omega_Y$. This allows non-zero mutual information between Y and C .

It is easy to see that the optimization problem (1) can have solutions for which $I(Z_p : Y) = H(Y)$. In such solutions, Z_p will contain at least some information about C , as $I(Y : C) > 0$. To avoid such solutions, we suggest two modifications. We replace the term $I(Z_p : Y)$ with $I(Z_p : Y|C)$, and add additional constraints: $I(Z_p : C) = I(Z_n : C) = 0$. The modified optimization problem can be solved by maximizing the following objective function:

$$\begin{aligned} J_{\text{sDSF}} = & \alpha I(Z_p : Y|C) + I(X : \{Z_p, Z_n, C\}) - \lambda I(Z_p : X) \\ & - \gamma I(Z_p : Z_n) - \gamma I(Z_p : C) - \gamma I(Z_n : C) \end{aligned} \quad (2)$$

We call this model *supervised* DSF (sDSF), as we have access to supervision signal on C . Following DSF, we propose two implementations of sDSF. The first one uses Hilbert-Schmidt independence criterion (HSIC [2]) for minimizing the last three terms of (2) (sDSF-H), while the second one follows the ‘‘independence through compression’’ strategy (sDSF-C). The training objective becomes:

$$\begin{aligned} \hat{J}_{\text{sDSF-H}} = & \alpha \mathbb{E} \log p(Y|Z_p, C) + \mathbb{E} \log p(X|\{Z_p, Z_n, C\}) + \lambda \mathbb{E} \log |\det S_p(X)| \\ & - \gamma_1 \text{HSIC}(Z_p, Z_n) - \gamma_2 \text{HSIC}(Z_p, C) - \gamma_3 \text{HSIC}(Z_n, C) \end{aligned} \quad (3)$$

$$\begin{aligned} \hat{J}_{\text{sDSF-C}} = & \alpha \mathbb{E} \log p(Y|Z_p, C) + (1 + \gamma) \mathbb{E} \log p(X|\{Z_p, Z_n, C\}) \\ & - (\lambda + \gamma) \mathbb{E} \log |\det S_p(X)| - \gamma \mathbb{E} \log |\det S_n(X)| \end{aligned} \quad (4)$$

Where $S_p(X)$ and $S_n(X)$ are sigma matrices of variational autoencoders (VAE) used for obtaining Z_p and Z_n . $\mathbb{E} \log |\det S(X)|$ come from Echo loss, thoroughly discussed in [1].

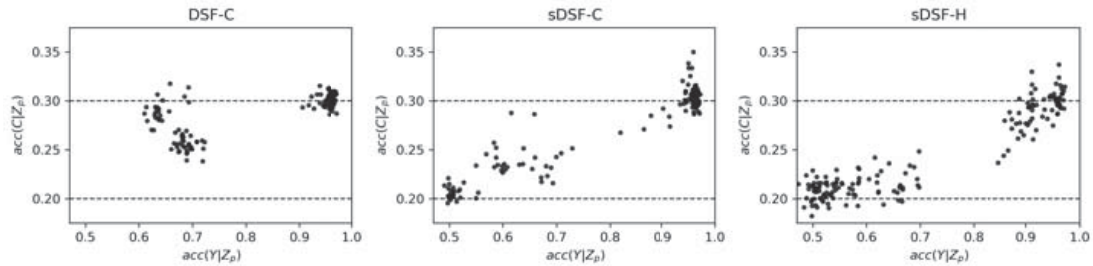


Fig. 1. Results of the three models on modified MNIST-ROT. Horizontal and vertical axes show $acc(Y|Z_p)$ and $acc(C|Z_p)$, respectively.

3 Experiments

For our experiments we use a modified version of MNIST-ROT dataset used in prior work. We keep only two digits, $\Omega_Y = \{4, 9\}$ and use larger rotation angles $\Omega_C = \{0, \pm 45, \pm 90\}$, as MNIST digits already have small inherent variability in rotation angles. In our dataset, the rotation angles are not uniformly distributed for each digit. In particular, $p(C|Y=4) = (0.3, 0.25, 0.2, 0.15, 0.1)$ and $p(C|Y=9) = (0.1, 0.15, 0.2, 0.25, 0.3)$. This makes sure the marginal distribution $p(C)$ is uniform, but a model with a perfect accuracy on Y will have at least 30% accuracy according to Proposition 1. Training, validation and test sets contain 7594, 1903 and 2294 samples, respectively. We perform experiments with three models: regular DSF-C and two versions of supervised DSF: sDSF-C, sDSF-H.

Fig. 1 shows our results. Plots correspond to the baseline model and two versions of our approach described in the previous section. The points correspond to all checkpoints of all hyperparameter choices. The coordinates of the points in the plot show accuracies of predicting Y and C from the Z_p at the specified checkpoint. It can be seen that there are a few sDSF-C checkpoints which have around 22% accuracy for predicting C , while $acc(Y|Z_p)$ is around 70%. These are already superior to the baseline models. On the other hand, sDSF-C checkpoints with less than 20.5% $acc(C|Z_p)$ appear only with $acc(Y|Z_p) < 55\%$, which means they have almost no information about the label. sDSF-H is slightly better than sDSF-C by two aspects. First, the best checkpoints with $acc(C|Z_p) = 25\%$ get up to 90% $acc(Y|Z_p)$. Second, the best checkpoints with $acc(C|Z_p) \leq 20\%$ have around 67% accuracy for predicting Y .

References

1. Brekelmans, R., Moyer, D., Galstyan, A., Ver Steeg, G.: Exact rate-distortion in autoencoders via echo noise. In: Advances in Neural Information Processing Systems. pp. 3884–3895 (2019)

2. Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., Smola, A.J.: A kernel statistical test of independence. In: Advances in neural information processing systems. pp. 585–592 (2008)
3. Jaiswal, A., Brekelmans, R., Moyer, D., Steeg, G.V., AbdAlmageed, W., Natarajan, P.: Discovery and separation of features for invariant representation learning. arXiv preprint arXiv:1912.00646 (2019)
4. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication. pp. 1–6. IEEE (2009)
5. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)

Excess-Risk consistency of group-hard thresholding estimator in Robust Estimation of Gaussian Mean

Arshak Minasyan

Yerevan State University, YerevaNN
minasyan@yerevann.com

Abstract. In this work we introduce the notion of the excess risk in the setup of estimation of the Gaussian mean when the observations are corrupted by outliers. It is known that the sample mean loses its good properties in the presence of outliers [5,6]. In addition, even the sample median is not minimax-rate-optimal in the multivariate setting. The optimal rate of the minimax risk in this setting was established by [1]. However, even these minimax-rate-optimality results do not quantify how fast the risk in the contaminated model approaches the risk in the uncontaminated model when the rate of contamination goes to zero. The present paper does a first step in filling this gap by showing that the group hard thresholding estimator has an excess risk that goes to zero when the corruption rate approaches zero.

Keywords: Robust estimation · Minimax estimation · Excess risk.

1 Introduction

In recent years, we witnessed a revival of interest in statistical methods that can efficiently deal with data sets corrupted by outliers. In particular, under the Huber contamination model in the problem of Gaussian mean estimation, [1] established the minimax rate and showed that it is attained by Tukey's median. Furthermore, [2] developed a general theory for obtaining the minimax rate (both upper and lower bounds) in a wide class of statistical models. These works are focused on statistical complexity of the estimators, without paying attention to the computational complexity. The latter has been addressed by [4] and [3], who analyzed the risk of computationally tractable estimators. Interestingly, the results proved in these papers only provide the order of magnitude of the minimax risk and do not tell anything about how fast the risk in the corrupted setting get close to the risk in the uncorrupted setting.

In this paper, we introduce the notion of the excess risk, which is defined as the difference between the risks in the corrupted and uncorrupted settings. Then, we present an analysis of this risk for a procedure that we call group hard thresholding estimator. It can also be seen as a version of the trimmed mean estimator. Our main result shows that this excess risk goes to zero, as the rate of contamination goes to zero.

To be more precise, let us assume that we observe n random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ in \mathbb{R}^p , which are assumed to satisfy

$$\mathbf{Y}_i = \boldsymbol{\mu} + \boldsymbol{\theta}_i + \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_p). \quad (1)$$

In the above formula, $\boldsymbol{\mu}$ is the unknown mean we wish to estimate, $\{\boldsymbol{\theta}_i\}$ are arbitrary deterministic vectors measuring the outlyingness of each data point and $\boldsymbol{\xi}_i$ are random errors. In this paper, we assume that $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_n]$ is a column-wise sparse matrix. All the observations with indices $i \in \mathcal{O} = \{\ell : \|\boldsymbol{\theta}_\ell\|_2 > 0\}$ are considered as outliers, while all the other are called inliers. In the sequel, we use notation

$$o = \text{Card}(\mathcal{O}), \quad \text{and} \quad \varepsilon = \frac{o}{n}.$$

The parameter ε , assumed to be strictly smaller than $1/2$, plays an important role in robust estimation. In particular, it is known that the minimax rate of estimation in model (1) is of order $\frac{p}{n} + \varepsilon^2$.

In this paper, we propose to consider a more precise measure of accuracy of an estimator, the excess risk. Recall that the risk of an estimator¹ $\widehat{\boldsymbol{\mu}}_n$ is given by

$$R[\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}; \boldsymbol{\Theta}] = [\mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Theta}} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2]^{1/2}.$$

In the above formula and in the sequel, the notation $\mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Theta}}[h]$ stands for the expectation with respect to the distribution of $\{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$ as defined by Eq. (1) (it is implicitly assumed that the function h depends on the observations $\{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$). It is a well-known fact that in the outlier-free setup, where $\boldsymbol{\Theta} \equiv \mathbf{0}_{p \times n}$ the minimax risk satisfies $\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathbb{R}^p} R[\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}; \mathbf{0}] = \sup_{\boldsymbol{\mu} \in \mathbb{R}^p} R[\overline{\mathbf{Y}}_n, \boldsymbol{\mu}; \mathbf{0}] = \sqrt{\frac{p}{n}}$, with $\overline{\mathbf{Y}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i$ being the sample mean of the observed vectors. Let us define the mixed matrix norm

$$\|\boldsymbol{\Theta}\|_{0,2} = \sum_{i=1}^n \mathbb{1}(\|\boldsymbol{\theta}_i\|_2 > 0).$$

Based on the expression of minimax risk in the outlier-free setup we define the worst-case excess risk of an estimator $\widehat{\boldsymbol{\mu}}$ by

$$\mathcal{E}(\widehat{\boldsymbol{\mu}}; n, p, \varepsilon) = \sup_{\boldsymbol{\mu} \in \mathbb{R}^p; \|\boldsymbol{\Theta}\|_{0,2} \leq \varepsilon n} R[\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}; \boldsymbol{\Theta}] - \sqrt{\frac{p}{n}}$$

as well as the minimax excess risk

$$\mathcal{E}(n, p, \varepsilon) = \inf_{\widehat{\boldsymbol{\mu}}} \mathcal{E}(\widehat{\boldsymbol{\mu}}, n, p, \varepsilon),$$

where the infimum is over all possible estimators $\widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$. Note that according to the definition, the estimators considered in the above formula can depend on n , p and $\varepsilon = o/n$. The main result of this paper shows that the excess risk of the group hard thresholding estimator, introduced in the next section, tends to zero as $\varepsilon = \varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$, as soon as $p = p_n$ is such that p_n/n is bounded by a constant.

¹ An estimator is any measurable function from $(\mathbb{R}^p)^n$ to \mathbb{R}^p

2 Group-hard Thresholding estimator

In this section we define the estimator $\hat{\boldsymbol{\mu}}_{\text{GHT}}$, called group hard thresholding estimator, and prove that this estimator has an excess risk that vanishes when the proportion of contamination ε tends to zero. Roughly speaking, $\hat{\boldsymbol{\mu}}_{\text{GHT}}$ is the arithmetic mean of a sample obtained from $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ by replacing all the vectors that are at a large distance from the coordinatewise median by the latter.

More specifically, let $\hat{\boldsymbol{\mu}}_{\text{Med}} := \text{Med}(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ be the coordinatewise median of the sample $\{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$. Let us fix a positive threshold $\lambda > 0$, which will be a tuning parameter of the method. For each $i \in \{1, \dots, n\}$, we put

$$\hat{\boldsymbol{\theta}}_i = HT_{\lambda}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{\text{Med}}) := (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{\text{Med}})\mathbb{1}(\|\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{\text{Med}}\|_2 > \lambda) \quad (2)$$

$$\hat{\boldsymbol{\mu}}_{\text{GHT}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \hat{\boldsymbol{\theta}}_i) := L_n(\mathbf{Y} - \hat{\boldsymbol{\Theta}}). \quad (3)$$

Next, we formulate the main theorem of the paper showing that the excess risk of the GHT estimator tends to zero if the proportion of outliers $\varepsilon = \varepsilon_n$ tends to 0 fast enough so that $\varepsilon_n p_n^{1/4}$ goes to zero. Notice that this condition holds when p is fixed, however this setup allows the infinite dimensional case, i.e. $p = p_n \rightarrow \infty$ under the constraint $\varepsilon_n p_n^{1/4} \log^{1/2} \varepsilon_n^{-1} = o(1)$ as the sample size n goes to infinity.

Theorem 1. For $\hat{\boldsymbol{\mu}}_{\text{GHT}}$ defined in (3) and $\lambda^2 = p + 8\sqrt{p \log \varepsilon^{-1}} + 16 \log \varepsilon^{-1}$ we have

$$\overline{\lim}_{n \rightarrow \infty} \mathcal{E}(\hat{\boldsymbol{\mu}}_{\text{GHT}}, n, p_n, \varepsilon_n) = 0$$

provided that $\varepsilon_n p_n^{1/4} \log^{1/2} \varepsilon_n^{-1} = o(1)$ and $p_n = O(n)$ as $n \rightarrow \infty$.

References

1. Chen, M., Chao, G., Ren, Z. A general decision theory for Huber's ε -contamination model. *Electronic Journal of Statistics*, 10, 3752–3774, 2016.
2. Chen, M., Chao, G., Ren, Z. Robust covariance and scatter matrix estimation under Huber's estimation model. *Annals of Statistics*, 46(5), 1932–1960, 2018.
3. Cheng, Y., Diakonikolas, I., Ge, R. High-dimensional robust mean estimation in nearly linear time. *arXiv:1811.09380*, 2018.
4. Collier, O., Dalalyan, A. Rate-optimal estimation of p-dimensional linear functionals in a sparse gaussian model. *Electronic Journal of Statistics*, 13(2), 2830–2864, 2019.
5. Huber, P. Robust estimation of a location parameter. *The annals of mathematical statistics*, 35(1) : 73–101, 1964.
6. Huber, P. A robust version of the probability ratio test. *The annals of mathematical statistics*, 36(6) : 1753–1758, 1965.

Current approaches and challenges for the two-party privacy-preserving record linkage (PPRL)¹

Yanling Chen

University of Duisburg-Essen, Germany
yanling.chen@uni-due.de

Abstract. Integrating data from diverse sources with the aim to identify similar records that refer to the same real-world entities without compromising privacy of these entities is an emerging research problem in various domains. This problem is known as privacy preserving record linkage (PPRL). Despite the abundant number of literature on PPRL, a commonly accepted formal framework is still missing. In this paper, we focus on the two-party PPRL and provide an overview of the currently existing approaches and related works. Several desired properties of two-party PPRL are discussed, which may lay the foundation for a formal description of the process of two-party PPRL and sound definitions of system performance that allow the comparative evaluation of different PPRL techniques.

Keywords: Record linkage, two-party protocol, semi-honest security model, approximate matching, privacy, scalability.

1 Introduction

1.1 Record linkage

Record linkage (RL), also known as *data linkage*, *data matching*, or *entity resolution*, aims to identify and link records that correspond to the same real-world entities within one or across several data sets. In particular, when records about the same entity need to be identified in a single data set, it is called *duplicate detection*, or *deduplication*.

In general, record linkage is a challenging task since a common entity identifier across the data sets to be linked is usually missing. Instead, the common attributes available have to be used for the linkage. Especially for those databases that contain records about people, the common identifying information that characterizes an individual typically are names, addresses, dates of birth, and so on (often referred to as quasi-identifiers or QIDs), which are not always stable over time and can also be missed or recorded with errors.

¹ The work is supported by the German Research Foundation, Deutsche Forschungsgemeinschaft (DFG), Germany, under grant AR 671/5-1 | SCHN 586/29-1.

1.2 Privacy concerns

The idea of data linkage was first described by Halbert Dunn in 1946 as a book of life for each individual [1]. Since then, several research domains have developed data linkage techniques using QIDs [2], [3]. A variety of linkage protocols have been proposed and especially the seminal work by Fellegi and Sunter [4] on probabilistic data linkage provided a sound theoretical basis. However, the problem is far from being completely solved especially with new challenges posed by the big data era as well as new privacy and confidentiality regulations and policies that prohibit the disclosure of personal identifiers.

For instance, the privacy rules of the Health Insurance Portability and Accountability Act (HIPAA) in the US and the Data Protection Directive of the European Union, restrict direct access to QIDs. If one would like to conduct a study of finding correlation between certain kind of automobile accidents and resulting injuries, data from police records, insurance companies and hospitals are needed (to be collected and analyzed). However, these institutions are not going to share data unless strong privacy is guaranteed.

1.3 PPRL=RL+Privacy

Protecting the privacy of personal sensitive information during the linkage process is the (additional) aim of the emerging research area of privacy-preserving record linkage (PPRL). It requires that the parties involved in a linkage learn only limited information about which record pairs are classified as a match, but nothing about the actual records and the values from any other party involved in the linkage.

Formally, PPRL can be defined as follows:

Definition 1 [32]. *Privacy-preserving record linkage (PPRL): Assume that there are m database owners and their respective databases are $\mathbf{D}_1, \dots, \mathbf{D}_m$. The linkage across all these m databases determines which of their records $r_1^{i_1} \in \mathbf{D}_1, r_2^{i_2} \in \mathbf{D}_2, \dots, r_m^{i_m} \in \mathbf{D}_m$, match according to a decision model $C(r_1^{i_1}, \dots, r_m^{i_m})$ that classifies record tuple $(r_1^{i_1}, \dots, r_m^{i_m})$ into one of the two classes, \mathbf{M} of matches and \mathbf{U} of non-matches.*

Moreover, the database owners do not wish to reveal their actual records with any other party; while they are prepared to disclose to a selected party (i.e., the data consumer such as a researcher) the actual values of some selected attributes of the record pairs that are in class \mathbf{M} to allow further data analysis.

1.4 PPRL in practice

PPRL is increasingly being required in many real-world application areas. Examples range from public health surveillance, business analytic, national censuses, population informatics, to crime and fraud detection, government services and national security. One of the currently most popular PPRL scheme was proposed by Schnell et al. [5] in 2009 that is based on Bloom filter encoding. This scheme and its variations have been

implemented in several real-world linkage applications in different countries, e.g., on medical data sets on newborn in Germany.

In more details, a Bloom filter is a space efficient data structure proposed by Bloom [24] in 1970 for checking element membership in a set. The PPRL based on Bloom filter encoding is usually done as follows: each database owner first generates the Bloom filters for its records. More specifically, each record, represented as a string, is split into subsets of length q (i.e., q -grams), where each q -gram determines k bit positions (e.g., by using k hashing functions) to be set to one in a binary vector of length l (i.e., Bloom filter, which initially consists of l zero bits). At the linkage phase, instead of comparing the plain record pair, the encoded Bloom filters are compared. For instance, in case of $q=2$ (i.e., bigrams), the pairs are classified as matches or non-matches based on their calculated Dice similarities.

2 Different forms of PPRL

Proposals to PPRL can be classified into those that require a third party for performing the linkage and those that do not. The former are known as ‘three-party protocols’ and the latter as ‘two-party protocols’. In three-party protocols, a (trusted) third party (which we call the ‘linkage unit’) is involved in conducting the linkage, while in two-party protocols only the two database owners participate in the PPRL process. Considering the difficulty of finding a trusted third party, we mainly put our focus on the two-party protocols.

Generally, two-party protocols start by the two database owners agreeing upon and exchanging any required information such as parameter settings, preprocessing methods, encoding or encryption methods, and any secret keys that are required, and further proceed by the secure exchange of encoded or encrypted attribute values to conduct the linkage. Followed by sending or exchanging the encoded records, the final step is to identify the matched records.

The advantages of two-party over three-party protocols is the fact that no database records are shared with any external party and thus there is no possibility of collusion between one of the database owners and the linkage unit. However, two-party protocols could require more sophisticated encoding or encryption mechanisms, because both database owners know the full details of the agreed parameters or encoding/encryption techniques and therefore they can potentially perform attacks on the exchanged (encrypted) data between them to infer actual values from each other’s data. In other words, the core encryption/encoding techniques need to ensure that each database owner cannot infer any sensitive information (on the non-linked records) from the other database with knowledge of both encrypted data sets and shared system parameters.

For the privacy analysis, often a semi-honest threat model is assumed. A two-party PPRL protocol is secure in a semi-honest model when neither party is able to gain any information from the execution of the protocol, other than the information gained from the protocol’s output (and the size of the other party’s input). The semi-honest security model is contrasted to the malicious security model, where the latter allows adversaries to arbitrarily deviate from the specified protocol while attempting to non-consensually

gain information from the protocol's execution. Nevertheless, it is shown by Goldreich et al. [6] that any protocol that is secure in the semi-honest security model can be made secure in the malicious security model, albeit inefficiently, through the use of what are known as zero-knowledge protocols.

3 Current approaches and related work on two-party PPRL

To gain a comprehensive understanding in the state-of-art solutions and challenges to the two-party PPRL, let us first have a close look into the current approaches and related work that have been proposed.

3.1 Current approaches on two-party PPRL

Yakout et al. 2009 [7] Yakout et al. [7] in 2009 gave a design for two-party PPRL protocol, assuming that the database records have been transformed into numeric vectors by each party as described in Scannapieco et al. [8]. Their protocol works as follows: first, each record (represented as a numeric vector) is transformed into a complex number, mapping to a point in the complex plane. These complex numbers are then exchanged between the two database owners, such that each can generate the pairs of complex numbers that are within a maximum distance from each other. These pairs correspond to pairs that likely correspond to matches. In a final step, the database owners calculate the actual distances between the vector representations of all likely matched pairs using a secure scalar product protocol, and decide whether they are linked based on the computed distance.

Inan et al. 2010 [9] Inan et al. [9] in 2010 proposed a hybrid approach combines differential privacy and cryptographic methods using secure multi-party computation (SMC) techniques to solve the PPRL problem in a two-party protocol. More specifically, a blocking protocol is developed that provides strong data protection compliant with differential privacy; and the pairs not filtered during blocking are compared by using SMC based matching.

Vatsalan et al. 2011 [10] Vatsalan et al. [10] in 2011 proposed an efficient two-party approach for PPRL. Their protocol is based on (1) the use of reference values that are available to both database owners, and allows them to individually calculate the similarities between their attribute values and the reference values; and (2) the binning of these calculated similarity values to allow their secure exchange between the two database owners.

Vatsalan et al. 2012 [11] Vatsalan et al. [11] in 2012 developed a two-party approach based on the use of Bloom filters for approximate private matching. They proposed an iterative classification approach where the database owners iteratively reveal bits from their Bloom filters. At each iteration they calculate the minimum similarity based on the revealed bit positions using the Dice-coefficient, and classify the pairs into matches,

non-matches, and possible matches. The pairs that are classified as possible matches are taken to the next iteration where more bit positions are revealed to classify the pairs. A length filtering method is used to reduce the number of record pair comparisons.

Vatsalan et al. 2013 [12] Vatsalan et al. [12] in 2013 proposed a two-party private blocking technique for PPRL based on sorted nearest neighborhood clustering. Privacy is addressed by a combination of two privacy techniques, i.e., the k -anonymous clustering and public reference values. More specifically, first clusters are generated by the two database owners using a selected set of reference values. Then quasi-identifying attribute values from records are added into these clusters such that each cluster will contain at least k quasi-identifying attribute values (thus providing k -anonymous privacy). Using a sorted nearest neighborhood approach of sliding a window over the reference values, similar clusters are identified and record pairs are generated from all records in the corresponding clusters that are in the same window.

He et al. 2017 [13] He et al. [13] proposed a privacy model based on differential privacy, named output constrained differential privacy, to construct efficient linkage protocols for sensitive databases that offer an end-to-end privacy guarantee for any non-matching record after the matching record pairs have been identified. Following this privacy model, the authors proposed a two-party protocol where two database owners collaborate to identify matching records in their databases. The protocol hides non-matching records by adding Laplace noise to the records in the blocking process.

Chen et al. 2018 [14] Chen et al. in 2018 [14] proposed a record linkage method for two parties. Their basic approach is based on a classical implementation of garbled circuits and a computationally efficient approach using a filtering strategy, which can be readily adopted in small-to-medium scale linkage tasks with a strong security guarantee.

Khurram 2019 [15] Khurram in 2019 [15] introduced an efficient two-part PPRL that runs in sub-quadratic time, provides high accuracy, and guarantees cryptographic security in the semi-honest security model. The security of the scheme is due to the application of a secure two-party computation using binary or arithmetic secret shares, however, the communication cost can be high.

3.2 Related work to two-party PPRL

Song et al. 2000 [16] Song et al. [16] in 2000 presented several cryptographic schemes that enable searching on encrypted data without leaking any information to the untrusted server. The problem can be described as follows: assume that Alice has a set of documents and stores them on an untrusted server Bob. Because Bob is untrusted, Alice wishes to encrypt her documents and only store the cipher text on Bob. Each document can be divided up into ‘words’. Later she wishes to retrieve the documents which contain the word W ; and Bob can determine with some probability whether each document contains the word W without learning anything else.

The proposed techniques for remote searching on encrypted data using an untrusted server and provided proofs of security for the resulting systems. Note that it is the client who encrypts the data and stores the encrypted data on the untrusted server. It is also the client who later performs searches on its own encrypted data. If we considered the PPRL as a search problem on encrypted databases. The difficulty lies in the fact that one party is searching on the database which is encrypted by the other party.

Freeman 2005 [17] Freedman et al. [17] in 2005 presented a privacy-preserving keyword search algorithm, which uses SMC techniques (homomorphic encryption) and oblivious pseudo-random functions. The keyword search problem can be described as follows: suppose that the server holds a database of n pairs (x_i, r_i) , each consisting of a keyword x_i and its payload r_i . The client's input is a search keyword w . If there is a pair where the keyword x_i is equal to the search keyword w (i.e., exact matching), then the corresponding payload r_i will be returned to the client.

This privacy-preserving keyword search algorithm could be applied to design a two-party PPRL for the case that two database holders share a common entity identifier by considering the common entity identifier as keywords. Straightforwardly, the record linkage process can be regarded as the keyword search with multiple queries.

Atallah et al. 2003 [18] Atallah et al. [18] in 2003 proposed a two-party protocol where the edit distance algorithm is modified for providing privacy to sequence approximate comparisons.

Ravikumar et al. 2004 [19] Ravikumar et al. [19] in 2004 used SMC techniques for secure computation of several distance functions. The protocol is developed in the setting of two parties. Note that the use of SMC computations for achieving privacy makes the protocol computationally intensive.

Li et al. 2011 [20] Li et al. [20] in 2011 introduced an approach for privacy-preserving group linkage (PPGL) to measure the similarity of groups of records rather than individuals.

4 Desired properties of two-party PPRL

4.1 Error-tolerant matching

The techniques for linking data involve ways to match pairs of data records based on the value of personally identifying information such as names, birth dates, addresses, and national or local identifying codes, which are not always stable over time and/or can be recorded with errors.

Exact matching by definition does not tolerate any errors in these attributes values. In fact, Winkler [21] reported that 25% of true matches in a US census operation would have been missed by exact matching. To deal with the data quality problem, approximate or error-tolerant matching is desirable.

4.2 An efficient solution to an unbalanced problem

Data linkage is generally a highly unbalanced classification problem because there will be many more non-matching than matching record pairs. For example, assuming two data sets of l records each, where each record refers to one entity, there will be a maximum of l matching but a minimum of $l^2 - l$ non-matching record pairs if no blocking has been applied.

So a solution that could efficiently filter out the non-matches is desired. In fact, such a solution is possible if a unique identifier exists across databases (where a non-match is signaled by any disagreement in a single digit/character between record pairs). However, it is problematic to deal with the multiple error-prone quasi-identifiers since an overall comparison is always needed before a judgment on the matching status of the pairs could be made.

4.3 Consistent matching results

It is known that the PPRL schemes could lead to inconsistent matching results due to non-transitivity [3]. Namely, if record pair (r_1, r_2) is classified as a match and (r_1, r_3) as a match, then the pair (r_2, r_3) should also be a match. However, when the pairs are looked at separately, the third pair might have been classified as a non-match or even not compared at all if blocking was used.

In the two-party PPRL, this may happen when either or both databases contain multiple records that correspond to the same real-world entity. To avoid this problem, both database owners might be required to conduct deduplication before the record linkage process. However, a PPRL scheme is expected to produce consistent results in the general setting, which imply the same set of the linked pairs for both database owners.

4.4 Provable, affordable and comparable privacy

Existing privacy solutions to PPRL are either provably secure but using computation techniques (such as SMC) that is prohibitively expensive in practice, or affordable but vulnerable to attackers, both internal and external. A PPRL solution enjoys both provable and affordable privacy is still missing.

Moreover, we note that for the linkage quality, precision and recall are widely accepted measures to evaluate different linkage techniques. However, for privacy, currently there are no commonly accepted measures, although k -anonymity and differential privacy are employed in the blocking step to attempt to enhance the privacy of PPRL. To develop adequate privacy measures that allow the comparative evaluation of different PPRL techniques so far remains an open problem [22], [23].

4.5 Low communication overhead

It is known that provable security could be provided by employing SMC techniques, however they generally have not only high computational cost but also high communication cost (since large numbers of messages will need to be exchanged

between the parties that participate in such computations). Expensive communication cost makes the protocol impractical for real-world applications, and less scalable for the linkage of very large sensitive databases.

4.6 Scalability

Another key factor that determines the usefulness of a linkage algorithm is whether it can scale to large datasets. In the era of big data, databases with records about many millions or even billions of individuals are not unusual. It is an imperative need to develop novel blocking techniques (to reduce the quadratic complexity of the database size) and efficient matching (both error-tolerant and private) algorithms (with regard to the length of the records) with low communication requirements.

5 Concluding remarks

Although several solutions have been proposed to solve the PPRL problem, no current solution offers a satisfying performance with a provable cryptographic security guarantee while maintaining both high accuracy, and low computational and communication complexity. Besides, a theoretical framework that allows the comparative evaluation of different PPRL techniques is still missing, unlike the scenario for the record linkage problem (without privacy concerns).

References

1. H. Dunn, "Record linkage", *American Journal of Public Health*, **36**(12), 1412, 1946.
2. T. N. Herzog, F. J. Scheuren, and W. E. Winkler, *Data Quality and Record Linkage Techniques*, 1st ed, Springer Publishing Company, Incorporated, 2007.
3. P. Christen, *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Berlin: Springer, 2012.
4. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, **64**(328), pp. 1183–1210, 1969.
5. R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," *BMC Medical Informatics & Decision Making*, **9**, p. 41, 2009.
6. O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game", *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, STOC'87, pp. 218-229, New York, NY, USA, 1987.
7. M. Yakout, M. Atallah, A. Elmagarmid, "Efficient private record linkage", *IEEE ICDE*, pp. 1283-1286, Shanghai, 2009.
8. M. Scannapieco, I. Figotin, E. Bertino, A. Elmagarmid, "Privacy preserving schema and data matching", *ACM SIGMOD*, pp. 653-664, Beijing, 2007.
9. Inan, M. Kantarcioglu, G. Ghinita, E. Bertino, "Private record matching using differential privacy", *EDBT*, Lausanne, Switzerland, pp.123-134, 2010.

10. D. Vatsalan, P. Christen, V. S. Verykios, "An efficient two-party protocol for approximate matching in private record linkage", *AusDM, CRPIT*, **121**, pp.125-136, Ballarat, Australia, 2011.
11. D. Vatsalan, P. Christen, "An iterative two-party protocol for scalable privacy preserving record linkage", *AusDM, CRPIT*, **134**, Sydney, 2012.
12. D. Vatsalan, P. Christen, and V. S. Verykios, "Efficient two-party private blocking based on sorted nearest neighborhood clustering," *CIKM*, San Francisco, CA, USA, 2013.
13. X. He, A. Machanavajhala, C. Flynn, D. Srivastava, "Composing differential privacy and secure computation: A case study on scaling private record linkage", *ACM Conference on Computer and Communications Security*, pp. 1389-1406. Dallas, 2017.
14. F. Chen, X. Jiang, S. Wang, L. M. Schilling, D. Meeker, T. Ong, M. E. Matheny, J. N. Doctor, L. Ohno-Machad, and J. Vaidya, "Perfectly secure and efficient two-party electronic-health-record linkage", *IEEE Internet Comput.*, **22**(2), p. 32-41, 2018.
15. M. Basit Khurram, "SFour: A Protocol for Cryptographically Secure Record Linkage at Scale", *Master Thesis*, 2019.
16. D. Song, D. Wagner, A. Perrig, "Practical techniques for searches on encrypted data", *IEEE Symposium on Security and Privacy*, pp. 44-55, 2000.
17. M. Freedman, Y. Ishai, B. Pinkas, O. Reingold, "Keyword search and oblivious pseudorandom functions", *Theory of Cryptography*, pp. 303-324, 2005.
18. M. Atallah, F. Kerschbaum, W. Du, "Secure and private sequence comparisons", *Workshop on Privacy in the Electronic Society*, ACM, Washington, DC, USA, 2003.
19. P. Ravikumar, W. Cohen, S. Fienberg, "A secure protocol for computing string distance metrics", *Workshop on Privacy and Security Aspects of Data Mining at IEEE ICDM*, Brighton, UK, 2004.
20. F. Li, Y. Chen, B. Luo, D. Lee, P. Liu, "Privacy preserving group linkage", *Scientific and Statistical Database Management*, pp.432-450, Springer, 2011.
21. W.E. Winkler, "Record linkage", *Handbook of Statistics*, D. Pfeffermann, C. Rao (eds.), vol. 29, pp. 351-380, Elsevier, 2009.
22. D. Vatsalan, P. Christen, and V. S. Verykios, "A taxonomy of privacy-preserving record linkage techniques," *Information Systems*, **38**(6), pp. 946-969, 2013.
23. K. Harron, H. Goldstein and C. Dibben, *Methodological Developments in Data Linkage*, John Wiley & Sons Inc., 2015.
24. B. Bloom. "Space/time tradeoffs in in hash coding with allowable errors", *Communications of the ACM*, **13**(7), pp. 422-426, 1970.

Inner Bound of E-capacity-Equivocation Region for the Generalized Wiretap Channel

Mariam Haroutunian

Institute for Informatics and Automation Problems, National Academy of Sciences of Armenia, Yerevan, Armenia armar@sci.am

Abstract. The problem of information theoretic security recently has attracted great attention. One of the problems concerns secure communication over a wiretap channel. The aim in the general wiretap channel model is to maximize the rate of the reliable communication from the source to the legitimate receiver, while keeping the confidential information as secret as possible from the eavesdropper (wiretapper).

We introduce and investigate the E-capacity-equivocation region for the wiretap channel, which is the generalization of the capacity-equivocation region studied by Csiszár and Körner. It is the closure of the set of all achievable rate-reliability and equivocation pairs, where the rate-reliability function presents optimal dependence of rate from error probability exponent (reliability). Previously the outer bound of this region was obtained. Here the inner bound of that region is constructed.

Keywords: Wiretap channel · information-theoretic security · equivocation rate · E-capacity.

1 Introduction

Security is an important topic in communications. The information theoretic security is an approach, that demonstrates the possibility of transmitting confidential messages without using an encryption key. The main idea of the information theoretic security is to exploit the inherent noises and difference between the channels to a legitimate receiver and eavesdropper. In addition, the transmitter intentionally adds randomness to prevent eavesdroppers from accepting useful information while guaranteeing the legitimate receiver to obtain the information. Such an approach to guarantee secrecy has the advantage of eliminating the key management issue, resulting in lower complexity and savings in resources. Such an approach was initiated by Wyner [1], who studied the most basic model called a wiretap channel. Later Csiszár and Körner [2] studied the broadcast channel with confidential messages, the special case of which is the more general model of wiretap channel. It is named as the generalized wiretap channel because the model from [1] is a special case of it when the channel to the eavesdropper is a degraded version of the main channel.

In this paper we consider the generalized model of wiretap channel (see Fig. 1), which is defined as follows.

The source wishes to transmit a message m to a legitimate receiver while keeping it as secret as possible from an eavesdropper. The confidential message m is assumed to be randomly and uniformly distributed over a message set \mathcal{M} . The encoder f_N maps each message m to a codeword $\mathbf{x}(m) = (x_1, \dots, x_N) \in \mathcal{X}^N$, where \mathcal{X} is the input alphabet and N is the transmission length. The codeword $\mathbf{x}(m)$ is transmitted over a discrete memoryless channel (DMC) with transition probability $W(y, z|x)$. The noisy version $\mathbf{y} \in \mathcal{Y}^N$ is accepted by legitimate receiver and $\mathbf{z} \in \mathcal{Z}^N$ by eavesdropper, respectively. The decoder g_N at the receiver maps the received sequence \mathbf{y} to an estimate \hat{m} of the message.

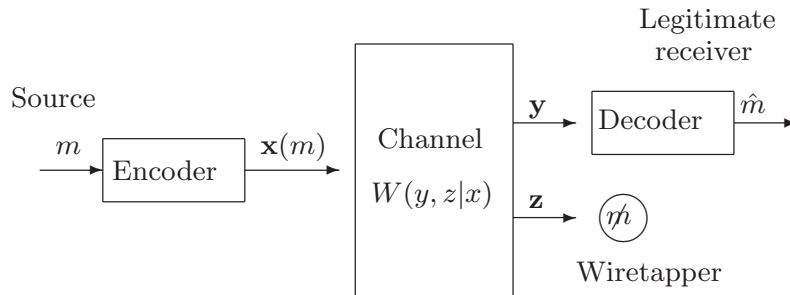


Fig. 1. The model of generalized wiretap channel.

The capacity-equivocation region $\mathcal{C}(W)$ as well as the secrecy capacity $C_s(W)$ of this model were obtained in [2]. Other models with secrecy constraints are surveyed in [3].

We investigate the E - capacity - equivocation region $\mathcal{C}(E, W)$, which is the closure of the set of all achievable rate - reliability - equivocation pairs $(R(E), R_e)$, where the function $R(E)$ presents optimal dependence of rate R from reliability (error probability exponent) E . It is the analogy of E - capacity (rate -reliability function) suggested by E. Haroutunian [4] and investigated for various channel models [5].

The outer bound of E - capacity - equivocation region was constructed in [6]. Here we present the inner bound of this region. When E tends to zero, both bounds coincide with the capacity-equivocation region obtained in [2].

2 Notations, Definitions and Formulation of Results

The DMC $W(y, z|x)$ with finite input alphabet \mathcal{X} , finite output alphabets \mathcal{Y} and \mathcal{Z} is memoryless

$$W^N(\mathbf{y}, \mathbf{z}|\mathbf{x}) = \prod_{n=1}^N W(y, z|x)$$

Let us denote

$$W_1(y|x) = \sum_z W(y, z|x),$$

$$W_2(z|x) = \sum_y W(y, z|x),$$

and

$$P_1 W_1(y|u) = \sum_x P_1(x|u) W_1(y|x). \quad (1)$$

To formulate the problem consider auxiliary random variables U and Q with values in finite \mathcal{U} and \mathcal{Q} , correspondingly, that satisfy the Markov chain relationship: $Q \rightarrow U \rightarrow X \rightarrow (Y, Z)$.

Let the probability distributions (PD) of random variable (RV) U be $P_0 = \{P_0(u), u \in \mathcal{U}\}$ and $P_1 = \{P_1(x|u), x \in \mathcal{X}, u \in \mathcal{U}\}$ be conditional PD of RV X for the given value u . Joint PD of RV U, X we denote by $P_{0,1} = \{P_{0,1}(u, x) = P_0(u)P_1(x|u), u \in \mathcal{U}, x \in \mathcal{X}\}$. and the marginal PD of X is $P = \{P(x) = \sum_u P_{0,1}(u, x), u \in \mathcal{U}, x \in \mathcal{X}\}$. We shall use also the following PD

$$V = \{V(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\},$$

$$P \circ V = \{P \circ V(x, y) = P(x)V(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$$

and

$$PV = \{PV(y) = \sum_x P(x)V(y|x), y \in \mathcal{Y}\}.$$

For N length code (f_N, g_N) the code rate is

$$R(f, g, N) = \frac{1}{N} \log |\mathcal{M}_N|$$

(log and exp functions are taken to the base 2) and the average error probability is

$$e_N(f, g, W_1) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} W_1^N \{\mathcal{Y}^N - g^{-1}(m) | \mathbf{x}(m)\},$$

where $g^{-1}(m) = \{\mathbf{y} : g(\mathbf{y}) = m\}$.

The secrecy level of confidential message m at the wiretapper is measured by the equivocation rate defined as

$$R_e^N = \frac{1}{N} H(M|Z^N),$$

where $H(X|Y)$ is the conditional entropy [7]. In other words, the equivocation rate indicates the eavesdropper's uncertainty about the message m given the channel outputs Z^N . Hence, the larger the equivocation rate, the higher the level of secrecy.

A rate – equivocation pair (R, R_e) is **achievable** if there exists a sequence of message sets \mathcal{M}_N with $|\mathcal{M}_N| = \exp NR$ and encoder – decoder (f_N, g_N) such that the average error probability tends to zero as N goes to infinity, and the equivocation rate R_e satisfies

$$R_e \leq \liminf_{N \rightarrow \infty} R_e^N.$$

The rate – equivocation pair (R, R_e) indicates the confidential rate R achieved at certain secrecy level R_e .

The **capacity - equivocation region** $\mathcal{C}(W)$ is defined to be the closure of the set that consists of all achievable rate – equivocation pairs (R, R_e) .

The following result was obtained in [2] as a special case of a more general result for the broadcast channel with confidential messages.

Theorem 1. *The capacity - equivocation region of wiretap channel is given by*

$$\mathcal{C}(W) = \bigcup_{P_{0,1}W} \left\{ \begin{array}{l} (R, R_e) : Q \rightarrow U \rightarrow X \rightarrow (Y, Z), \\ R \leq I_{P_{0,1}, W_1}(U; Y), \\ 0 \leq R_e \leq R, \\ R_e \leq I_{P_{0,1}, W_1}(U; Y|Q) - \\ -I_{P_{0,1}, W_2}(U; Z|Q), \end{array} \right\} \quad (2)$$

where for generic random variables X and Y , $I(X; Y)$ denotes the mutual information between X and Y [7]. The auxiliary random variables Q and U are bounded in cardinality by $|\mathcal{Q}| \leq |\mathcal{X}| + 3$ and $|\mathcal{U}| \leq |\mathcal{X}|^2 + 4|\mathcal{X}| + 3$, respectively.

From that theorem the following corollary was obtained in [2] on **secrecy capacity**, which is defined as the maximum rate at which the message M can be transmitted while being kept perfectly secret from the eavesdropper.

Corollary 1. *The secrecy capacity of the wire-tap channel is given by*

$$C_s(W) = \max_{P_{0,1}W} [I_{P_{0,1}, W_1}(U; Y) - I_{P_{0,1}, W_2}(U; Z)],$$

where the auxiliary random variable U satisfies the Markov chain relationship: $U \rightarrow X \rightarrow (Y, Z)$, and is bounded in cardinality by $|\mathcal{U}| \leq |\mathcal{X}| + 1$, respectively.

We investigate the E - **capacity - equivocation region** $\mathcal{C}(E, W)$, which is defined as the closure of the set that consists of all E -achievable rate – equivocation pairs $(R(E), R_e)$, $E > 0$ with the average error probability satisfying $e \leq \exp\{-NE\}$. In [6] the following theorem is proved.

Theorem 2. *For $E > 0$, the outer bound for E - capacity - equivocation region of generalized wiretap channel is given by*

$$\mathcal{C}(E, W) \leq \mathcal{R}_{sp}(E, W)$$

with

$$\mathcal{R}_{sp}(E, W) = \bigcup_{P_{0,1}W} \left\{ \begin{array}{l} (R(E), R_e) : Q \rightarrow U \rightarrow X \rightarrow (Y, Z), \\ R(E) \leq \\ \leq \min_{P_1V : D(P_1V || P_1W_1 | P_0) \leq E} I_{P_{0,1}, V}(U; Y), \\ 0 \leq R_e \leq R(E), \\ R_e \leq I_{P_{0,1}, W_1}(U; Y|Q) - I_{P_{0,1}, W_2}(U; Z|Q), \end{array} \right\} \quad (3)$$

where $D(P_1V || P_1W_1 | P_0)$ denotes the divergence between conditional distributions P_1V and P_1W_1 given PD P_0 [7].

Here the following result is obtained.

Theorem 3. *For $E > 0$, the inner bound for E - capacity - equivocation region of generalized wiretap channel is given by*

$$\mathcal{R}_r(E, W) \leq \mathcal{C}(E, W)$$

with

$$\mathcal{R}_r(E, W) = \bigcup_{P_{0,1}W} \left\{ \begin{array}{l} (R(E), R_e) : Q \rightarrow U \rightarrow X \rightarrow (Y, Z), \\ R(E) \leq \\ \leq \min_{P_1V: D(P_1V||P_1W_1|P_0) \leq E} |I_{P_{0,1},V}(U; Y)|^+, \\ D(P_1V||P_1W_1|P_0) - E|^+, \\ 0 \leq R_e \leq R(E), \\ R_e \leq I_{P_{0,1},W_1}(U; Y|Q) - I_{P_{0,1},W_2}(U; Z|Q), \end{array} \right\} \quad (4)$$

where $|a|^+ = \max(a, 0)$.

The proofs are using the method of types [8]. The set of all $\mathbf{u} \in \mathcal{U}^N$ of the type P_0 is denoted by $\mathcal{T}_{P_0}^N(U)$ and $\mathcal{T}_P^N(X|\mathbf{u})$ is the set of all vectors $\mathbf{x} \in \mathcal{X}^N$ with conditional type $P_1(x|u)$ given $\mathbf{u} \in \mathcal{T}_{P_0}^N(U)$.

To prove the Theorem 3 we must show that the rate region specified in (4) is E - achievable for $E > 0$. This is done by constructing a code of length N with certain properties based on the random coding technique.

The proof consists of 2 steps. In step 1 the existence of a code with required properties is proved. In step 2 the estimation of the equivocation rate is given.

For encoding the stochastic encoder f is considered, similar to [2]. It can be considered as a mapping $f : \mathcal{M} \times \mathcal{T} \rightarrow \mathcal{X}^N$, which maps (m, t) to a codeword $\mathbf{x} \in \mathcal{X}^N$, where T is a randomizer, independent of M . We consider the case when the realization of T is unknown to the receiver and eavesdropper. In the case when T is known to the receiver, the model differs, because T serves as a secret key shared by the sender and receiver and the secrecy rate is larger. The code is constructed using the message splitting approach, when the source message is split into two parts. The first part can be decoded by both the receiver and the wiretapper, while the remaining part is only for the legitimate receiver to decode and needs to be kept as secret as possible from the eavesdropper. This rate splitting technique is useful only for the channel models with secrecy constraint.

For decoding we use the divergence minimization criterion suggested by E. Haroutunian [4] and successfully applied for various models [5]. Error probability of the random code constructed with this encoding and decoding strategies is estimated

$$e_N(f, g, W_1) \leq \exp\{-NE\}.$$

The proof is completed by estimation of the equivocation rate.

Corollary 2. *When $E \rightarrow 0$ the inner and outer bounds of E -capacity equivocation region coincide with capacity - equivocation region (2) obtained in [2].*

3 Conclusion and Future Work

A new notion E - capacity - equivocation region of the generalized wiretap channel is investigated, the inner and outer bounds of this region are derived. When $E \rightarrow 0$ this bounds coincide with capacity - equivocation region (2) obtained in [2]. The next step of investigations is to introduce by analogy a new concept of E - secrecy capacity and study it. The bounds of E - capacity - equivocation region and E - secrecy capacity can have simpler forms for some special classes of channels (physically degraded, stochastically degraded, less noisy, more capable), which we will address in future work.

References

1. Wyner, A. D.: The wire-tap channel, Bell System Technical Journal, **54**(8), 1355—1387 (1975)
2. Csiszár, I., Körner, J.: Broadcast channel with confidential messages, IEEE Transactions on Information Theory, **24**(3), 339—348 (1978)
3. Liang, Y., Poor, V., Shamai (Shitz), S.: Information theoretic security, Foundations and Trends in Communications and Information Theory, **5**(4-5), 355–580 (2008)
4. Haroutunian, E.: E-capacity of DMC, IEEE Transactions on Information Theory, **53**(11), 4210–4220 (2007)
5. Haroutunian, E., Haroutunian, M., Harutyunyan, A.: Reliability criteria in information theory and in statistical hypothesis testing, Foundations and Trends in Communications and Information Theory, **4**(2-3), 97–263 (2007)
6. Haroutunian, M.: Outer bound for E-capacity – equivocation region of the wiretap channel, In: 12th International Conference on Computer Science and Information technologies, pp. 129—131. Yerevan, Armenia (2019). Reprint In: IEEE Revised selected papers, pp. 93–95. (2019)
7. Cover, T. M., Thomas, J. A.: Elements of Information Theory. 2nd edn. A Wiley-Interscience Publication, USA (2006)
8. Csiszár, I.: Method of types, IEEE Transactions on Information Theory, **44**(6), 2505—2523 (1998)

A survey on deep semi-supervised learning algorithms

Ani Vanyan, Hrant Khachatryan

YerevaNN

Department of Informatics and Applied Mathematics, Yerevan State University
ani@yerevann.com, hrant@yerevann.com

Abstract. Semi-supervised learning is a branch of machine learning focused on improving the performance of models when the labeled data is scarce, but there is access to large number of unlabeled examples. Recently, there has been a remarkable process in designing algorithms which are able to get reasonable image classification accuracy having access to labels for only 0.5% of the samples on relatively small datasets like CIFAR-10 and SVHN. The downside of these algorithms is that they require expensive tuning of hyperparameters for each dataset, and the hyperparameters tuned for one dataset do not generalize to others. In this work, we survey most of the recently proposed semi-supervised algorithms designed to work in the scope of deep learning. We highlight novelties and problems related to the robustness.

Keywords: Semi-supervised learning.

1 Introduction

In this paper we describe the latest advances in semi-supervised learning, a branch of machine learning focused on improving the performance of an algorithm using a small set of labeled and a large set of unlabeled samples.

These algorithms rely on the premise that obtaining unlabeled data is cheap. Although we have to note that most of these algorithms fail when the distribution of the unlabeled data is different from the distribution of the labeled data. In fact, some methods use special tricks that explicitly require the distributions to be the same.

There is a lot of literature on semi-supervised learning. Most recent papers refer to two classical works for a general overview [2] [22]. In this report we will focus on more recent algorithms which build on top of existing neural network architectures designed for regular supervised learning. These algorithms introduce regularization terms to the loss functions, augment the inputs and perform various kinds of ensembling tricks. This report notably does not cover transductive SVMs (which were a popular method in early 2000s), graph-based methods and methods based on generative models.

Most of the algorithms described in this report are being tested on popular image classification datasets: CIFAR-10, CIFAR-100 [11], SVHN [14] and ImageNet [7]. All these datasets are designed for supervised learning, so to use

them in semi-supervised setup, part of their labels is hidden from the algorithms during the training. On the other hand, the validation sets are usually kept intact, which makes these setups a little bit unrealistic. Some papers also perform experiments on STL-10 dataset [4] which by design has a large subset of unlabeled examples. SVHN also has a special extension called SVHN-extra with 531K additional images.

For the rest of this report, X denotes the labeled dataset with samples $(x, p) \in X$. Here, p is a one-hot vector. U denotes the dataset without the labels. $f_\theta(x)$ is a function (neural network) with parameters θ . It outputs a probability distribution on the labels. $Augment(x)$ is a *stochastic* operation that augments the sample x so that its label remains the same. $H(\cdot, \cdot)$ denotes the cross entropy: $H(p, q) = -\sum_i p_i \log(q_i)$

2 Consistency regularization

The main concept that drives research in semi-supervised learning for the past five years is called consistency regularization. The core idea is to make sure the neural network produces similar results for the augmented versions of the same unlabeled image. It is enforced by an additional term in the loss function:

$$L_U = \frac{1}{|U|} \sum_{x \in U} \|f_\theta(Augment(x)) - f_\theta(x)\|_2^2$$

Note that $Augment(x)$ is a stochastic function, and f_θ might also be stochastic (e.g. due to dropout). So the difference is most likely non-zero.

2.1 Π -model

As far as we know, consistency regularization was first introduced in an algorithm called Π -model [12]. In particular the authors used the following loss function:

$$\begin{aligned} L &= L_X + \lambda(t)L_U \\ L_X &= \frac{1}{|X|} \sum_{(x,p) \in X} H(p, f_\theta(x)) \\ L_U &= \frac{1}{|U|} \sum_{x \in U} \|f_\theta(Augment(x)) - f_\theta(x)\|_2^2 \end{aligned}$$

Here, $\lambda(t)$ is the relative weight of the consistency loss term, which slowly grows from zero to its final value λ over the training process. λ is a hyperparameter. In Π -model, $Augment(x)$ means just two operations:

- Translation by $a \sim Uniform(-2, 2)$ pixels
- Horizontal flip (for all datasets, except SVHN)

2.2 The problem of the unstable target

One critical problem with this formulation of the consistency loss is that it is not stable. This was discovered in the same paper and a partial solution was given. The authors suggested to update one of the terms in the consistency loss (the one with no augmentation) less often and slowly. In particular, they update it once per epoch, and use exponential moving average of the outputs of the snapshots taken at each epoch. This trick is called temporal ensembling.

$$f_{temp.ens}(x) = \alpha f_{temp.ens}(Augment(x)) + (1 - \alpha) f_{\theta}(Augment(x))$$
$$L_U = \frac{1}{|U|} \sum_{x \in U} \|f_{\theta}(Augment(x)) - f_{temp.ens}(x)\|_2^2$$

The first formula is computed once per epoch.

2.3 Mean Teacher

The authors of Mean Teacher algorithm [18], presented in NIPS 2017, gave a better solution to the unstable target problem. They use two separate models: a Student network with θ parameters and a Teacher with θ' parameters. Student is trained as usual. Teacher is not trained via backpropagation. Instead, its weights are updated at each iteration using the weights from the Student network:

$$\theta'_t = \alpha \theta'_{t-1} + (1 - \alpha) \theta_t$$

On unlabeled examples, the Teacher network provides the learning target:

$$L_U = \frac{1}{|U|} \sum_{x \in U} \|f_{\theta}^{student}(Augment(x)) - f_{\theta'}^{teacher}(Augment(x))\|_2^2$$

This work highlighted another problem in the space of semi-supervised learning algorithms. The number of hyperparameters is huge:

- The choice of the neural architecture (backbone)
- Ratio of labeled and unlabeled examples in a batch
- Early stopping criteria
- Decay rate α in exponential moving average formula
- Learning rate schedule
- Weight decay

In such conditions, one way to make comparisons with earlier work more fair is to re-implement the old models in the same codebase. The authors of Mean Teacher re-implemented H -model. Mean teacher consistently worked better than the previous methods.

2.4 Virtual Adversarial Training and Entropy Minimisation

In [13], the authors suggest to change the way images are augmented to be used in consistency loss. Instead of using data-specific augmentation functions, they generate an adversarial example. The idea is similar to the adversarial training method introduced in [9], when the regular loss function is applied to the perturbed version of the input sample:

$$L_{adv} = H(p, f_{\theta}(x + r_{adv}))$$

$$r_{adv} = \arg \max_{r: \|r\| < \epsilon} H(p, f_{\theta}(x + r))$$

Note that r_{adv} can be approximated using Fast Gradient Sign method introduced in the same paper: $r_{adv} = \epsilon \text{sgn}(\nabla_x H(p, f_{\theta}(x)))$. Also note that this operation requires access to the correct label p . For unlabeled examples, we do not have p , so we calculate r_{adv} by taking the perturbation which changes the prediction of the network by the largest magnitude (measured by cross-entropy):

$$r_{adv} = \arg \max_{r: \|r\| < \epsilon} H(f_{\theta}(x), f_{\theta}(x + r))$$

The authors of [13] suggest a fast approximation of this operation.

Another notable difference in VAT and previous methods is that VAT uses cross-entropy instead of Euclidean distance in the consistency loss term.

Another problem with semi-supervised learning methods is the lack of confidence in predictions on unlabeled examples. On the other hand, it is assumed that each unlabeled image belongs to only one class, so the prediction on each image should have low entropy. There are several ways to achieve that, but the first one, as far as we could find, was introduced in the same paper. It adds an additional term into the loss function to minimize the entropy of predictions.

$$L_{ent} = \frac{1}{|X| + |U|} \sum_{x \in X \cup U} H(f_{\theta}(x))$$

$$H(p) = - \sum_i p_i \log p_i$$

3 Evaluation challenges

The number of variables in semi-supervised setups is so large that is increasingly hard to compare different algorithms. In [15], the authors attempted to create a fair comparison setup. In particular, they

- Re-implemented the best known methods in a single code repository
- Fixed the backbone classifier network: WideResNet-28-2 with batch normalization and leaky ReLU
- Fixed the optimizer: Adam with fixed β_1 and β_2

- Used fixed data augmentation and preprocessing strategy, although there are slight differences between SVHN and others. In particular, horizontal flips are not used for SVHN.
- Used equal hyperparameter tuning budget for all algorithms. This is implemented by running 1000 trials of Gaussian-Process-based black box optimization in Google Cloud.

The model selection was performed on the full validation set, which is not a realistic scenario, and it is acknowledged by the authors. Their hyperparameter search resulted in different initial learning rates for Adam optimizer for different methods. Also, in case of VAT, the best value for ϵ (which controls the magnitude of adversarial perturbation) turned out to be different for CIFAR-10 and SVHN.

The authors report the following issues they discovered in their analysis:

1. Fully-supervised baselines are not tuned correctly in many papers. The authors suggest to use the same budget for tuning the hyperparameters of the fully supervised setup. They discover that with more fair experimental setup, the difference between the supervised baselines and the new algorithms is actually lower. The authors also showed that with much stronger regularization it is possible to reach 13.4% error rate on CIFAR-10 with 4000 labels.
2. Transfer learning from (resized) ImageNet is a strong baseline, and it is ignored in most papers. The best result they got from transfer learning (12.09% error) is better than the best result in semi-supervised learning (13.13% error).
3. All models assume that the distribution of unlabeled examples follows the distribution of labeled examples. It is shown that when this assumption does not hold, using unlabeled examples might hurt the performance.
4. To analyze the role of additional unlabeled examples, the authors use SVHN-Extra dataset and monitor the performance on SVHN given different number of unlabeled examples. They show that some methods get worse performance when exposed to too many unlabeled examples. The authors also analyze the effect of the number of labeled examples.
5. Finally, the authors show that in a more realistic treatment of validation data, when the size of the validation set is just 10% of the (labeled) training set, then it is not feasible to reliably distinguish between strongly and weakly performing models.

4 Multi-stage algorithms

The paper described in the previous section had some positive impact on further research in semi-supervised learning. Almost all subsequent papers working on CIFAR-10 and SVHN used the same underlying architecture, most authors re-implemented previous best models in the same codebase. Unfortunately, this is still not the case with experiments on ImageNet, experimental setups still vary a lot. Also, the authors still continue to use full validation sets for model selection and for comparing different algorithms.

A number of papers introduced more complicated algorithms with multiple stages to beat the state-of-the-art for semi-supervised learning on CIFAR-10, SVHN, ImageNet and others. In this section we cover the most notable ones.

4.1 MixMatch

MixMatch is an algorithm that combines many ideas, including consistency regularization, exponential moving average of network weights and a special trick for obtaining new unlabeled examples called MixUp.

MixUp, introduced in [21], is a way to construct new samples by taking a convex combination of existing samples. For each pair of samples (x_1, p_1) and (x_2, p_2) , MixUp performs the following steps:

1. Sample $\lambda \sim \text{Beta}(\alpha, \alpha)$
2. $\lambda' = \max(\lambda, 1 - \lambda)$ to make sure it's close to 1
3. $x' = \lambda'x_1 + (1 - \lambda')x_2$
4. $p' = \lambda'p_1 + (1 - \lambda')p_2$
5. Return (x', p')

The second step was introduced in MixMatch and makes sure that the samples from $\text{MixUp}(A, B)$ are “closer” to A .

The semi-supervised learning algorithm used in MixMatch paper is essentially the same as in other papers, except there is an additional stage of modifying both the labeled and unlabeled sets. This stage is called MixMatch.

$$\begin{aligned}
 X', U' &= \text{MixMatch}(X, U, T, K, \alpha) \\
 L_X &= \frac{1}{|X'|} \sum_{(x', p') \in X'} H(p', f_\theta(x')) \\
 L_U &= \frac{1}{|U'|} \sum_{(x', q') \in U'} \|q' - f_\theta(x')\|_2^2 \\
 L &= L_X + \lambda(t)L_U
 \end{aligned}$$

In short, $\text{MixMatch}(X, U)$ function applies MixUp to both labeled and unlabeled examples, and uses the average prediction of multiple augmented versions of the same unlabeled image. As taking average might reduce the entropy in the predicted distribution, the authors perform an additional step of sharpening the probabilities with temperature T which is another hyperparameter.

The authors attempt to follow the setup used in [15]. They note, that the best values for two of the new hyperparameters they have introduced do not vary in the datasets they have tested on: sharpening temperature is always set to $T = 0.5$ and the number of augmentations performed on the same unlabeled image is always $K = 2$. Instead, the best values for α hyperparameter of the Beta distribution used in MixUp and the coefficient λ in the main loss function are different for CIFAR versions and SVHN.

To make evaluations more stable, they use an exponential moving average of the model parameters with a decay rate of 0.999 when evaluating on the

validation set. They report state-of-the-art results on all benchmarks. This is the first paper which tests the performance of SSL algorithms on CIFAR-10 with only 250 labels.

4.2 ReMixMatch

The team behind MixMatch made their algorithm even more complicated by adding two more components. The resulting system, called ReMixMatch [1], will be presented at ICLR 2020 conference. The two main additions are:

1. Distribution alignment. The predicted probabilities on a batch of unlabeled examples are scaled to match the distribution of the labels present in the labeled subset. This allows to get significantly higher accuracies in CIFAR-100 with very limited labeled examples. In practice, the scaling coefficients are estimated using a running average of 128 batches.
2. Anchored augmentation. The target label (or probability distribution over labels) for unlabeled examples is determined using *weakly* augmented versions of the images, while the prediction for the same images is computed by using *strongly* augmented versions. Weak augmentation is the same augmentation used in previous works. Strong augmentation used in ReMixMatch is called CTAugment. CTAugment uniformly samples transformations from Python Image Library to apply to the images (similarly to RandAugment) but dynamically infers magnitudes for each transformation during the training process. Since CTAugment does not need to be optimized on a supervised proxy task and has no sensitive hyperparameters, it can directly be included in semi-supervised models to experiment with more aggressive data augmentation. Intuitively, for each augmentation parameter, CTAugment learns the likelihood that it will produce an image which is classified correctly. Using these likelihoods, CTAugment then only samples augmentations that fall within the network tolerance.

There are few other tricks, like using a self-supervised loss of predicting the rotation angle (idea borrowed from S4L model, see Section 4.4). ReMixMatch shares values with MixMatch for multiple hyperparameters, but notably, the number of strongly augmented samples used in the consistency loss term is changed from $k = 2$ to $k = 8$. In addition to the experimental setups used in previous papers, the authors report performance on CIFAR-10 with only 40 labels, although they mention that they had to change one hyperparameter to make their model work in that setup: the coefficient for the loss term responsible for rotation prediction.

4.3 Methods based on Contrastive Predictive Coding

In [16], a novel method for unsupervised representation learning was introduced. It is based on the so called InfoMax principle, which attempts to maximize mutual information between representations of different “views” of the image

(usually defined as various patches extracted from the same image). The representation is trained by predicting the representation of the closest patch using a contrastive loss, inspired by ideas from metric learning. The method is called Contrastive Predictive Coding. The representations learned using this method act as high quality features for downstream tasks. The model has many technical details, e.g. using PixelCNN for aggregating representations of patches.

[10] extends this model with several tricks (e.g. layer normalization, random flipping of patches, etc.) and applies it to semi-supervised setups. In particular, they learn an unsupervised representation based on contrastive predictive coding using all images from ImageNet dataset, and then use the labeled subset to train another classifier on top of the learned representations. Note that the classifier is not a linear shallow model, it is another ResNet. This setup is pretty hard to compare with other semi-supervised models. The authors report results with various percentages of labeled examples. There are two commonly used benchmarks for semi-supervised setup for ImageNet: with 1% labeled data (10 samples per class) and with 10% labeled data. CPCv2 is applied to both setups.

This paper was submitted to ICLR 2020, but was rejected. Later, another paper critically analyzed the results obtained using CPC and similar methods based on maximizing mutual information, and concluded that the experimental successes presented in those papers are mostly due to similarities of these methods to deep metric learning (the triplet loss, hard negative mining etc.) and not because of the quality of mutual information maximization [19]. Some progress in this direction is reported in [3].

4.4 Semi-supervised Self-supervised Learning

Another complicated and multi-stage algorithm was described in [20], published in ICCV 2019. They suggest to integrate self-supervised learning techniques into semi-supervised learning. In particular, they focus on two known self-supervised learning methods:

1. Rotation. Each unlabeled image is rotated by 90, 180 and 270 degrees and along with the original one are given to a classifier which attempts to predict the rotation angle (4-class classification). The classifier has a ResNet backbone, so it learns to extract useful features.
2. Exemplar. Two augmented versions of the same image are passed through the classifier and the learned representations are trained to be similar. Triplet loss is used to avoid collapse of representations.

These methods produce representations without using any labels. The new method suggested in this paper, called semi-supervised self-supervised learning (S4L) adds a regular classification loss term to the loss function, which is computed only for labeled examples. At each iteration, two equal sized batches are sampled: one from the set of labeled examples, another one from the set of unlabeled examples. The loss for rotation prediction or exemplar is computed either on unlabeled batch only, or on both batches. This choice does not affect the final performance of these models.

In the last part of the paper the authors describe a three-stage system which brings state-of-the-art results on ImageNet:

1. Train a semi-supervised model using Virtual Adversarial Training (along with entropy minimisation) with an additional classifier to predict rotation of the image.
2. Use the model obtained from the first stage to generate pseudo labels for all images of ImageNet. The labels are generated by taking the average of predictions across five random crops and four rotations of the same image. Train the same algorithm on the dataset using the predicted labels. Initialize the weights from the network obtained in the first stage, and then train for 18 epochs while decaying the learning rate after 6th and 12th epochs.
3. Fine tune the model obtained in the second stage by using only the original labels. This step is trained with weight decay $3 \cdot 10^{-3}$ and learning rate $5 \cdot 10^{-4}$ for 20 epochs. Learning rate is decayed 10x every 5 epochs.

The resulting model is called MOAM (mix of all models). The number of design choices made in MOAM make it impractical to use for other datasets. Still, its results were state-of-the-art as of January, 2020.

5 Current State-of-the-art: Back to Basics

Another branch of research in semi-supervised learning considers relatively simple, single-stage models. By changing a few details from previous approaches, these papers reach new state-of-the-art results.

5.1 Unsupervised Data Augmentation

Unsupervised Data Augmentation (UDA) is a new model quite similar to VAT, but replaces virtual adversarial example generation with a very strong augmentation. In particular, they use RandAugment [6], which at the time was the strongest data augmentation method known for CIFAR datasets. RandAugment is inspired by AutoAugment [5]. AutoAugment uses a search method to combine all image processing transformations in the Python Image Library (*PIL*) to find a good augmentation strategy. In RandAugment, search is not used, instead the augmentations are uniformly sampled from the same set of transformations in PIL. Basically, RandAugment is simpler and requires no labeled data as there is no need to search for optimal policies. It is important to note, that it is not obvious how RandAugment should be configured for other datasets. As with many other models, the branch of the network which guesses the label on the non-augmented version of the image uses a fixed copy of weights and does not pass the gradient through.

Additionally, UDA uses a training technique, called Training Signal Annealing (*TSA*), to reduce overfitting when there is a huge gap between the amount of unlabeled data and that of labeled data. TSA gradually releases the “training signals” of the labeled examples as training progresses. It utilizes a labeled

example if the model’s confidence on that example is lower than a predefined threshold, which increases according to a schedule. The threshold for the confidence is increased during the training by one of the three rules: logarithmic, linear and exponential.

UDA is tested on CIFAR-10 and SVHN, but also on several sentence classification tasks. To perform data augmentation on sentences, they used back-translation: translated each sentence into French and back into English using an existing machine translation model. Although they reported state-of-the-art results on almost all benchmarks, this paper was also rejected from ICLR 2020 due to the lack of novelty.

5.2 FixMatch

FixMatch [17] is the most recent algorithm in the line of relatively simple algorithms and maintains the state-of-the-art scores for almost setups of CIFAR-10 and SVHN. Similar to UDA, it uses weakly augmented version of an image to guess the label, and forces the network to output the same label on a strongly augmented version of the same image. FixMatch uses CutOut [8] along with RandAugment or CTAugment as a strong augmentation procedure known from previous papers (UDA and ReMixMatch, respectively). In contrast to UDA and other methods, FixMatch performs $\arg \max$ on the guessed label, so it essentially becomes equivalent to pseudo-labeling. Additionally, FixMatch ignores the guessed labels if the confidence is lower than $\tau = 0.95$ threshold.

Unlike MixMatch, FixMatch does not change the λ weight of the consistency loss term during the training. The thresholding operation likely compensates for that. λ is always set to 1. FixMatch uses SGD with a cosine annealing schedule. It does not use MixUp, sharpening or distribution alignment.

The paper does extensive analysis on the role of various components. The outcomes of their analysis might be helpful in subsequent research:

1. Stochastic gradient descent with momentum performs better than Adam optimizer
2. Nesterov momentum is not significantly better than the regular momentum
3. Weight decay is extremely important. Changing weight decay value by 10x might result in 10% absolute increase in error rate.
4. Sharpening instead of pseudo-labeling does not significantly change the results, so pseudo-labeling is chosen for simplicity.
5. The batch size for unlabeled data is set to be $\mu = 8$ larger than the one for labeled data. $\mu < 8$ results in worse performance. $\mu > 8$ does not improve the performance.
6. Cosine decay of learning rate performs better than no decay. The difference between linear and cosine decay schemes is not significant.
7. ImageNet requires a completely different set of hyperparameters (See Appendix C of [17]). The experiments are performed with batch size 1024 for labeled and 5120 for unlabeled samples.

Finally, FixMatch reports performance on an extremely low label scenario which the authors call “barely supervised learning”. When only one image is available per class, the performance of FixMatch strongly depends on the choice of that single image. They show that when the samples are “representative” of the class, the performance can reach 80% using only 10 labels (one per class). On the other hand, if the samples are poorly chosen, the accuracy is below 40%. The “representativeness” is measured by a technique which requires all available labels, so this paper does not suggest an automatic method to determine the best samples.

6 Conclusion

In this paper we have reviewed the recent developments in semi-supervised learning algorithms designed for image classification tasks. Most of the methods are based on consistency regularization. Few methods attempt to combine it with ideas from unsupervised representation learning and metric learning. Despite the significant progress in terms of accuracy on a couple of benchmarks, the methods still suffer from severe instability with respect to hyperparameters and fail when the distribution of unlabeled samples is shifted. We expect the future work in this area to focus on making the methods more practical in real-world applications. On the other hand, the methods developed for image classification are expected to be transferred to other tasks and domains.

References

1. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 / accepted at ICLR’2020 (2019)
2. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press (2006)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
4. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223 (2011)
5. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 113–123 (2019)
6. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical data augmentation with no separate search. arXiv preprint arXiv:1909.13719 (2019)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

10. Henaff, O.J., Srinivas, A., Fauw, J.D., Razavi, A., Doersch, C., Eslami, S.M.A., van den Oord, A.: Data-efficient image recognition with contrastive predictive coding (2020), <https://openreview.net/forum?id=rJerHlrYwH>
11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
12. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. ICLR (2017)
13. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **41**(8), 1979–1993 (2018)
14. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
15. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: *Advances in Neural Information Processing Systems*. pp. 3235–3246 (2018)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
17. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685* (2020)
18. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *Advances in neural information processing systems*. pp. 1195–1204 (2017)
19. Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M.: On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, accepted at ICLR 2020 (2019)
20. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1476–1485 (2019)
21. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017)
22. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* **3**(1), 1–130 (2009)

On Machine Learning Powered Theorem Prover for Propositional Fragment of Minimal Logic

Ashot Baghdasaryan¹ and Hovhannes Bolibekyan²

¹ Russian-Armenian University, Yerevan, Armenia
baghdasaryana95@gmail.com

² Yerevan State University, Yerevan, Armenia
bolibekhov@ysu.am

Abstract. There are three main problems for theorem proving with a standard cut-free system for the propositional fragment of minimal logic. The first problem is the possibility of looping. Secondly, it might generate proofs which are permutations of each other. Finally, during the proof some choice should be made to decide which rules to apply and where to use them. In order to solve the rule selection problem, recurrent neural networks are deployed and they are used to determine which formula from the context should be used on further steps. As a result, it yields to the reduction of time during theorem proving.

Keywords: Automated theorem prover · minimal logic · loop detection · recurrent neural network.

1 Introduction

The sequent system GM^- for minimal logic was introduced in [1]. GM^- is a permutation-free sequent system; it avoids the problems of permutations in the cut-free sequent system of Gentzen. GM^- partly addresses the looping problem and hence is advantageous as a system for theorem proving. However, the naive implementation of GM^- will lead to the possibility of looping. Some looping mechanisms have been considered earlier in [2], [3], [4].

In this paper following [2] one type of history mechanism is considered. In this system the problem of looping is removed, but the problem of rule selection remains unsolved. There is a stoup selection rule, when a formula from the context should be selected to be considered as a stoup.

There are different kind of systems in which rule selection problem leads to proof search inefficiency issues. Because of that problem automated theorem provers based on that systems experience some difficulties. [5] and [6] show some approaches of applying machine learning methods in the field of automated theorem proving.

2 Loop Detection

Further in the text we follow well known definitions of a formula, sequent, proof, context, stoup, equivalence of the systems as in [4], [7].

$$\frac{A, \Gamma \Rightarrow B; \epsilon}{\Gamma \Rightarrow A \supset B; H} (\supset R_1) \quad \text{if } A \notin \Gamma \qquad \frac{\Gamma \Rightarrow B; H}{\Gamma \Rightarrow A \supset B; H} (\supset R_2) \quad \text{if } A \in \Gamma$$

$$\frac{A, \Gamma \Rightarrow \perp; \epsilon}{\Gamma \Rightarrow \neg A; H} (\neg R_1) \quad \text{if } A \notin \Gamma \qquad \frac{\Gamma \Rightarrow \perp; H}{\Gamma \Rightarrow \neg A; H} (\neg R_2) \quad \text{if } A \in \Gamma$$

$$\frac{\Gamma \Rightarrow A; (C, H) \quad \Gamma \xrightarrow{B} C; H}{\Gamma \xrightarrow{A \supset B} C; H} (\supset L) \quad \text{if } C \notin H$$

$$\frac{\Gamma \Rightarrow A; (C, H) \quad \Gamma \xrightarrow{\perp} C; H}{\Gamma \xrightarrow{\neg A} C; H} (\neg L) \quad \text{if } C \notin H$$

$$\frac{\Gamma \xrightarrow{A} C; H}{\Gamma \xrightarrow{A \wedge B} C; H} (\wedge L_1) \qquad \frac{\Gamma \xrightarrow{B} C; H}{\Gamma \xrightarrow{A \wedge B} C; H} (\wedge L_2)$$

$$\frac{\Gamma \Rightarrow A; H \quad \Gamma \Rightarrow B; H}{\Gamma \Rightarrow A \wedge B; H} (\wedge R)$$

$$\frac{A, \Gamma \Rightarrow C; \epsilon \quad B, \Gamma \Rightarrow C; \epsilon}{\Gamma \xrightarrow{A \vee B} C; H} (\vee L) \quad \text{if } A, B \notin \Gamma$$

$$\frac{\Gamma \Rightarrow A; H}{\Gamma \Rightarrow A \vee B; H} (\vee R_1) \qquad \frac{\Gamma \Rightarrow B; H}{\Gamma \Rightarrow A \vee B; H} (\vee R_2)$$

$$\frac{A, \Gamma \xrightarrow{A} B; H}{A, \Gamma \Rightarrow B; H} (C)^* \qquad \frac{\Gamma \Rightarrow A; (A, H)}{\Gamma \xrightarrow{\perp} A; H} (\perp) \qquad \frac{}{\Gamma \xrightarrow{A} A; H} (ax)$$

* B is either a propositional variable, \perp or a disjunction.

A, B, C are formula. Γ , H are sets of formula.

B, Γ is shorthand for $\{B\} \cup \Gamma$.

Fig. 1. The propositional system SwMin

One way to prevent loops is to add a history to each sequent. The history is the set of all sequents that have occurred so far in a proof tree. After each backwards inference the new sequent (without its history) is checked to see whether it is a member of this set. If it is we have looping and we backtrack. If not the new history is the union of the new sequent (without its history) and the old history, and we try to prove the new sequent, and so on.

The approach requires lots of sequents to be stored and on every step the list should be used for specific checkings. All that is quite inefficient as the sequents being stored contain much more information than actually needed to proceed. To prevent looping we can keep few information and satisfy the requirements.

The main idea behind to reduce the history and check the loops is the fact that only goal formula need to be stored. The rules of GM^- are such that the context cannot decrease; once a formula is in the context it will remain in the context of all sequents above it in the proof tree. For two sequents to be the same they obviously need to have the same context. We may empty the history every time the context is extended, since we will never get any of the sequents below the extended one again. Goal formulas are the only ones to be stored in the history. If we come across a goal already in the history we have the same goal and the same context as another sequent, that is, a loop.

There are two slightly different approaches to doing this. There is the straightforward extension and modification of the system which we shall call a *SwMin*, and there is an approach which involves storing more formula in the history, but that detects loops more quickly. This we will call as *ScMin*, and the implementation is in some cases more efficient than the *SwMin*.

In scope of considered systems sequent $\Gamma \Rightarrow C; H$ has context Γ , goal C , history H and no stoup, and sequent $\Gamma \xrightarrow{A} C; H$ has context Γ , goal C , history H and stoup A . When the history has been extended we have parenthesised (C, H) for emphasis, while by ϵ we denote empty history. The *SwMin* system is displayed in Figure 1, and the *ScMin* system in [2].

Theorem 1. [2] *The system SwMin and GM^- are equivalent.*

The idea is focused on constructing proof trees based on a given pattern in a first system and considering inference rule under the focus. Detailed proof can be found in [2].

3 Rule Selection

In *SwMin* the problem of rule selection remains unsolved. There is a stoup selection rule, when a formula from the context should be selected to be considered as a stoup. Though this is inefficient as it requires many branches to prove, which may be unnecessary. We developed prover *SwProv* based on *SwMin* system. To avoid the rule selection problem neural networks are deployed in the *SwProv* prover (*SwNNProv*), which helps us to make "right" decisions. At each step of the proof, if there are multiple choices of the inference rule to be applied at the current step, neural network is used to determine which formula from the context will become a stoup.

3.1 Sequent To Vector Transformation

Firstly, all formulas in sequents are represented in prenex normal form. To be able to use neural networks in the proof search it is necessary to train network model against provable sequents. To proceed with that we introduce numerical representation for the sequents assigning a specific number to each symbol. Based on that representation similar formulas will get identical vectors.

Autoencoders [8] are trained to get fixed length encoding for each sequent. Two slightly different approaches are experimented. In the first one (AE) under-complete autoencoder is used in order to get encoding of the numerical representations of the sequents using L2 loss function:

$$Loss_{AE} = \frac{1}{2N} \sum_{i=1}^N (x^{(i)} - \hat{x}^{(i)})^2, \quad (1)$$

where N is the number of examples, x is the numerical representation of the sequent, while \hat{x} is the output of the autoencoder.

Second approach (CAE) differs from the first one in terms of the loss function. As representations of the similar sequents are close to each other, hidden layer of the autoencoder has to be less sensitive to the inputs changes. So the intuition of the contractive autoencoders [9] is used to modify the loss function:

$$Loss_{CAE} = \frac{1}{2N} \sum_{i=1}^N (x^{(i)} - \hat{x}^{(i)})^2 + \lambda \frac{1}{N} \sum_{i=1}^N \left\| \nabla_{x^{(i)}} h(x^{(i)}) \right\|_F^2, \quad (2)$$

where

$$\left\| \nabla_{x^{(i)}} h(x^{(i)}) \right\|_F^2 = \sum_{j=1}^M \sum_{k=1}^L \left(\frac{\partial h(x^{(i)})_j}{\partial x_k^{(i)}} \right)^2 \quad (3)$$

h is the representation of the hidden layer, N is the number of examples, M is the size of hidden layer, L is the size of input layer, x is the numerical representation of the sequent, \hat{x} is the output of the autoencoder, λ is the hyper-parameter that controls the strength of the regularization.

So, $\nabla_{x^{(i)}} h(x^{(i)})$ is the jacobian of the encoder and the minimisation of the Frobenius norm of the jacobian matrix, which is the sum squared over all elements inside the matrix, results the encoder to keep only the "useful" information. The comparison between two approaches and the retention quality each of the methods shown in Figure 2.

As a result we get numerical representation (encoded vectors) for the sequents.

3.2 Recurrent Neural Networks in Proof Search

Standard library of propositional logic problems is used as a training dataset. More than 100000 formulas are generated with the help of 5000 predefined formulas of minimal logic. Some formulas are added to the dataset from the TPTP problem library [10] too. For each element in training set *SwProv* prover is run

and training examples are generated. At each point of proof tree, where a stoup formula has to be selected, all sequents in that branch of tree are considered and sequence of vectors is generated by "Sequent to Vector" transformation. To differentiate successful outcomes while training neural network one needs to take numerical representation for each stoup candidate formula and corresponding ground truth label (whether this is the right selection).

Two types of recurrent neural network are deployed, where the first one consists of gated reclified unit (GRU) [11] as recurrent layer and 2-dense layers with skip connections [12], while the second one consists of LSTM [13] layer and 2-dense layers with skip connections. The output of recurrent layer (feature vector extractor module) is concatenated with numerical representation of stoup candidate and then is mapped into 2-length one-hot encoded vector via dense layer with softmax activation function. As a final step cross entropy is used as a loss function:

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i * \log(p(y_i)) + (1 - y_i) * \log(1 - p(y_i)), \quad (4)$$

where y is the label and $p(y)$ is the predicted probability of the candidate formula being right for all N examples.

Method		Precision	Recall	F1	Accuracy
AE	Retention Loss	0.7	0.74	0.72	72%
AE	0.015	0.7	0.75	0.72	73%
CAE	0.014	0.71	0.75	0.73	73%
CAE + GRU		0.72	0.78	0.75	75%
CAE + LSTM					

Fig. 2: Autoencoders retention comparison.

Fig. 3: Quality metrics for different approaches of the problem.

As it is more important not to drop the right subtrees (misclassify as 0) rather than keep more erroneous subtrees (misclassify as 1), loss function is changed by adding penalization term for recall, and also prediction threshold is changed to get higher value of recall. Figure 3 shows quality metrics for different approaches calculated on the testing dataset. It is obvious, that LSTM network with contractive autoencoder compression is the best one with 75% accuracy and 0.78 recall.

As a result, the unnecessary branches of the proof tree are removed in almost 75% of cases, which leads to the reduction of time during automated theorem proving.

3.3 Inference

In order to reduce processing time during automated theorem proving, recurrent neural networks are deployed. As the network inference is computationally very expensive and the main goal was to reduce consumed time during rule selection, some inference reductions and benchmarkings are done. NVIDIA TensorRT is

applied to the model for getting the benefits like layer/tensor fusion and automatic precision calibration. It turns out, that using float16 instead of float32 during inference does not significantly affect to the accuracy, while it speeds up the inference. Figure 4 shows how many inferences per second can be done for different GPUs with different architectures, also with and without using TensorRT and with float32 or float16 arithmetics.

GPU	Architecture	without TensorRT (inf/sec)	TensorRT float32 (inf/sec)	TensorRT float16 (inf/sec)
Nvidia P100	Pascal	870	1200	1400
Nvidia K80	Kepler	250	350	400
Nvidia V100	Volta	1100	1600	3100
Nvidia T4	Turing	740	1000	1900
Nvidia Jetson Nano	Maxwell	50	65	70

Fig. 4: Inference benchmark.

4 Results

In result of constructing new proof systems for propositional fragment of the minimal logic and deploying concept of neural network in the prover experiments revealed proof search space reduction and the level of accuracy up to 75% training 150 epochs.

Example	SwProv	SwNNProv
$(A \wedge \neg A) \supset B$	2.1	1.5
$(A \supset B) \supset (A \supset C) \supset (A \supset (B \supset C))$	2.6	1.2
$((\neg\neg A \supset A) \supset A) \vee (\neg A \supset \neg A) \vee (\neg\neg A \supset \neg\neg A) \vee (\neg\neg A \supset A)$	4.4	4.6
$\neg\neg((\neg A \supset B) \supset (\neg A \supset \neg B) \supset A)$	42	25
$(((((A \supset B) \supset ((B \supset C) \supset (A \supset C)))) \supset C) \supset ((B \supset C) \supset (((((A \supset B) \supset ((B \supset C) \supset (A \supset C)))) \vee B) \supset C)))$	37	16
$(((((A \supset B) \supset ((C \supset B) \supset ((A \vee C) \supset B)))) \supset C) \supset (((((A \supset B) \supset ((C \supset B) \supset ((A \vee C) \supset B)))) \supset \neg C) \supset \neg(((A \supset B) \supset ((C \supset B) \supset ((A \vee C) \supset B))))))$	129	15
$((((G \supset A) \supset J) \supset ((P \vee (Q \& P)) \supset P) \supset E) \supset (((H \supset B) \supset I) \supset C \supset J \supset (A \supset H) \supset F \supset G \supset (((C \supset C) \supset I) \supset ((P \vee (Q \& P)) \supset P)) \supset (A \supset C) \supset (((F \supset A) \supset B) \supset I) \supset E)$	869	174
$(((((G \supset A) \supset J) \supset D \supset E) \supset (((H \supset B) \supset I) \supset C \supset J \supset (A \supset H) \supset F \supset G \supset (((C \supset B) \supset I) \supset D) \supset (A \supset C) \supset (((F \supset A) \supset B) \supset I) \supset E)) \& B) \supset (((G \supset A) \supset J) \supset D \supset E) \supset (((H \supset B) \supset I) \supset C \supset J \supset (A \supset H) \supset F \supset G \supset (((C \supset B) \supset I) \supset D) \supset (A \supset C) \supset (((F \supset A) \supset B) \supset I) \supset E))$	1359	96

Fig. 5: Proving time comparison.

Compared to the prover without neural network time spent for the proof is reduced for almost twice. Figure 5 shows the comparison between SwProv (based on SwMin system, with the rule selection problem) and SwNNProv (LSTM network powered prover) automatic theorem provers. In complex formulas SwNNProv obviously is performing much efficient and faster.

5 Conclusion and Future Work

In this paper, three main problems of theorem proving with a gentzen-style cut-free system of minimal logic are considered. The main contribution of this work is to solve the rule selection problem, in the way of expressing it as a machine learning problem and proposing methods for solving it based on recurrent neural networks. A discussion on different representations of sequents has been provided. In particular, representations using two types of autoencoders have been proposed. A new approach based on different types of recurrent neural networks was introduced for solving the rule selection problem. Our contribution in this part was to adapt these algorithms to the automated theorem provers for reducing the proof tree, which to our knowledge have never been addressed before. Also, some optimizations and benchmarkings are done in order to have a faster model during inference.

From an experimental point of view, our contribution lies in the comparison of models for rule selection using two types of autoencoders with LSTM and GRU cells. These experiments were performed for the different types of minimal logic formulas and it's superpositions. Results show that our approach obtains good results and it reduces the proof time for almost twice.

Future research should be devoted to the development of new types of machine learning models (bidirectional RNNs, attention mechanisms) and to the training of new models based on enriched dataset of minimal logic formulas. Regardless, future research could continue to explore the first-order and modal fragments of minimal logic.

References

1. Bolibekyan H.R., Chubaryan A.A.: On the sequent systems of weak arithmetics, Doklady National'noy Akademii Nauk RA, vol. 102, N3, pp. 214-218 (in Russian), (2002)
2. H.R. Bolibekyan, A.R. Baghdasaryan: On some systems of propositional minimal logic with loop detection, Reports of NAS RA, Volume 119, Number 2, pp. 110-115 (2019)
3. Bolibekyan H., Baghdasaryan A.: On some systems of minimal predicate logic with history mechanism, The Bulletin of Symbolic Logic, Volume 24, Number 2, pp. 232-233 (2018)
4. Howe J.M.: Two Loop Detection Mechanisms: a Comparison., Springer Lecture Notes in Artificial Intelligence, Volume 1227, pp. 188–200. (1997)
5. Bridge, J.P., Holden, S.B., Paulson, L.C.: Machine Learning for First-Order Theorem Proving. J Autom Reasoning 53, 141–172 (2014)

6. C. Kaliszyk, F. Chollet, C. Szegedy: HolStep: A machine learning dataset for higher-order logic theorem proving. ICLR (2017)
7. Kleene S.C.: Introduction to metamathematics., D. Van Nostrand Comp., Inc., New York-Toronto, (1952)
8. P. Baldi: Autoencoders, Unsupervised Learning, and Deep Architectures, (2012)
9. S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio: Contractive Auto-Encoders: Explicit Invariance During Feature Extraction, ICML'11: Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 833–840, (2011)
10. G. Sutcliffe: The TPTP Problem Library and Associated Infrastructure. From CNF to TH0, TPTP v6.4.0, Journal of Automated Reasoning, Volume 59, Number 4, pp. 483-502, (2017)
11. K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio: On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, (2014)
12. P. Fischer, O. Ronneberger, T. Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, (2015)
13. M. Sundermeyer, R. Schluter, H. Ney: LSTM neural networks for language modeling. INTERSPEECH, pp. 194–197, (2012)

Estimating Efficient Sampling Rates of Metrics for Training Accurate Machine Learning Models

Tigran A. Bunarjyan¹, Ashot N. Harutyunyan¹, Arnak V. Poghosyan¹,
A.J. Han Vinck², Yanling Chen², and Narek A. Hovhannisyan³

¹VMware Eastern Europe, ²University of Duisburg-Essen, ³TeamViewer Armenia
{tbunarjyan;aharutyunyan;apoghosyan}@vmware.com,
{han.vinck;yanling.chen}@uni-due.de,
narek.hovhannisyan@teamviewer.com

Abstract. Cloud management solutions provide full real-time visibility into modern software-defined data centers (SDDC) of high complexity and sophistication through measuring millions of indicators with increasingly high sampling rate. This high frequency monitoring of metrics allows capturing the expected ever-growing dynamism of business-critical applications resulting in huge bases of time series data to be stored for analysis, pattern detection, and training predictive/forecasting models. That causes high analytics overhead and product performance issues. Therefore, identifying optimal sampling rates of time series data subject to preserving their main information content could mitigate this issue. A particular use case is tuning the sampling rates to be efficient for training ML models accurate enough in analytics tasks, such as anomaly detection. In this paper, we analyze a large collection of cloud application metrics and show that the sampling rate can be substantially reduced with a small information divergence. Moreover, we show that those anomaly detection modules perform sufficiently/tolerably accurate for the reduced data sets.

Keywords: Time series, sampling rate, information loss, information divergence, forecasting, anomaly detection, ML model training.

1 Introduction and Motivation

Cloud management products (see Wavefront [1] and vR Ops [2]) are aimed at designing monitoring solutions of high precision and increasingly wider coverage of data center administration aspects. Often these solutions are enabled with a high frequency of sampling rate of data center indicators that is targeted to acquire a maximum level of information to take actions toward many product frontiers (such as performance sustainability, scale optimization etc.). On the other hand, samples acquisition at maximum possible rate implies various costs that affect efficient resource management and design of data-driven analytics. Therefore, tuning the monitoring solutions according to “adequate” or “efficient” sampling rates will result in a reduction of data management overheads, noise, and save extra compute and storage resources for various on-demand tasks.

Our investigation deals with experimental evaluation of efficient sampling rates of time series data in Wavefront subject to two important criteria:

- preserving the original distribution of the metric with minimum information loss, while reducing its sampling rate;
- preserving the information value of the metric in terms of accuracy of ML models we train on to deliver important analytics features in the product, while reducing its sampling rate.

Based on such an analysis, we can categorize our initial data base of time series into classes, where each class is characterized by its own efficient sampling rate (reflecting the nature/dynamism of individual flows).

We investigate the problem from anomaly detection and forecasting perspectives (using ARMA-based [3] and W-TSF [4] forecasting and anomaly detection systems in AI Genie [3]) to guarantee that the information loss does not affect the extreme behaviors in the reduced data set and quality of predictions. Our prototype algorithms are applied on more than thousand time series metrics from Wavefront to perform experimental validations. We demonstrate that significant reductions in sampling rates can be achieved while still providing accurate ML models for performing those important tasks in the product.

The application monitoring solution by Wavefront is designed to get deep insights from the underlying system with high-frequency sampling rate of time series (1 p/s) for all monitored metrics. This makes it an effective tool for application performance analysis and optimization, efficient capacity management and proactive planning, etc., with an intelligent time series query language. Wavefront as a high-performance streaming analytics platform also provides an alternate view (called AI Genie, see Fig. 1) to customer chart data that mainly focuses on anomaly detection and forecasting of monitored data.



Fig. 1. Wavefront's AI Genie. An anomaly detection chart on forecasted time series data.

2 Sampling Rate Reduction

The dataset we experimented with consists of 1530 metric time series (44,392 observations each) collected by Wavefront while monitoring the environment providing the product's cloud service to the customers. As more interesting are high variable time series, we filter out the dataset metrics with some level of consecutive constant behavior.

2.1 Analyzing Information Loss

To analyze the data distribution of the corresponding sampling rate, we specifically look into only the data values that were collected under lower frequencies. Fig. 2 illustrates ten times sample reduction principle to alerting.query.reissue_latency.duration.s metric original data.

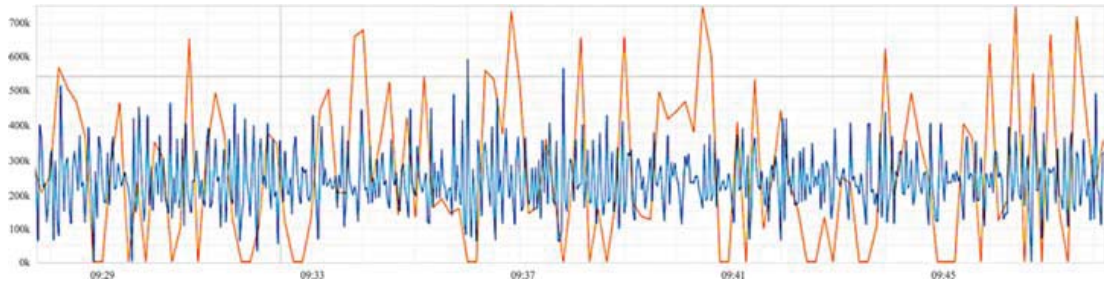


Fig. 2. alerting.query.reissue_latency.duration.s metric data values under 1 p/s (blue) vs. 1 p/10s (yellow) rates.

The incremental reduction of sampling rates leaves with ten distinct datasets corresponding to data with sampling frequency ranging from 1 sample per second to 1 sample per 10 second. Fig 3a and 3b show an example of the visual difference between ten- and zero-times reduced alerting.alerting_period.duration.max metric data that Wavefront has gathered from monitoring sample cloud application.

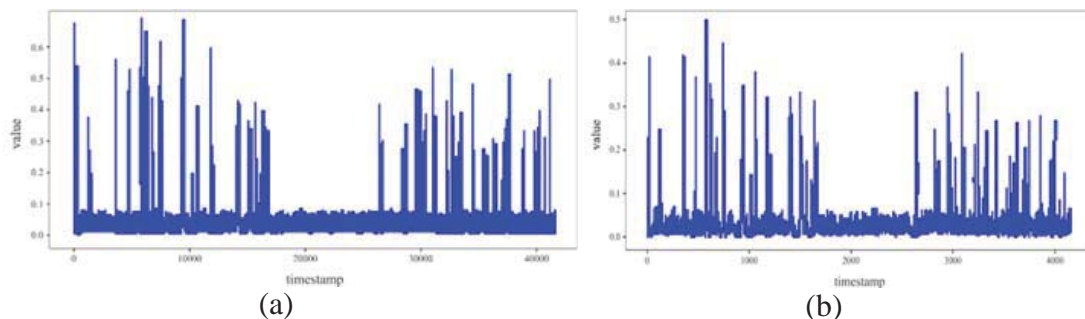


Fig. 3. alerting.alerting_period.duration.max metric sampled at (a) 1p/s and (b) 1p/10s.

Each metric of time series dataset with respect to each sampling rate from 1 to 10 per second, consists of different values spread in different range of intervals. Our approach is to divide the range of each metric into high-granularity (thousand) sub-intervals and compute the relative frequency of data points that fall into those for constructing relevant histogram distributions. The goal is to get an estimate of the probability distribution/mass function of the metric and see how the sampling rate reduction distorts it. Then the obtained relative frequencies for each metric and its reduced version can be interpreted as probability distributions of those. Fig 4 depicts histogram distributions of the original and 10-times reduced data.

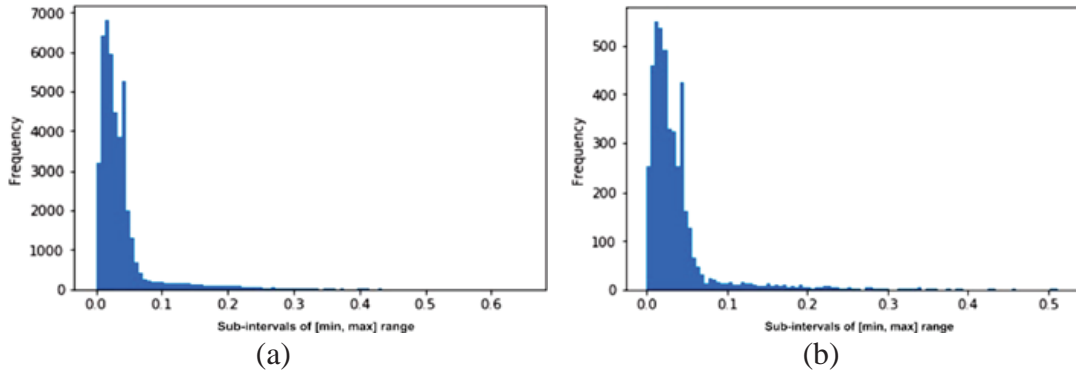


Fig. 4. Histogram distributions of (a) original and (b) reduced data of the same metric.

As our incentive is to measure in what significance does the sampling rate reduction affect the loss of information, we look into the Jensen-Shannon divergence/distance (also referred to as JSD) that is a method based on Kullback-Leibler divergence [5] measuring the closeness between two probability distributions. For any probability distributions P and Q , the JSD is defined by the formula:

$$JSD(P, Q) = \frac{1}{2}D(P, M) + \frac{1}{2}D(Q, M),$$

where $M = \frac{P+Q}{2}$ and $D(P, Q)$ is the KL divergence between probability distributions P and Q . JSD varies between 0 and 1. To see how similar the probability distributions of original and reduced sampling rated data are, we compute the relevant JSD.

As a result of information loss analysis, we found out that despite the dramatic change in sampling rate, the significant information content is preserved in most of the metrics that are monitored. We verified that 1290 out of 1312 metrics experience no more than 4% information loss (see Fig. 5) while reducing the sampling rate ten times. In this way, we were able to categorize application metrics based on their sensitivity to the sampling rate and refer to ten times lower sampling rate as general sampling rate for 1290 metrics to guarantee the information loss tolerance.

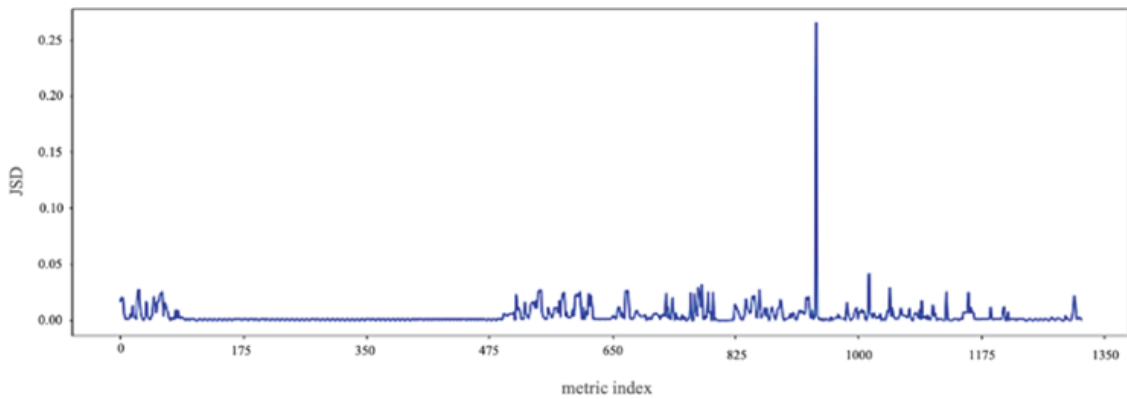


Fig. 5. Information loss experienced by all metrics of sampled at 1p per 10s rate.

Although most metrics turned to be indifferent to sampling rate reduction in terms of information loss, there is a group of metrics which need preserving the predefined sampling rate. Those metrics express high variability and dynamic changes of the system; hence, high-frequency sampling will be required to detect their abnormality behaviors and other important patterns for reliable management purposes.

2.2 Seasonal Trend Decomposition

The sampling rate reduction may affect the monitored object data from the perspective of change in seasonality and trend. We conducted Seasonal Trend Decomposition (also referred to as STL) filtering procedure for decomposing a seasonal time series. The comparison of STL results applied on the original and ten-times reduced data set demonstrated that none of the metrics in our dataset experience variations in seasonal and other effects being sampled at 1 point per 10 seconds frequency. Thus, we see that such a high reduction of sampling rate with low information loss is still tolerable in terms of a quality statistical analysis of the data.

3 Performance Analysis of Algorithms and Related Art

Performance analysis of algorithms for time series data has been mainly conducted in the context of cross-validation strategies. An interesting empirical evaluation of forecasting algorithms by Cerqueira *et al* [6] demonstrates that cross-validation approaches can be applied to stationary time series. However, according to their study, in case of many real-world data sets with presence of different sources of non-stationary behavior, “the most accurate estimates are produced by out-of-sample methods that preserve the temporal order of observations”. Compared to such a classical cross-validation setting, we are interested in validating ML models subject to information loss induced by sampling rate reduction. Data reduction in production analytics is an important technology challenge (see Poghosyan *et al* [7]). Relevant ideas linking to *information bottleneck* principle can be found in Harutyunyan *et al* [8].

3.1 Time Series Forecasting and Anomaly Detection

Wavefront also provides advanced analytical functionality in terms of time series forecasting and anomaly detection - mechanisms that enable its customers to promptly discover anomalous patterns in the data indicating possible misbehaviors within the workflows of the monitored environment. In particular, AI Genie applies two different anomaly detection and forecasting algorithms (ARMA-based [3] and W-TSF [4]), see Figs 6 and 7, respectively. The ARMA-based approach of anomaly detection incorporates multiple competitive models using online forecast engine to discover the set of anomalies in a whole testing window of forecasts in individual metric streams with respect to anomaly sensitivity of data in a given window. This algorithm describes the short-term temporal dependent patterns in the time series using autoregressive moving-average (ARMA) model. It calculates the forecast confidence bounds based on the

residual error of the selected ARMA model and the user-provided confidence interval parameter. The algorithm performs anomaly detection on complete set of forecasted window values directly.

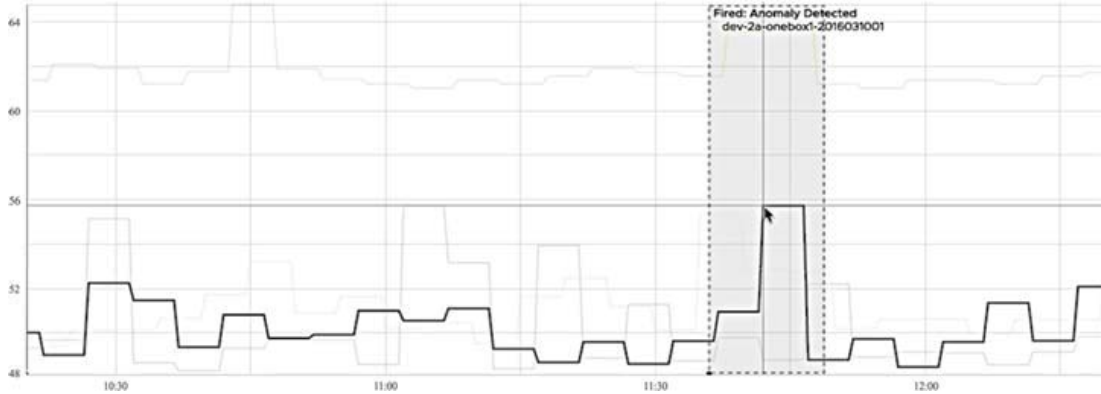


Fig. 6. ARMA-based anomaly detection in AI Genie.

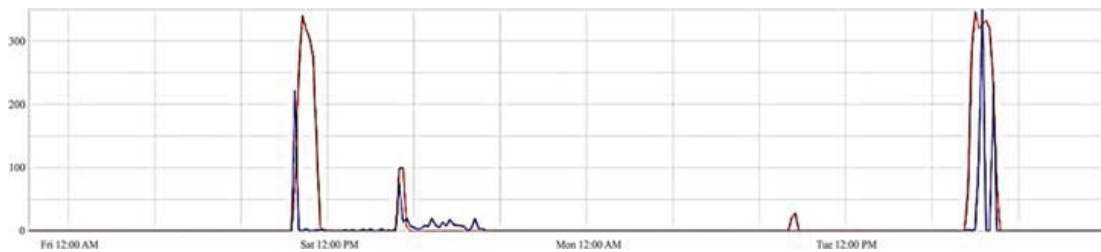


Fig. 7. W-TSF anomaly detection in AI Genie.

W-TSL technology leverages offline pre-trained neural network models and hypothesis testing procedures to achieve confidence bound-assist anomaly detection using transformations of data from non-stationary into a stationary process [4]. In contrast to ARMA-based algorithm, W-TSL starts iterating over the time series with sliding test window principle (Fig. 8) and calculates anomaly scores for each of the slide window using the set of forecasted values only for candidate slide window itself.

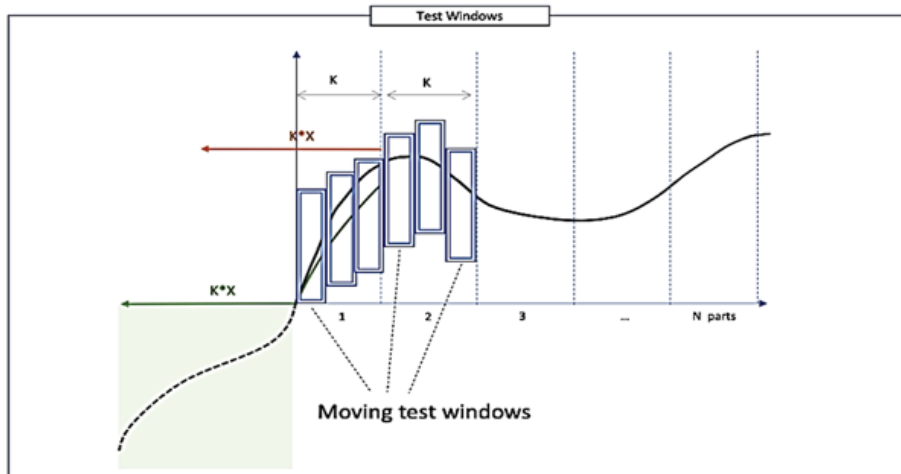


Fig. 8. Example of moving test windows (blue regions) used by NN based anomaly detection algorithm

If the anomaly score for a time window (e.g., 5 min) crosses a threshold (80% as a default value), then the window is declared to be anomalous (and alert is triggered). The anomaly score is computed by the percentage of data points outlying the confidence bounds of the anomaly detection algorithm, so it varies from [0,1].

Even though the above mentioned two algorithms introduce non-similar architecture designs and approaches to forecasting and anomaly detection task, they still rely on substantial amount of historical data to provide fundamental and accurate analytics on top of that.

3.2 Validating Efficient Sampling Rates for Forecasting and Anomaly Detection

To estimate if the reduced time series data set is still preserving its utility for training ML models (such as ARMA-based and W-TSF), we run the following experiment. We use 1290 metrics, each with its original (44392 data points) and reduced sampling rate with the tolerable information loss. Then we compare the results of anomaly detection performed by two algorithm using AI Genie running on production. The algorithms function by taking 20% of historical data at the start for learning and continuously moving forward to produce forecast and anomaly detection for the rest of the data. Figures 9 and 10 illustrate combination charts of the number of anomalies in the original and sampled datasets provided by the algorithms, respectively, across all metrics, after declaring the total count of anomalies at the end of experiment for each of the metrics.

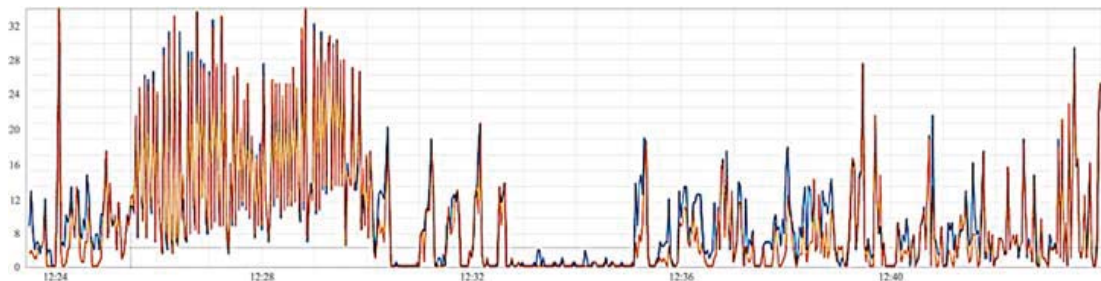


Fig. 9. ARMA-based anomaly count chart of original vs. sampled (ten times reduced) 1290 metrics data.



Fig. 10. W-TSF anomaly count chart of original vs. sampled (ten times reduced) 1290 metrics data.

As a result of our experiment (see. Table 1) we discover that W-TSF anomaly detection on ten-times reduced metric dataset provides almost identical distributions of anomalies in the test window. ARMA-based algorithm was more sensitive to the reduction with loss of 11% of anomalies.

Table 1. Number of anomalies detected by algorithms on test windows of original and ten times sampled datasets.

ML Algorithms	Original data	Ten times reduced data
ARIMA-based	7561	6669
W-TSF	8530	8571

Figs 11 and 12 show that the mean squared errors (MSE) of anomaly scores between original and reduced data (that both algorithms adopt as a measure to define anomalies/alerts on). The average metric MSE for ARMA-based algorithm is 0.0083 and the same average for W-TSF is 0.0135. So, for both algorithms we get a low difference in anomaly scores, which implies that anomaly detection algorithms still provide adequate predictions with much sparser data sets. Additionally, with reduced data sets we get significantly (10 times) less memory utilization and Disk IO.

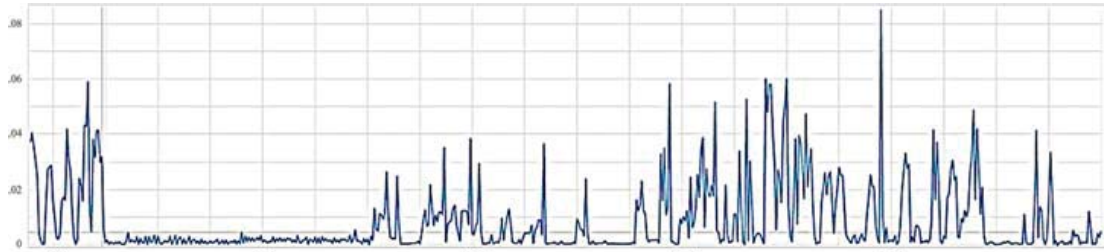


Fig. 11. MSE between anomaly scores in test window of original and sampled (ten times reduced) metrics data using ARMA-based anomaly detection.

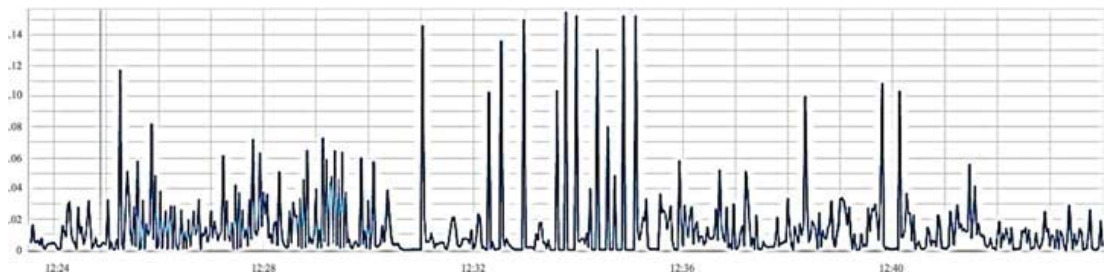


Fig. 12. MSE between anomaly scores in test window of original and sampled (ten times reduced) metrics data using W-TSF anomaly detection.

This verifies our intuition that the data reduction subject to tolerable information loss (according to Section 2) could also provide anomaly predictions enough accurate compared to baselines. Such a reduction implies an essential gain in overall performance for every streaming operation with the data. Based on the proposed analysis and related algorithms, we recommend the product the efficient sampling rates learned for effective

data management and analytics. Our algorithms can be regularly run to re-estimate those efficient rates with application evolution/dynamics.

4 Conclusion

We described an information-theoretic approach to estimating efficient sampling rates of monitoring flows while preserving their information content. Our experiments on a large data set measured by Wavefront demonstrate that significant reduction levels can be achieved with very low information loss. With such an approach we can substantially reduce the data management and analytics overhead forced by high-frequency monitoring and real-time analysis/representations of data center processes in daily operations. Those experiments prove that complex ML models can be still trained within the product with acceptable accuracies on substantially reduced data sets. It also improves performance of AI features of a cloud management product in terms of forecasting and anomaly detection.

References

1. Wavefront by VMware: <https://cloud.vmware.com/wavefront>.
2. vRealize Operations Manager: <http://www.vmware.com/products/vrealize-operations.html>.
3. Forecasting and anomaly detection with AI Genie: https://docs.wavefront.com/ai_genie.html.
4. Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Pang, C., Oganessian, G., Ghazaryan, S., Hovhannisyan, N.: W-TSF: Time series forecasting with deep learning for cloud applications. Submitted to CODASSCA-20 (2020).
5. Cover, T., Thomas, J.: Elements of Information Theory. Wiley, 1991-2006.
6. Cerqueira, V., Torgo, L., Mozetic, I.: Evaluating time series forecasting models: An empirical study on performance estimation methods. <https://arxiv.org/pdf/1905.11744.pdf> (2019).
7. Harutyunyan, A.N., Chen, Y., Han Vinck, A.J.: Thoughts on information-theoretic aspects of several problems in data science. In: Collaborative Technologies and Data Science in Smart City Applications (CODASSCA), pp. 118-122, AUA, Yerevan, Armenia, Sep. 12-15, (2018).
8. Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M.: Managing cloud infrastructures by a multi-layer data analytics. In: IEEE Int. Conference on Autonomic Computing (ICAC), pp. 351-356, Würzburg, Germany, July 19-22, (2016).

W-TSF: Time Series Forecasting with Deep Learning for Cloud Applications

Arnak Poghosyan¹, Ashot Harutyunyan¹, Naira Grigoryan¹, Clement Pang¹,
George Oganessian¹, Sirak Ghazaryan¹, and Narek Hovhannisyan²

¹VMware, Inc.

{apoghosyan; aharutyunyan; ngrigoryan; pangc; goganesyan;
sghazaryan}@vmware.com

²TeamViewer Armenia

narek.hovhannisyan@teamviewer.com

Abstract. One of the main targets of application performance managers is monitoring of cloud environments with high-velocity custom metrics and analytics. The key components of time series data analytics are forecasting and anomaly detection. The classical methods of time series forecasting were recently empowered by neural network-based models which gain increasing popularity due to their flexibility and ability to tackle complex non-linear problems. Meanwhile, some of the disadvantages of that approach mitigate expectations and require specific solution for SaaS applications. The first challenge for network-based models is resource utilization due to GPU trainings. SaaS applications are extremely sensitive to luxury resources due to high costs. The second challenge is inability of the networks to handle non-stationary time series data that behave with trend and/or seasonality. In this paper, we propose W-TSF, a time series forecasting engine that was preliminary designed for Wavefront by VMware, a monitoring tool for cloud environments. W-TSF resolves all mentioned problems. Implementation and testing for real-customer time series data proved its acceptable capabilities for cloud environments in terms of prediction accuracy and resource consumption.

Keywords: time series data, neural networks, forecasting, anomaly detection.

1 Introduction

One of the key components for application performance monitoring/management (APM) software to appear in the Gartner’s magic quadrant [1] is the availability of Analytics with “artificial intelligence for IT operations (AIOps)”. To accomplish this, APM solutions employ event correlation, anomaly detection and root cause analysis algorithms on APM-acquired data. Time series data (known also as metrics), together with logs, traces, histograms and events is an intrinsic APM-acquired data type heavily utilized by all APM leaders like New Relic, AppDynamics, Dynatrace, Broadcom and others [1].

Fast and accurate time series analysis and forecasting is of great importance for various reasons like anomaly detection, anomaly prediction and capacity planning. When metric data are collected and analyzed by a monitoring system, administrators of a cloud environment may desire automated forecasts of future metric-data values indicative of likely future states of applications or infrastructure components. Data related to computing-resources and capacities may include trends indicating that additional processor bandwidth or mass-storage capacity may be needed, in the near future, due to increasing workloads, in order to prevent delays and failures and/or to maximize economic efficiency. Time series data can also be used for correlation analysis and as a source of anomaly events for further root cause analysis. In this paper, we focus our attention to time series forecasting capabilities.

Analysis of time-series data is a significant branch of mathematics and computing that includes a variety of different types of analytic procedures, computational tools, and forecasting methods. It is sufficient to mention the well-known and powerful approaches like SARIMA and Holt-Winters' (see [2] with references therein). However, certain applications require relatively quick forecasts and are associated with significant temporal constraints, forestalling lengthy and computationally intensive analyses. In other applications, including cloud-computing applications, the price of complex computational processes needed for accurate forecasting may outweigh the benefits of the forecasts produced by the computational processes.

Application of neural networks (NN), and other machine-learning techniques, may produce an efficient approach to time-series analysis and forecasting [3]. However, naïve implementation of a neural-network-based system in a cloud-computing environment would likely fail to provide adequate response times and would likely be far too expensive for most clients. Training and storing of neural networks are both time-consuming and expensive with respect to the necessary resources.

Hence, it is not feasible to train those models in demand for the specified time series data. From the other side, it would not be feasible to train and store special-purpose neural networks for all of the different possible types of time series. A naïve attempt to train a single neural network to analyze all of the various different types of time-series data would also likely fail, since different types of time-series data exhibit different types of behaviors and temporal patterns, and because a single neural network would need a vast number of nodes and even vaster sets of training data to produce reasonable forecasts for general time-series data.

In this paper, we suggest a NN-based system [4] (named as W-TSF) for a time series forecasting that essentially differs from other well-known classical techniques. The purpose was training a generic model applicable to a wide range of real-customer time series data. Our efforts were devoted to development of such procedures that would allow practical realization of the idea. Here, we discuss the theoretical foundation of the approach and show the results for real cloud-computing environments. Implementation and testing are performed in Wavefront by VMware. Wavefront offers a real-time metrics monitoring and analytics platform designed for optimization of cloud and modern applications that rely on containers and microservices.

Worth noting that our main goal is the performance of the approach for cloud environments rather than accuracy compared to the well-known techniques that perform

individual training for each specified time series data in the GPU accelerated environments. For us, the performance is balance between accuracy and resource utilization. We observed that the accuracy of the forecasts is comparable to the classical ARMA related approaches while preserving resource consumption on acceptable levels. In particular, application of the pretrained network to a specified time series in a cloud environment can be performed without GPU acceleration and with moderate number of CPU cores.

2 Our Approach and Related Work

As we already concluded, the utilization of NN-based models for time series forecasting faces two main milestones. The main milestone is resource utilization, firstly, availability of GPUs for model trainings. The solution is straightforward for some types of problems. For example, computer vision typically utilizes 2D convnets by training generic networks on powerful data centers and further applying the pretrained networks to specific problems. There are well-known models available for practical applications and this approach was proved to be effective for many similar applications.

We follow this common idea of training a general network and storing it for further application to forecasting. This will help to transfer the problem of resource utilization from a customer side to our side as training should be performed in our private cloud environments where we can utilize enough powerful GPUs. Figure 1 shows implementation of this approach. The entire system consists of two separated parts: off-line and on-line modes. Off-line mode performs model training for a special time series database consisting of time series data across different customers. It corresponds to the training in a powerful private datacenter. The weights of the trained network and its configuration is stored in a cloud for further utilization as a file in “json” format. On-line mode corresponds to a customer cloud-computing environment. The weights and configuration of the pretrained network can be restored from the file and applied to the specified time series data. Hence, the forecasting can be performed without GPU acceleration just by some CPU cores.

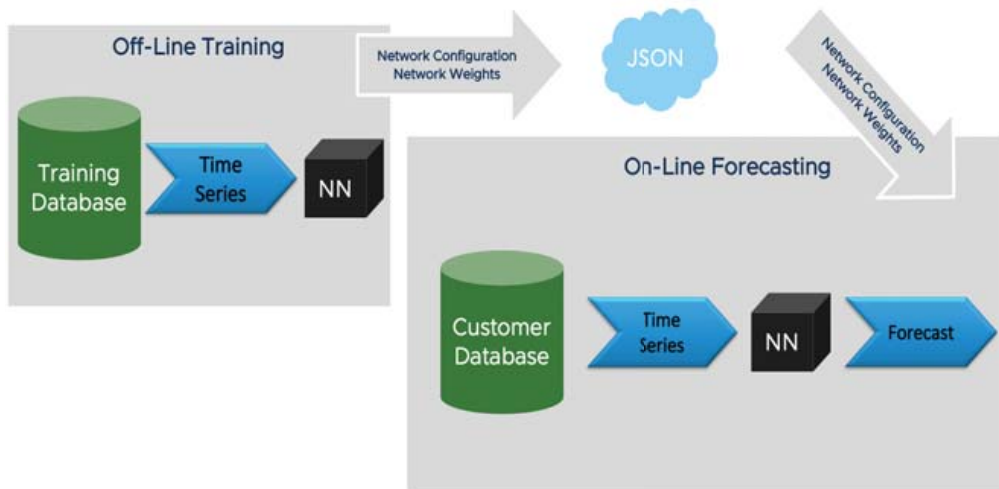


Fig. 1. The general flow describing interconnection of off-line and on-line modes.

A crucial milestone connected with the system of Fig. 1, is the diversity of time series data behaviors that probably will not allow to handle all of them with a unique trained model. Natural way should be classification of time series into some well-known classes for which class-specific models can be trained. This idea we keep for checking in our future works. More simplified version of the latest idea is to train a model for a specified class and transform all other time series to this specific class.

How to find the class with the best trainable characteristics? Naturally, models based on networks should have better performance due to their non-linearity, flexibility and ability of generalization. It is assumed, that with NNs, no any specific assumptions need to be made about the model which should be one of the most important advantages. Different authors showed in their studies and experiments [5-10] that better results compared to SARIMA and related models could be achieved only by combination of transformations that ‘stabilize’ the behavior (e.g. detrending, deseasoning) of the specified time series. However, to be honest, the results regarding the forecasting of non-stationary time series data directly via NN models are very controversial (see [9] with references therein).

In other words, trying to be mathematically more rigorous, the NN models have weak forecasting capabilities while applied to non-stationary time series data. It means that the training of NN models and their application to time series forecasting will provide with acceptable accuracy only in case of stationary data. The set of stabilizing transformations is time series class specific. Say, the class of time series data with a deterministic trend can be stabilized via detrending by a regression (linear or non-linear), the class with a stochastic trend by a differencing of the proper order, etc.

Our implementation applies different well-known hypothesis testing algorithms for time series classification. Deterministic versus stochastic trend classification can be performed via KPSS [11] and ADF [12-14] tests. Deterministic versus stochastic periodicity analysis can be performed via PDM [15,16] and Canova-Hansen [17] tests. As a result, NNs can be trained only for stationary time series data. Application of pre-trained neural networks to a user specified non-stationary time series needs passing through classification, transformations, forecasting and final reverse transformations stages as Fig. 2 describes.



Fig. 2. The process of non-stationary time series forecasting.

3 Implementation

As already described, we deal with two separate processes – off-line training of a neural network model and on-line application of the model to a user specified time series data.

Off-line model training was performed in VMware private datacenters equipped with powerful GPUs. Experimental training database included 3000 time series, taken

from real customer cloud environment. Time series had 1-minute monitoring interval and, in average, 1-month duration. We tried different models based on multilayer perceptron (MLP) and recurrent neural network (RNN) architectures with different number of layers and nodes and checked their accuracies on a test data via root mean square measure. We decided to continue working with the MLP models as found nonsignificant difference compared to RNN.

Actually, the MLP models have the simplest known architecture of NN, and they are easy to implement, train and use. The current model has 2 hidden layers with 256 nodes in each. ‘relu’ activation function is applied to the hidden layers. The latest layer has linear activation function. Input layer consists of 40 nodes and output layer of 20 points. Hence, the network forecasts 20 future data points based on 40 historical data points. Overall, 81,428 weights were trained for this specific model. ‘Adam’ optimizer and mean average error (‘mae’) as a loss function were used. We applied 5 epochs for each time series and 20 epochs for the entire database. The idea was in getting a generic model for a database rather than overfitting a specific time series. Also, we used *batch_size* = 1500.

On-line mode applies the pretrained NN to a user specified time series. This part is implemented in Wavefront by VMware. AI Genie UI of Wavefront (see Figs. 3 and 4) simplifies and automates time series forecasting and anomaly detection capabilities. It requires minimal set of parameters to start running the AI engine.

In the case of forecasting a user can specify (or use defaults) a time series, select the forecast horizon, and the corresponding sensitivity of the confidence bounds (for drawing narrow or wide bounds). In Figs. 3 and 4, the red curve corresponds to the historical data, the black curve to the forecast, and the green area to the confidence bounds. The forecast horizon is 1 week. It means that the historical data is 2 weeks as the pretrained network works with 2:1 ratio (the historical period is twice longer than the forecast period). Confidence bounds correspond to “moderate” setting (the others are “conservative” and “aggressive” settings).

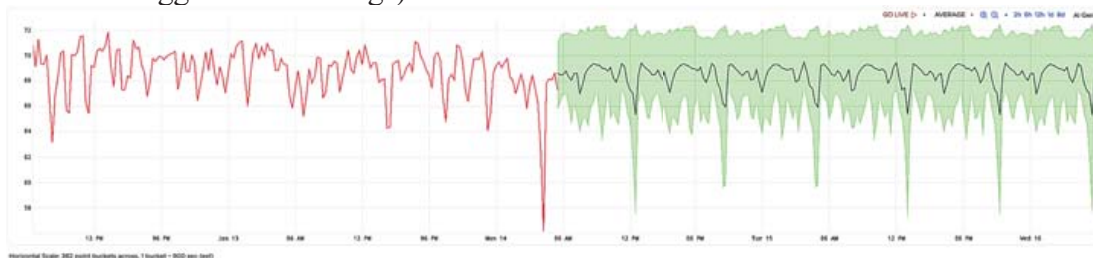


Fig. 3. Wavefront AI Genie UI for time series forecasting. Example of a stationary data.

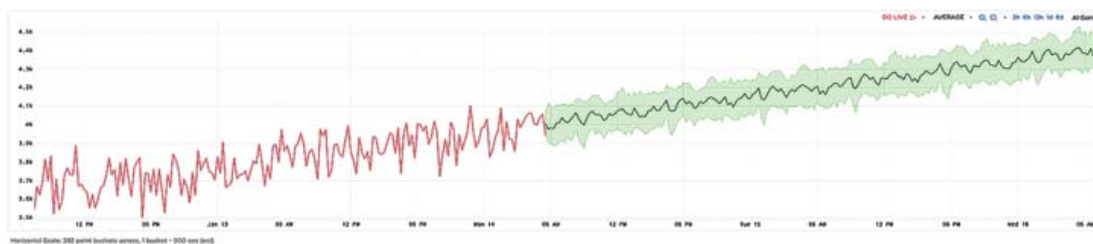


Fig. 4. Wavefront AI Genie UI for time series forecasting. Example of a trendy data.

Confidence bounds can be used for estimating the accuracy of the forecast (wider means less accurate) and also for time series anomaly detection. They allow to define anomaly scores meaning the percentage of data points that violate confidence bounds within a test window. Anomaly should be detected, and the corresponding alert triggered for the test window if the “anomaly score” is bigger than a predefined threshold.

4 Discussion

The first problem that we encountered during the implementation is the limited number of input and output nodes of the NN model. It meant that 20 future points could be estimated based on 40 historical points. This was very strong limitation that allowed to handle trend very well but not smaller peculiarities especially for longer historical periods. We decided to take grids with number of points multiple to 40, divided the entire grid into sub-grids with 40 points and sequentially fed the NN model. For each those sub-grids, the NN model provided with 20 forecasts which were regrouped together to get forecast data points with denser resolution.

By default, we used predefined number of grid points connected with the length of the horizon window. For example, we set 16,000 points for the 3 months window. In future, we need to optimize those values as experiments showed that the same accuracy of the forecasts is possible to derive even by smaller number of grid points.

Experiments revealed some interesting characteristics of the proposed system. The first important result was comparison with the well-known ARMA related approaches. We selected a big dataset composed of thousands time series and performed three groups of experiments with different lengths of historical data like 2 hours, 1 week and 4 weeks and then calculated the corresponding relative root mean square errors for the corresponding forecasts for both approaches.

Then, we calculated the percentages of time series data for which the error is smaller than 1, bigger than 1 but smaller than 10 and bigger than 10. It is important that the percentages in each group were almost similar for both approaches. Of course, for different time series data the winners were different, but overall the forecasting powers across the entire dataset were equal.

The result is very important as ARMA-based models construct a specific model for each selected time series data while our approach uses the same model for the entire dataset with preliminary specific transformations for non-stationary data. By the way, ARMA-based models are doing quite the same for each specified time series to detect the corresponding trend and seasonality. We didn't compare the running times of each approach as they actually utilized different number of historical points and the comparison would be unfair.

We investigated also the complexity of our approach in different stages of execution. The entire process consists of three main parts – data collection and preprocessing, application of the neural network, and visualization of the forecasted data values. We estimated that the most complex part is the first stage which takes almost 80% of the entire execution time. Worth noting also, that the NN model requires equidistant sampled historical points which we derived via linear interpolation.

References

1. Magic Quadrant for Application Performance Monitoring, <https://www.gartner.com/doc/reprints?id=1-1YTYAJJ4&ct=200422&st=sb>, last accessed on 2020/06/23.
2. Hyndman, R.J., Athanasopoulos, G.: *Forecasting: Principles and Practice*, 2018, Monash University, Australia. Available online: <https://otexts.com/fpp2/>, last accessed on 2020/06/23.
3. Lewis, N.D.: *Deep Time Series Forecasting with Python: An Intuitive Introduction to Deep Learning for Applied Time Series Modeling*. CreateSpace Independent Publishing Platform, 2016.
4. Poghosyan, A., Hovhannisyan, N., Ghazaryan, S., Oganessian, G., Pang, C., Harutyunyan, A., Grigoryan, N.: *Neural-Network-Based Methods and Systems that Generate Forecasts from Time-Series Data*. Filed Jan 15, 2020. US patent application No: 16/742,594.
5. Farway, J., Chatfield, C.: Time Series Forecasting with Neural Networks: A Comparative Study using the Airline Data. *Applied Statistics* 47, 231–250 (1995).
6. Kolarik, T., Rudorfer, G.: Time Series Forecasting using Neural Networks. *APL Quote Quad* 25, 86–94 (1994).
7. Nelson, M., Hill, T., Remus, T., O'Connor, M.: Time series Forecasting using NNs: Should the Data be Deseasonalized First? *Journal of Forecasting* 18, 359–367 (1999).
8. Hansen, J.V., Nelson, R.D.: Forecasting and recombining time-series components by using neural networks. *Journal of the Operational Research Society* 54(3), 307–317 (2003).
9. Zhang, G.P., Qi, M.: Neural Network Forecasting for Seasonal and Trend Time Series. *European Journal of Operational Research* 160, 501–514 (2005).
10. Wang, X., Smith, K.A., Hyndman, R.J.: Characteristic-Based Clustering for Time Series Data. *Data Mining and Knowledge Discovery* 13(3), 335-364 (2006).
11. Kwiatkowski, D., Phillips, P., Schmidt, P., Shin, Y.: Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure are We that Economic Time Series have a Unit Root? *Journal of Econometrics* 54, 159-178 (1992).
12. Dickey, D.A., Fuller, W.A.: Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *J. Amer. Stat. Assoc.* 74, 427–431 (1979).
13. Dickey, D.A., Fuller, W.A.: Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica* 49, 1057-1072 (1981).
14. Said, E., Dickey, D.A.: Testing for Unit Roots in Autoregressive Moving Average Models of Unknown Order. *Biometrika* 71, 99–607 (1984).
15. Stellingwerf, R.F.: Period Determination using Phase Dispersion Minimization. *The Astrophysical Journal* 224, 953-960 (1978).
16. Davies, S.R.: An Improved Test for Periodicity. *Mon. Not. R. Astr. Soc* 244, 93-95 (1990).
17. Canova, F., Hansen, B.E.: Are Seasonal Patterns Constant Over Time? A Test for Seasonal Stability. *Journal of Business and Economic Statistics* 13, 237-252 (1995).

Fingerprinting Data Center Problems with Association Rules

Ashot N. Harutyunyan, Naira M. Grigoryan, and Arnak V. Poghosyan

Office of CTO of Cloud Management, VMware Eastern Europe
{aharutyunyan;ngrigoryan;apoghosyan}@vmware.com

Abstract. Cloud management technologies increasingly automate different aspects of data center administration, where the final goal is to make self-driving solutions. Learning fingerprints of KPI- or SLO-impacting performance problems in IT infrastructures is a relevant task towards such a vision. Instead of defining problem types for data center components (resources/objects of various kinds) using domain knowledge, which is hard to obtain and unreliable because of complexities and sophistication of modern cloud systems, we propose a ML framework to detect those issue categories. Then alerting engines can run on top of those patterns to notify the users on conditions that are impacting system's KPIs thus providing explainability for troubleshooting and long-term performance optimization of the infrastructure. We consider several scenarios for learning problem definitions in terms of constructs by vRealize Operations – one of the leading solutions in the cloud management market. Using association rules mining concepts we can recommend problem patterns (fingerprints) in form of minimum size attribute combinations that constitute core structures highly associated with degradation of the KPI or SLO loss. We demonstrate experimental insights on virtualized environments applying our prototype algorithm.

Keywords: Data Center Management, Problem Fingerprint, Association Rules.

1 Introduction

One of central tasks in management of cloud computing environments/applications is maintaining required “health” of those systems measured by behavior of Key Performance Indicators (KPI) or Service Level Objectives (SLO) as thresholds on those KPIs. Normally, cloud operations solutions (including vRealize Operations (vR Ops) [1] and vRealize Log Insight (vR LI) [2]) allow manual configuration of alerts/problems with abnormality conditions that user perceives as precursors of important deviations from typicality. In those products users define their own problems with an appropriate alert workflow. They can deal with composite anomalies with conditions defined manually and get notified in case of their occurrence. It is thus assumed that those self-defined conditions might impact the data center operations worth paying attention to and hence introduce a user-controlled alerting noise. However, it brings a new issue in terms of unfeasible manual and ad-hoc configuration efforts for large-scale data centers hardly tractable by expert knowledge. Moreover, unreliable problem definitions may introduce

high rate of missed (false negative) alerts. Our framework introduces a learning approach for identifying the core (performance) problem patterns or fingerprints subject to SLO loss. Those environment-specific problem specifications as min-size anomaly ensembles represent the basic anomaly definitions capable of detecting deteriorating health or a potential SLO violation. We propose criteria to rank problems to effectively handle the alerts in terms of system remediation and show results on a real environment.

Several scenarios towards automation of problem definitions at attribute layer of IT resources/objects are discussed. This means that we propose to recommend the user those attributes of the infrastructure/application which are highly associated (ranked) with KPI degradation and need to be reflected in problem definitions. In other words, the user is offered to “codify” his/her problems primarily with atomic abnormality (symptoms) conditions on the indicated attributes which are able to decrease the KPI by some pre-specified degree that the user does not want to be ignorant of. This will forewarn the user regarding potentially significant KPI decline to prevent unwanted loss and take actions to depress the active anomalies. Moreover, by mining association rules we can recommend problem patterns (fingerprints) in form of minimum size attribute combinations that constitute core structures highly associated with degradation of the KPI or its loss. This approach can lead to full definitions of problem alerts in an automatic way, while learning rules with specific abnormality conditions (symptoms) occurring on those attributes.

The continuity assumption. In general, we assume that the SLO loss is a gradually evolving process impacted by increasing number of anomalies in the system. So, if some abnormality conditions/combinations are associated with tolerable KPI degradations, they are precursors of higher degradations and SLO loss.

In section 2 we discuss the concept of the problem fingerprint. In Section 3 we introduce criteria for ranking objects attributes/indicators in terms of their association to real performance issues, to characterize their ability to leave fingerprints. Based on those criteria we can build different association rules using the relevant machine learning framework and foresee several implementation scenarios of different complexity and depth (Section 4). In Section 5 we demonstrate experimental results from a real environment with attribute-level analysis of the problem fingerprinting and their transaction patterns as rules of associations that define attribute fingerprints.

2 The Concept of Problem Fingerprinting

Based on the Dynamic and Hard Thresholding techniques ([3], [4]) employing sophisticated statistical inference methods on time series metrics measured from the entire data center, vR Ops is capable to detect every atomic change/outlier (against historically representative behavior of the monitoring flow) or anomaly occurring in the system not primarily yielding a malfunction. Extra sources for atomic anomalies can be different monitoring platforms such as log management products (e.g., vR LI [2], see relevant ML approaches [5]-[7] developed in this area). Those anomalies are then hierarchically employed to estimate the overall statistical health of IT resources and entire infrastructure stack based on their volume and distribution. Within this approach to

anomaly detection, those atypical events (alarmed to the user) are not necessarily linked to a performance indicator that would allow qualifying them as real performance problems (i.e., situations with significant drops in the KPI) as well as quantifying their impact on the system.

With historical analysis of DT and HT violations (as anomalies with symptoms DT-above/below, HT-above/below, respectively) for an object kind as basic anomaly space in vR Ops, as well as other relevant events of different types (configuration, properties, log-related alarms), we are interested in identifying those object attributes that

- 1) are highly associated with the KPI degradation according to different criteria;
- 2) constitute patterns in their combinations/transactions that are collectively well-associated with the KPI drop.

Those *patterns* are called *problem fingerprints* as the most consistent structures that we observe in the system such that they make a “rule”. For determination of those rules we apply the machine learning framework of association rules. These constructions will logically lead to an efficient control over the critical anomalies and their blocks making a fundamental layer of the management analytics to proactively prevent potential regression of the system state into a serious performance issue or dysfunction. In the experimental part of our work we show that those core patterns, characterizing the Virtual Machine (VM) object kind, exist and can be recommended to feed the alerting engine.

Problem fingerprinting is of high importance for users of very large and complex environments enabling them to gain deeper insights into performance specifics of own infrastructures and applications. Prior related art in this domain was mainly focused on similarity analysis of data center incidents in their reoccurrence (see PhD thesis by Bodik [8], references therein, including a direct ascendant work by Cohen *et al* [9] on identifying *crisis signatures*). Our approach differs in nature, it is about learning important combinations (rules) of atypical states of data center flows subject to their potential impact on a KPI. This provides a linkage between all statistically unexpected events and their consequences on the system performance thus reducing event noise. Moreover, we quantitatively characterize several aspects of those rules to be used for weighing their risk/impact. The main use case applications of this learning are automation of alert definitions (a recommender system for performance-oriented alerting) in cloud management products and explainability of the situations with SLO loss for targeted troubleshooting and root cause localization.

3 Characterization Criteria for Attributes

To measure which indicators/attributes (type of an IT process) of a resource/object kind (e.g., VM, or any custom group of infrastructure elements) are highly important to include in a problem definition or able to make a fingerprint individually, first we look at

1. in what frequency their events (DT/HT violations, etc.) are historically associated with the KPI degradation;
2. the degree of KPI degradation those attributes are associated with;
3. how many other attributes have co-occurring events during the KPI decline;

Based on those criteria we summarize 3 scores for each of an attribute in a resource kind, namely

- *Participation rate*;
- *KPI metric (health) degradation rate*;
- *Co-occurrence index*.

In particular,

- the *participation rate* is the prior probability of the attribute association with SLO loss;
- the attribute's *health degradation* rate shows the average degradation in the KPI when both the SLO loss and attribute's events are observed;
- the *co-occurrence index* of an attribute explains the weighted average count of co-occurring attributes (in terms of their events) at SLO losses.

The weights in the co-occurrence index are measured by frequencies of the attribute participation in event groups (occurring at the SLO loss) of different lengths. Final rank of an attribute as a priority recommendation index, to consider it as an individual fingerprint, is defined by a function on those scores. We omit this part for further study.

4 Learning Association Rules

The machine learning framework of association rules makes a relevant fit to our goal of finding attribute-transactions-based association patterns with the KPI behavior, in addition to individual association scores. The work [10] which introduces this transactions mining approach of different interestingness criteria has been applied in various domains, initially emerged as a discovery mechanism of relations between variables in large databases. The sales data of a supermarket is an example. There are many interestingness criteria studied in machine learning literature, such as confidence ("strength") or conviction of the transaction rule, etc. In the example of the supermarket sales, for instance, if "butter" is frequent in transactions where "bread" is bought, then there is a rule: "butter is bought with bread". The conditional probability of this link is the confidence of the rule. Using this framework we can study what combinations of "interesting" attributes make a rule transaction (with "enough" confidence as a conditional participation rate criterion) in terms of association with a KPI degradation and its degree, as well-as of the co-occurrence index.

As a particular scenario we indicate a procedure that identifies rules with *maximum confidence transitions* in the following sense Assume we are looking for rules that could contain attributes $a(1), a(2), \dots, a(r)$ of our interest in terms of their individual scores (such as high rates in SLO degradation). Starting from $a(1)$ we search for an attribute from the specified list that can make a rule within the overall transactions set subject to maximum confidence above some level. In other words, we declare that $[a(1); a(2)]$ is a rule, if $a(2)$ occurs with $a(1)$ at SLO losses above a conditional probability which is the maximum over the rest of possible attributes. Then the third attribute component is chosen with the same principle of maximum conditional probability conditioned on the pair $[a(1); a(2)]$, and so on, until we are not able to add the next component with enough confidence.

In view of this machine learning framework and the criteria we introduced in Section 3 to build fingerprint rules, we shortly indicate some execution scenarios of different depths for our algorithm.

Execution Scenario 1 (pure attribute ranking). We can restrict ourselves with recommending only high rank attributes for problem constructions based on the criteria in Section 3.

Execution Scenario 2 (association rules). Using association rules, we can mine attribute co-occurrence rules of different interestingness. Those combinations of attributes users take as problem definition structures, adding abnormality conditions on them. Certain SLO loss may only occur when multiple attributes experience abnormalities. The maximum co-occurrence index of an attribute set indicates the upper size of association rules reasonable to search for that attribute. For each of association rule discovered, we analogously measure relevant scores of participation, degradation, and co-occurrence to produce their importance rank for resource kind performance.

Execution Scenario 3 (Attribute plus symptom fingerprinting). The approach can be applied to the events space of an object kind to detect the exact symptoms (abnormality conditions) that need to be assigned to fingerprint attributes.

Execution Scenario 4 (ranking user-defined problems subject to KPI metric). One particular scenario is to apply the above described association approach to scoring/ranking the user-defined alerts against performance indicators and come up with recommendations that indicate which problems are relevantly defined and which are of low relevance. For that purpose, the scoring of alerts according to the criteria of participation, health degradation, and co-occurrence is identical to the procedure described for Attributes in Section 3. Moreover, it is possible to look for alert rules that are “interesting” in terms of higher impact on the KPI metric, applying the same algorithms we use for identification of attribute rules.

5 Experimental Results

We ran experiments on an internal data center of active usage for the VM object kind to investigate attributes of those objects in terms of their individual scores, as well as patterns making association rules according to predefined interestingness. The input data parameters are:

- number of VM's: 50;
- number of attributes: 700;
- number of attributes violated SLO: 86;
- duration of monitoring metrics: 30 days;
- number of events analyzed: 454,437.

In the experiments, the Anomalies score of those VM's in vR Ops as a KPI metric is considered. The Anomalies score represents how abnormal the behaviour of the object is, based on its historical metrics data. This score is calculated using the total number of threshold violations for all metrics for the selected object and its child objects. A low Anomalies score indicates that an object is behaving in accordance with its historical normalcy - most or all of the object metrics are within their thresholds. A high

number of Anomalies usually indicates a problem (statistically) or at least a situation that requires attention. Therefore, within this setting, looking for attribute-level fingerprint transactions is equivalent to identification of patterns that statistically gravitate “lot of” other attribute anomalies.

In Table I we illustrate those attributes that are observed in SLO violation with larger than 0.01 participation rate, where SLO=50 as threshold on the Anomalies scores of all 50 VM’s.

In Table II we show some association rules of attributes which are composed of the top 5 attributes from Table I that we are interested in (scores are rounded). In this case, the participation rate of an attribute vector (a defining factor in declaring fingerprint rule) is measured as a conditional probability of its occurrence over all assembles associated with the KPI degradation with equal or larger size. In the meantime, the KPI degradation index of an attribute rule is defined by the worst-case scenario – by the component with maximum KPI degradation score. Therefore, we see that most of the rules included in Table II have the same KPI degradation index, because in all those patterns the attribute `guestfilesystem|percentage` is present, which has the highest such a score among other fellow attributes. As to the co-occurrence score of a rule, it is calculated as the average of co-occurring attributes when the rule’s transaction is observed.

Table I. Attributes by Association Criteria.

Attribute Name	Part. Rate	KPI Degrada-tion	Co-Occur-rence
<code>guestfilesystem percentage</code>	0.07	69.27	21.60
<code>guestfilesystem usage</code>	0.04	70.57	22.61
<code>diskspace activeNotShared</code>	0.03	72.52	18.95
<code>net bytesTx_average</code>	0.03	73.73	24.49
<code>cpu usagemhz_average</code>	0.03	74.15	30.68
<code>net transmitted_average</code>	0.03	74.60	24.49
<code>cpu idle_summation</code>	0.03	72.73	29.06
<code>cpu used_summation</code>	0.03	73.90	29.60
<code>net usage_average</code>	0.03	72.34	27.80
<code>cpu readyPct</code>	0.02	71.16	24.34
<code>cpu ready_summation</code>	0.02	71.16	24.34
<code>cpu wait_summation</code>	0.02	71.97	29.08
<code>net packetsTxPerSec</code>	0.02	70.77	20.45
<code>net received_average</code>	0.02	73.31	30.07
<code>net packetsRxPerSec</code>	0.02	73.32	34.97
<code>datastore write_average</code>	0.02	74.30	24.36
<code>guestfilesystem freespace_total</code>	0.02	70.34	22.91

datastore maxObserved_Write	0.02	78.09	24.20
guestfilesystem usage_total	0.02	71.16	23.00
guestfilesystem percentage_total	0.02	70.57	22.68
diskspace used	0.02	72.50	18.98
mem host_usage	0.02	71.24	23.67
net bytesRx_average	0.02	71.82	30.07
datastore maxObserved_NumberWrite	0.02	78.01	20.73
mem guest_usage	0.01	70.92	23.22
mem consumed_average	0.01	70.92	23.22
datastore totalLatency_average	0.01	72.99	36.62
mem guest_dynamic_entitlement	0.01	70.80	23.19
mem shared_average	0.01	71.35	26.80
virtualDisk commandsAveraged_average	0.01	69.04	29.49
cpu numberToRemove	0.01	74.86	25.68
datastore commandsAveraged_average	0.01	69.38	30.99
disk usage_average	0.01	72.95	32.37
datastore maxObserved_NumberRead	0.01	78.17	31.20
virtualDisk totalReadLatency_average	0.01	78.46	33.54
virtualDisk numberWriteAveraged_average	0.01	74.32	35.39
cpu swapwait_summation	0.01	69.05	23.85
virtualDisk usage	0.01	73.07	33.78

In general, we observe that only a small portion of attributes are associated with the Anomalies score degradation with the specified threshold above. Within those we don't see strictly dominant group of attributes in terms of the participation rate. They also have close KPI degradation scores because of averaging effect, with some diversity in the co-occurrence. However, in case of rules/fingerprints the diversity becomes significant, both in terms of the participation rate, as well as the co-occurrence index, where we see that the rule of 5 attributes

cpu|usagemhz_average
diskspace|activeNotShared
guestfilesystem|percentage
guestfilesystem|usage
net|bytesTx_average

gravitated the maximum number of other attributes. It is an intuitive example that shows that there are dangerous combinations of attributes that can result in higher rate of anomaly propagation over the object.

In Table III we demonstrate some rules discovered with the principle of *maximum confidence transition* described in Section 4. Their scores of participation rate, KPI

degradation rate, and the co-occurrence index can be computed similarly with the case of Table II. Table III shows, for instance, different parameters of guestfilesystem that represent various quantifications of the same process. This seems a redundant rule, however, it is natural for our experiment, meaning that those attributes have simultaneously generated anomaly events that increased the Anomalies score of VM's above the acceptable degree.

Table II. Some Attribute Rules/Fingerprints.

Size	Rules	Participation Rate	KPI Degradation	Co-Occurrence
2	[guestfilesystem percentage, guestfilesystem usage]	0.695	69.27	22
	[guestfilesystem percentage, net bytesTx_average]	0.396	69.27	24
	[cpu usagemhz_average, guestfilesystem percentage]	0.362	69.27	30
	[cpu usagemhz_average, net bytesTx_average]	0.232	73.73	35
	[cpu usagemhz_average, diskspace activeNotShared]	0.216	72.51	28
3	[diskspace activeNotShared, net bytesTx_average]	0.20	72.51	22
	[diskspace activeNotShared, guestfilesystem percentage, guestfilesystem usage]	0.44	69.27	21
	[guestfilesystem percentage, guestfilesystem usage, net bytesTx_average]	0.4	69.27	24
	[cpu usagemhz_average, guestfilesystem percentage, guestfilesystem usage]	0.37	69.27	30
	[cpu usagemhz_average, guestfilesystem percentage, net bytesTx_average]	0.2	69.27	35
4	[cpu usagemhz_average, diskspace activeNotShared, guestfilesystem percentage, guestfilesystem usage]	0.22	69.27	28
	[diskspace activeNotShared, guestfilesystem percentage, guestfilesystem usage, net bytesTx_average]	0.201	69.27	23
	[cpu usagemhz_average, guestfilesystem percentage, guestfilesystem usage, net bytesTx_average]	0.20	69.27	35

5	[cpu usagemhz_average, diskspace activeNotShared, guestfilesystem percentage, guestfilesystem usage, net bytesTx_average]	0.17	69.27	36
---	---	------	-------	----

All those patterns that we exemplified are “consistent” transactions of attributes that when occurring with abnormality conditions make a precursor of SLO violations. This consistency is defined by the confidence/strength of association rules, as well as the rank scores of KPI degradation and co-occurrence.

Table III. Rules with maximum confidence transitions.

Attribute_Name	Other Attributes in the Rule
guestfilesystem percentage	guestfilesystem usage guestfilesystem usage_total guestfilesystem percentage_total guestfilesystem freespace_total diskspace activeNotShared
cpu usagemhz_average	cpu used_summation cpu numberToRemove virtualDisk usage virtualDisk totalReadLatency_average datastore maxObserved_Write
cpu idle_summation	guestfilesystem percentage guestfilesystem usage cpu wait_summation cpu usagemhz_average cpu used_summation

6 Conclusions

We described an association rules approach to performance problem fingerprinting in data centers. It enables learning patterns of atomic anomalies leading to losses in the system’s performance indicator. Ranking and rule-making criteria according to their impact magnitudes on the KPI are also indicated. In a real data set, we observed a kernel set of attributes to constitute the problem definitions codebook by and specific rules that can be recommended as problem fingerprints for alerting modules.

References

1. VMware vCenter Operations Manager. <http://www.vmware.com/products/vrealize-operations/>.
2. VMware vRealize Log Insight. <http://www.vmware.com/products/vrealize-log-insight>.
3. Marvasti, M.A., Poghosyan, A.V., Harutyunyan, A.N., and Grigoryan, N.M.: An enterprise dynamic thresholding system. In: 11th Int. Conference on Autonomic Computing (ICAC'14), pp. 129-135, June 18-20, Philadelphia, US (2014).
4. Poghosyan, A.V., Harutyunyan, A.N., and Grigoryan, N.M.: Managing cloud infrastructures by a multi-layer data analytics. In: Int. Conference on Autonomic Computing (ICAC'16), pp. 351-356, July 18-22, Wuerzburg, Germany (2016).
5. Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., Hovhannisyan, N.A., and Kushmerick, N.: On machine learning approaches for automated log management, *Journal of Universal Computer Science*, special issue on Collaborative Technologies and Data Science in Smart City Applications, vol. 25, issue 8, pp. 925-945 (2019).
6. Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., and Kushmerick, N.: Learning baseline models of log sources. In: *Collaborative Technologies and Data Science in Smart City Applications (CODASSCA)*, pp. 145-156, Sep. 12-15, Yerevan, Armenia (2018).
7. Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., Kushmerick, N., and Beybutyan, H: Identifying changed or sick resources from logs. In: *IEEE Int. Workshop on Foundations and Applications of Self Systems*, pp. 86-91, Sep. 3-7, Trento, Italy (2018).
8. Bodik, P: Automating data center operations using machine learning. PhD thesis <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/thesis.pdf>, University of California, Berkeley, (2010).
9. Cohen, I, Zhang, S., Goldszmidt, M., Symons, J., Kelly, T, and Fox, A.: Capturing, indexing, clustering, and retrieving system history. In Andrew Herbert and Kenneth P. Birman, editors, *Symposium on Operating Systems Principles (SOSP)*, ACM (2005).
10. Agrawal, R., Imieliński, T, and Swami, A.: Mining association rules between sets of items in large databases. In: *ACM SIGMOD*, p. 207, 1993.

Intelligent Troubleshooting in Data Centers with Mining Evidence of Performance Problems

Ashot N. Harutyunyan, Naira M. Grigoryan, Arnak V. Poghosyan,
Sunny Dua, Hovhannes Antonyan, Karen Aghajanyan, and Bonnie Zhang

VMware

{aharutyunyan;ngrigoryan;apoghosyan;duas;
hantonyan;kaghajanyan;bonniez}@vmware.com

Abstract. Identifying actual root causes of a performance issue within a modern cloud infrastructure with high level of scale, sophistication, and complexity, is a hard task. It is especially complicated to diagnose a service or infrastructure degradation of an unknown nature, when no active alert is enough indicative about potential sources (be it an object, its metric, property, or an associated event) of the problem. In such a situation, the data center administration is intuitively looking for changes in the system that might reveal the causative factors. This requires costly investigations and results in business-critical losses. Cloud management vendors are building visions around AI Ops-enabled automation of the entire workflow of root cause analysis and troubleshooting. We propose a solution towards such a vision which is based on hypothesis testing and machine learning approaches for automatic mining “important changes” of various kinds in behavior of data center objects across time and infrastructure topology. Those are the most relevant evidence patterns expected to explain the performance issue. Our current implementation which is integrated into vRealize Operations runs on the available three sorts of monitoring data – *metrics*, *properties*, and *events*. However, the full vision is to extensively include more observability provided by other cloud management tools vertically scaled to capture the depth of a specific dimension of the data center administration. The implemented module produces lists of recommended patterns across those three dimensions rank ordered subject to different criteria for each, such as confidence (p -value) provided by hypothesis testing and magnitude of change in the metric data, event’s sentiment score or abnormality degree, unexpectedness/entropy of property variations, etc. We describe the main analytical concepts behind the solution and demonstrate its validation in an application troubleshooting scenario.

Keywords: Automated troubleshooting and root cause analysis, AI Ops, evidence mining, change point detection, time series (TS), hypothesis testing (HT), p -value, sentiment analysis, entropy, ranking, machine learning.

1 Introduction

Automated root cause analysis (RCA) and troubleshooting of various performance problems in increasingly complex cloud computing environments becomes an AI Ops challenge (see Gartner’s definition [1] of the term). Utilizing the power of Machine Learning in specific use cases can greatly help especially when human-annotated data sets are available for training supervised learning models. However, there is a naturally severe lack of such data sets and, therefore, unsupervised learning is the predominant framework in RCA. The thesis [2] provides an overview to this problem in the cloud with useful references therein. At the same time, projects targeted on building self-driving data centers and based on reinforcement learning (project Magna [3]), is an important initiative which can potentially lead to an effective solution of the problem. With our current work we make the next step forward to enhance our cloud management solution with data science and ML approaches to automate discovery of potential sources of performance issues in the customer data center. This paper introduces a novel approach to the problem which integrates various types of monitoring data to generate potential root cause recommendations prioritized across different layers of data center administration within a single UX scenario. It might potentially evolve into a recommender system that also tracks user feedback (direct or indirect) for data labeling and for training supervised learning models, thus making RCA and troubleshooting highly effective and personalized to the user environment.

The service degradation or non-optimal performance can originate both from the infrastructure and application layers of the cloud system. When it occurs, the user goes through a typical process (Fig. 1) of troubleshooting consisting of several stages. This process is subject to full automation within an intelligent troubleshooting. To adequately approach the problem, different information sources (obtained from monitoring of various aspects of the data center deployment – metrics, logs, properties, events, application traces, net flows, etc.) need to be combined within an intelligent analysis in an automatic manner. It means that all cloud monitoring products (e.g., vR Ops [4], Wavefront [5], vR Log Insight [6], and vR NI [7]) could bring their insights into such an analysis within an integrated cloud management platform.

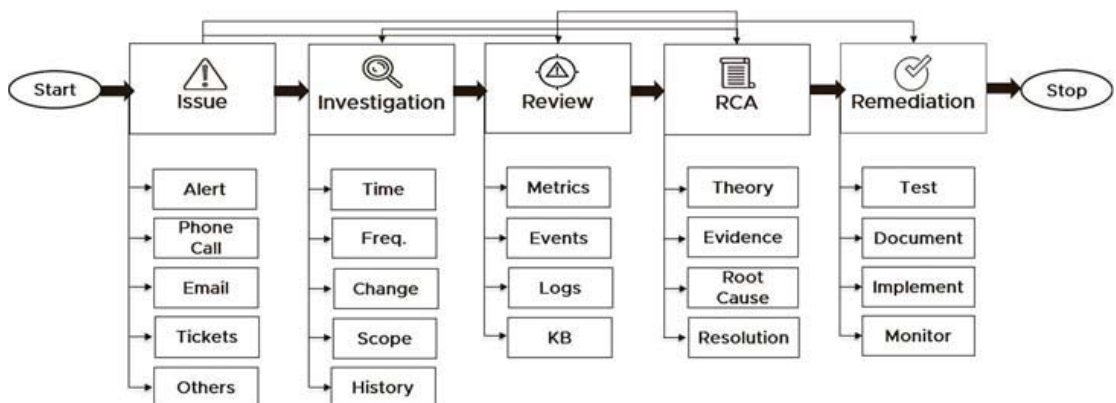


Fig. 1. Stages and modules of the troubleshooting process. Review layer might consist of all other data dimensions measured from the system.

With the current work, the following use case is addressed. There are performance issues of data center objects or related application KPIs, which are not self-explainable with the existing active alerts associated to those entities. It means the user faces an “unknown problem” for which there are no specific alerts/symptoms defined or available ones don’t point out the actual root cause (are effects or unrelated events). In RCA with an AI Ops product, the user expects a reference to the origin of the problem, at least automatic discovery/recommendation of the “candidate causes” to further investigate for the actual root cause. Our algorithms perform evidence mining – discovery of those candidate entities across different data types available in vR Ops, within a “relevant” time and topology scope for an object with a performance problem. These are the most important changes and potentially causative patterns preceded the issue the user wants to troubleshoot, prioritized according to “significance” of those changes. In Section 2 we motivate our work and discuss the prior art. Section 3 describes our algorithms for evidence mining based on HT and concepts for ranking detected patterns. Section 4 demonstrates how effective was the prototyped feature in an experimental diagnostic of an application performance issue. Section 5 contains information on user experience research and initial feedback from the customers. Section 6 concludes the paper with notes on the future work on enhancement of the solution.

2 Motivation and Prior Art

Various approaches have been developed for anomaly and change detection using history of monitoring data (both structured/metrics [8]-[9] and unstructured/logs [10]-[12]) to assist data center admins in faster RCA and troubleshooting. Prior related art in this domain was focused also on similarity analysis of data center incidents in their reoccurrence (see PhD thesis by Bodik [13], references therein, including a direct ascendant work by Cohen *et al* [14] on identifying *crisis signatures*). Compared to those specific settings, our view of the problem is based on integrating available information pipelines into a single troubleshooting platform that consolidates and prioritizes recommendations in any situation the user is concerned about the performance of data center objects. Fig. 2 demonstrates our AI Ops vision consisting of four main layers:

- *measuring* and selecting data for analysis,
- *discovery* of problem signals in time and topology scopes,
- *learning* importance of patterns,
- *ranking* those by various criteria.

A partial realization of this generic vision into vR Ops leverages only three types of data managed by the product – events, properties, and metrics. Specifically, it detects and visualizes change behaviors occurred in metrics, “unexpected” changes in properties and “interesting” (of highly negative sentiment) events within the time and topology proximity of the issue. In this troubleshooting workflow, the scope of objects, as well as time, to investigate based on topology hierarchy can be defined automatically (default setting), but it is also adjustable by the user. This automatic identification tries to capture the “problem coverage zone” across time and topology, determined by co-located alerts start time and their topology relationships/closeness. The

implementation recomputes/refreshes evidence collections upon user tuning of those default scopes.

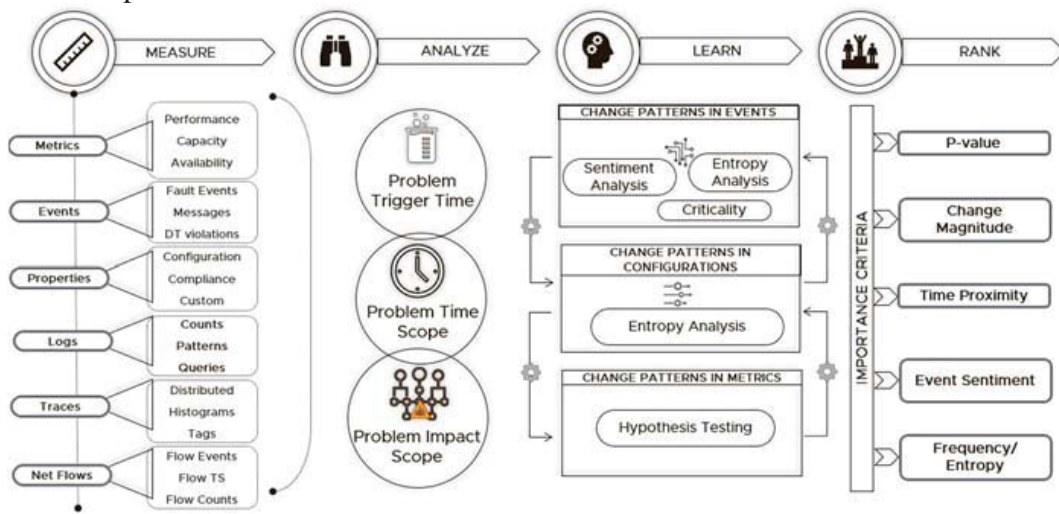


Fig. 2. AI Ops troubleshooting workflow for automated analysis of interesting patterns from all data dimensions.

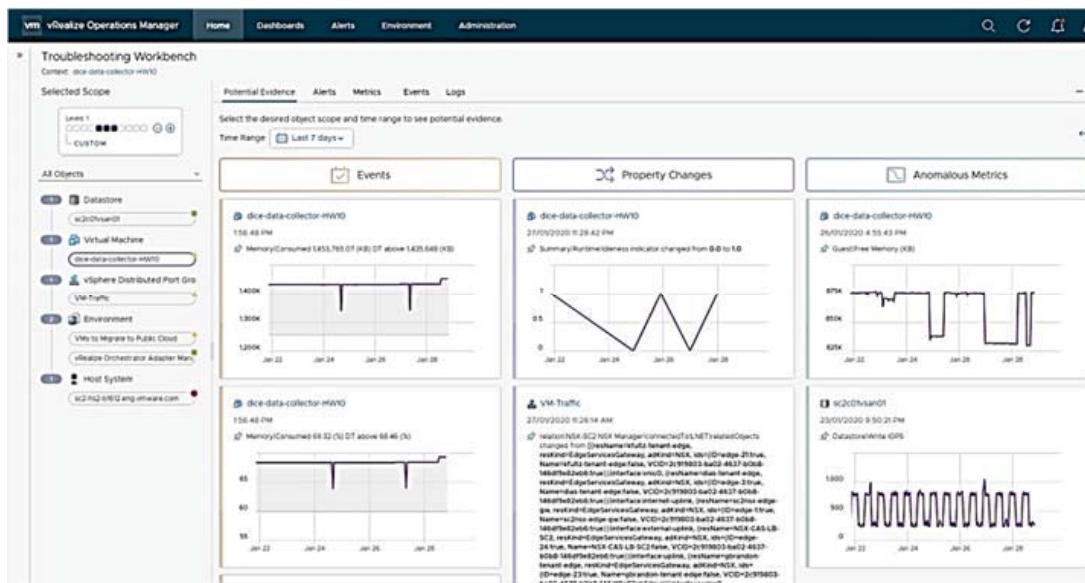


Fig. 3. Intelligent troubleshooting workbench in vR Ops: Potential evidence patterns by events/properties/metrics are shown.

3 Discovery of Evidence Patterns

3.1 Evidence in Metrics

For the identified problem coverage zone (both learned or adjusted by the user), a change point detection algorithm is applied to all metrics of objects within the related time and topology scopes. We assume that when the user is troubleshooting an issue

that is still active, spiky behaviors in the metrics are not very much interesting to the user. Instead, changes indicating distribution shifts (or any of its attributes like mean, variance, median, etc.) in the time series data are important. One realization of the change detection is based on Pettitt test [15]. Our algorithm iterates over data points through a sliding window,

- measures the *test statistic* (sign-based measure of data value difference) for the datapoints on the right and left windows;
- calculates the corresponding *p*-value (there is a very strong evidence on rejecting the null hypothesis, which is the “no change” hypothesis, for $p \leq 0.01$);
- identifies the most significant change according to the smallest *p*-value and/or test statistic.

The statistic computed for metric values in the left-hand and right-hand window is given by $U_{t,T} = \sum_{i=1}^t \sum_{j=t+1}^T D_{ij}$, where

$$D_{ij} = \text{sgn}(x_i - x_j) = \begin{cases} 1 & x_i < x_j \\ 0 & x_i = x_j \\ -1 & x_i > x_j \end{cases}$$

x_i 's are metric values in the left-hand window;

x_j 's are metric values in the right-hand window;

$1 \leq t < T$;

t is the largest time value in the left-hand window;

T is the number of points in the sliding time window.

Pettitt's non-parametric test statistic for the sliding time window is given by $K_T = \max_{1 \leq t < T} |U_{t,T}|$. A *p*-value of the non-parametric test statistic K_T is given (and justified in literature experimentally for specific data sets) by $p \cong 2 \exp\left(\frac{-6(K_T)^2}{T^3 + T^2}\right)$. A change point at the time t is significant when p is smaller than some confidence threshold. Further details on the algorithmic part include:

- we exclude those metrics that are “accumulative”, trendy (e.g. uptime metrics) by checking whether change was declared on all steps when we perform HT;
- for each of the changed/anomalous metrics, we compute magnitude of the change within the time scope, measured by difference of medians between the left and right windows of the change point, normalized over the range of the whole data. This way we filter out those metrics which experienced a distribution change but not a significant shift in the data range.

Another change point detection algorithm which is non-parametric and reasonable in terms of implementation feasibility with *randomization* for short time series data, and *p*-value estimation is based on Permutation Test [16]. The *p*-value can be combinatorically obtained by the following formula

$$p = \frac{1}{M!} \sum_{j=1}^{M!} I(\text{Test}_j > U_{\text{obs}})$$

where M is total number of data points within the left and right windows, which form $M!$ possible permutations, $Test_j$ is K_T statistic for time instance t for the j th permutation of the observation data, U_{obs} corresponds to the same statistic of the observation data in the regular order in the time series, and I is the indicator function. So, if the formula gives us a very small value, it means the original order is not likely to be from the same distribution for the left and right windows, and a “change of distribution” hypothesis should be accepted. Then metrics mined according to the above-mentioned procedures can be ranked according to p -value increase, as well as time-closeness to the issue detected (alert fire time, KPI deterioration, any other signal/side info referring unhealthy state).

3.2 Evidence in Events

Our module analyses events space in the same problem coverage zone to mine more evidence which can add extra insights along the changed metrics. It automatically queries all events that are active during the troubleshooting time frame. All types of events can be considered (Faults, Change Events, Notifications, Dynamic Threshold violations [8], etc.) except, for instance, those which are subject to symptoms included in alert definitions (for example, Hard Threshold violations). This is because in case of an "unknown" issue, the alerts appeared in the system are not self-explaining the cause of the problem. We apply sentiment analysis to narrow down the space of potential evidence patterns. The algorithm excludes the events with highly positive sentiments (from a predefined library), for instance: “completed with status ‘success’”, “restored”, “succeeded”, “sync completed”. Then, candidate evidence patterns are ranked according to the following criteria:

- *sentiment score* [-1,1] (from very negative to neutral (0) to very positive);
- *criticality* level – "0-25" are for info level events, "26-50" for warning, "51-75" for immediate, "76-100" for critical;
- *status* of the event (active or cancelled);
- *closeness* of the event to the problem start time;
- *frequency* of event: its occurrence during the troubleshooting time frame;
- *entropy* [17] of event (how unexpected/rare/uncertain or usual/expected is the event) for the object or application component (measured by $-\log p$, where p is the relative frequency of the change). Rare events get higher rank according to the entropy criterion, meaning uncertainty might imply higher risk or stronger evidence.

3.3 Evidence in Properties

Across this dimension, all configuration/compliance changes in property data within the time frame of interest are discovered. These are Boolean metrics or counter metrics. For importance ranking of property changes the following criteria are applied:

- *time-closeness* to the reported issue;
- *frequency* of the property (or property type) change within the troubleshooting time window;

- *entropy* of the property change (how rare or usual is this change) for the object or application component. Rare changes get higher rank according to the entropy concept.

3.4 Extensions of Evidence Mining

The proposed troubleshooting solution will substantially benefit from integration with other management products to include into analysis also *log data*, *network flows*, and *application traces*. For those data kinds we are developing specific algorithms to mine important patterns of interest with similar ranking measures explained in case of metrics, events, and properties.

Interesting patterns in log data. Log data of infrastructure objects and application components within the investigation scope might contain important evidence about the performance issue. The following signals are of interest:

- *trending* error/warning/info messages from log stream of object or application;
- *topics* detected in log messages correlated to other evidence discovered;
- *anomalous divergences* in event type distributions by vR LI [12];
- *deviations* from baseline event type distributions [10]-[11].

Interesting patterns in network flows. vR NI [7] provides network-related diagnostics of the system. Here important patterns are bottleneck flows, change points in flows detected using our analysis for metric data.

Interesting patterns in application traces. Wavefront's [5] distributed application tracing provides us with an extra dimension for this evidence discovery analysis. Potential evidence patterns might include traces that are "atypical" or simply of low-frequency signatures (entropy concept).

4 Experimental Evaluation

4.1 Background

During the validation phase of the troubleshooting workbench inside vR Ops, multiple experiments with real-life use cases were conducted to arrive at the potential root cause for issues which customers face on a day to day basis. The simulation of such use cases leveraging a real application and building real life conditions helped mimic what customers go through. The goal of this exercise was to measure the effectiveness of the capabilities of the troubleshooting workbench including automated scope definition, time proximity and most importantly the relevance of the mined evidences by the system.

4.2 Simulation Candidate

In one such experiment a real-life use case associated to a media services provider was simulated. The company ran a three tier CRM application consisting of a Web, App & DB on a VMware SDDC infrastructure. Within this CRM application a home-grown

survey application for running seasonal marketing campaigns was heavily leveraged by the marketing function. For a holiday season marketing campaign, a survey was rolled out to thousands of subscribers for critical inputs into the product and sales strategy. While the scale and load test of the survey application were successful, on eventual roll out in production, the application was extremely slow and often resulted on a 404 error for the end customers resulting in a kiosk in the marketing and line of businesses. The eventual root cause found by the organization was a rouge maintenance script which moved the Virtual Machine disk of one of the survey app VM to a local datastore which was unable to sustain the http requests coming from the web. The amount of time spent by the organization to find the root cause and remediate the issue took around 68 hours. This downtime of the application resulted in a survey drop rate of approximately 37% which was a major setback for the firm as inputs from many subscribers was missing.

4.3 Experiment Details

Using open source CRM and survey components, a 3-tier application named “Shudder-CRM-App” was deployed on a VMware SDDC environment backed by vSphere, NSX and vSAN. Using the open source survey module running on a virtual machine, a new survey was created to be rolled out to end users. The underlying resources deployed for the survey application could support up to 1500 concurrent users.

To generate the load equivalent to the real-world, a couple of tools were used. A web server stress tool was leveraged to generate http web requests on the survey URL. In order to simulate the real-world scenario, the VM was migrated over to a local datastore when the number of simulated users reached close to 450 users.

In addition to the application load, external load was generated by using synthetic I/O on the local datastore using I/O Meter to help create potential bottlenecks which could be detected as evidence using the change point detection algorithms. Upon reaching close to 500 users, the web service hosting the survey crashed and the users resulted in getting errors related to URL taking too long to respond. From this point on, to verify the evidence gathering capabilities of the troubleshooting workbench, the application in question was searched within vR Ops. Upon launching the troubleshooting workbench with the contextual application topology of the shudder application, potential evidence was presented along with signals of existing critical events which represented a high amount of storage read-write latency.

While the symptoms were clearly pointing towards a storage related issue, the key validation for the troubleshooting workbench capability was to find the potential evidence which resulted in the storage issue. In the workbench, several patterns pointed towards the potential root cause and the correlated consequences.

4.4 Root Cause Detection

The workbench was instrumental in pointing out at certain key evidence which helped validate the root cause by showcasing the key underlying changes resulting in a correlated event of storage performance degrading drastically. This was the root cause of the web application going down under drastic user pressure and underlying I/O bottlenecks. The first critical event which is a consequence of the issue, points at the storage outstanding I/O and Latency hitting the roof (Fig. 4). This was detected automatically as an evidence by the workbench using the change point detection algorithm.

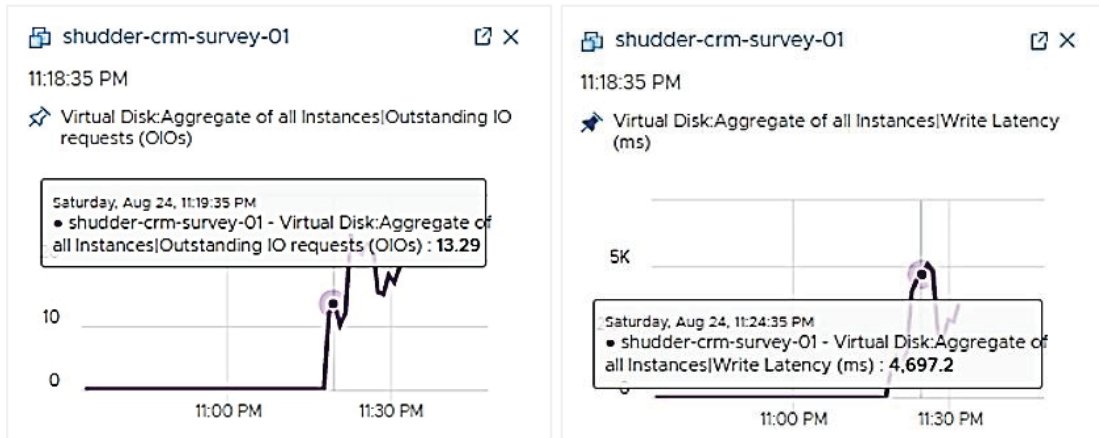


Fig. 4. Detected evidence showing jump in outstanding IO & Disk Latency.

Alongside the consequences, the key evidence of the root cause leading to this issue was listed. This root cause pointed out to a change which was triggered in the environment right before the KPIs were impacted and the application went down. This change was detected as a property change by the troubleshooting workbench with correlated timestamps for subsequent change points detected. Upon pinning the key evidences on a common scale, a perfect time and change pattern correlation was found across changes and causal evidence, hence solidifying the root cause of the problem (Fig. 5).

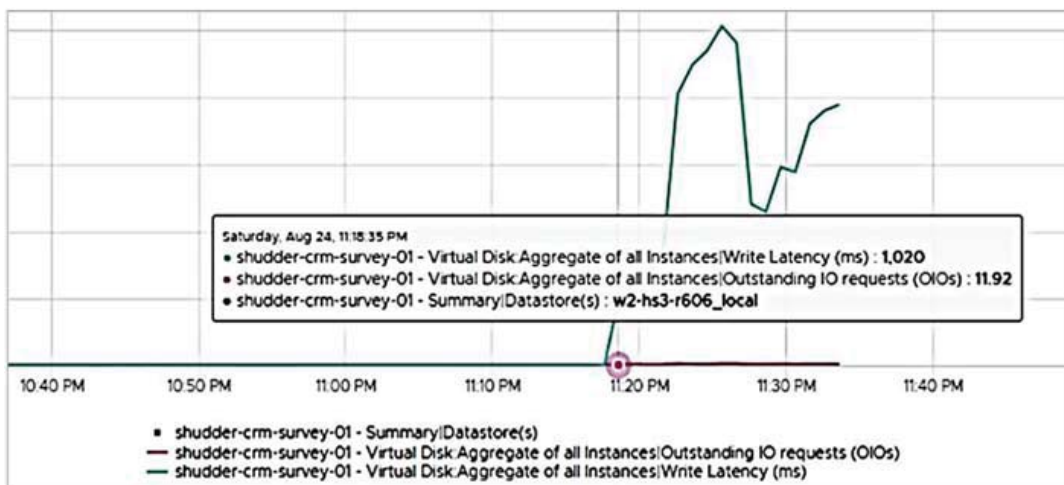


Fig. 5. Visual correlation of pinned evidences pointing towards the root cause.

The experiment proved the effectiveness of the troubleshooting workbench in detecting the root cause from thousands of metrics, events and log changes occurring in a dynamic environment over a large scope of objects hosted on a complex SDDC environment. The end to end issue detection, root cause analysis and remediation time reduced to a mere 30 minutes as compared to the 68-hour downtime faced by an equivalent application in real world environment, hence meeting the key objective of reducing the mean time to resolution (MTTR) and mean time to innocence (MTTI) with accurate and automated root cause analysis.

5 User Experience and Validation

5.1 Exploratory Research and Pain Points

Exploratory, qualitative research through contextual inquiry [18] was conducted to gain insights into existing user pain points while troubleshooting infrastructure and application issues. The research was conducted at with 25 users of various environment sizes. The goal for troubleshooting of the users is to reduce MTTR and MTTI; however, the users have too much data to look through and often find it difficult to understand the scope of the problem. Additionally, users complain of having too many context switches and depending on too many tools before finding the root of the issue. The general sentiment was there was too much tedious work involved in troubleshooting.

5.2 User Experience Considerations

Designing the user experience for presenting the evidence was important to ensure users can properly and effectively digest the intelligent data presented to them. The first determination was what 'knobs' to provide the user to help tune the evidence discovered and those were narrowed down to time and object scope, the two pieces of information that are crucial for troubleshooting. Another consideration was how a user interacts with individual pieces of evidence; a user will either find the evidence relevant or not relevant. To aid in the triaging of evidence, there is built-in functionality to pin or dismiss evidence.

To begin their flow, users have three distinct entry points into the troubleshooting workbench: through investigating an alert, through viewing details of a specific object, and by navigating directly to the troubleshooting workbench and searching for a target object. Once inside the troubleshooting workbench, the user experience allows for this following workflow, which matches the mental model of the troubleshooting process mentioned in Section 1.

- Step 1: View potential evidence (events, property changes, anomalous metrics).
- Step 2: Expand time and scope to view more evidence or reduce time and scope to narrow down evidence.
- Step 3: Pin evidence to metric viewer for comparison.
- Step 4: Investigate additional alerts, metrics, events, and logs within selected object scope.

5.3 Initial Validation

Following the general availability of the troubleshooting solution, initial validation with users was completed to understand the effectiveness of the new intelligence. A survey with an open discussion forum was held with 22 existing users, screened for those using the latest release of the product that premiered the troubleshooting functionality. Several participants have successfully used the troubleshooting workbench to troubleshoot and resolve a real issue. The functionality gained a satisfaction rating of 5.47/7 and a recommendation rating of 6.31/7. When asked to compare troubleshooting using the workbench compared to troubleshooting without it, 100% reported seeing improvement, with 87% noting significant improvement and 13% noting slight improvement. There was general positive sentiment on the functionality and the evidence presented; users did desire higher accuracy and less noise. Work can be done on both the user experience and the evidence mining to further reduce time required to complete troubleshooting issues.

Participants also validated thoughts on extending evidence mining with other management products (e.g., vR NI and LI). They also expressed interest in smart recommendations and more collaboration features. This feedback will help drive future work in this area.

6 Conclusion and Future Work

We introduced a novel intelligent troubleshooting framework for mining evidence of performance problems in data centers. It is based on combination of data science and ML algorithms to discover important patterns across various type of data that might explain the origin of the problem of an unknown nature. We also outlined how it can be further extended and enhanced. Initial implementation demonstrates significant power of the approach in automatically recommending accurate evidence in experimental application performance diagnostic and in real customer environments. Further plans include not only improving user experience on how indicatively we can organize the evidence patterns in terms of trend lining the evolution of the problem (their densities across time axis and across topology hierarchies), but also enhancing analytics power of the workbench in several directions:

- accurate learning of problem coverage zone is an important ML task, which will improve noise degree of our recommendations;
- incorporation of user feedback (ratings) on the recommended items collected over time would help us in filtering out non-indicative patterns (meaning, they are not likely to be causative);
- alternatively, ratings can be used in importance ranking. Moreover, ratings can be used for labelling data and training supervised ML models. Thus, we'll be able to identify/predict complex incidents composed of various type of evidence in the data. In this way, the algorithms will be tuned to the customer environment and application nature.

References

1. How to get started with AI Ops by Gartner: <https://www.gartner.com/smarterwithgartner/how-to-get-started-with-aiops/>.
2. Josefsson, T.: Root cause analysis through machine learning in the cloud: <https://uu.diva-portal.org/smash/get/diva2:1178780/FULLTEXT01.pdf>.
3. Tech preview announcement: Project Magna for vSAN continuous optimization: <https://blogs.vmware.com/management/2019/08/tech-preview-project-magna.html>.
4. VMware vRealize Operations Manager: <http://www.vmware.com/products/vrealize-operations.html>.
5. Wavefront by VMware: <https://www.wavefront.com/>.
6. VMware vRealize Log Insight: <https://www.vmware.com/products/vrealize-log-insight>.
7. vRealize Network Insight: <https://www.vmware.com/products/vrealize-network-insight.html>.
8. Marvasti, M.A., Poghosyan, A.V., Harutyunyan, A.N., and Grigoryan, N.M.: An enterprise dynamic thresholding system. In: 11th Int. Conference on Autonomic Computing (ICAC'14), pp. 129-135, June 18-20, Philadelphia, US (2014).
9. Poghosyan, A.V., Harutyunyan, A.N., and Grigoryan, N.M.: Managing cloud infrastructures by a multi-layer data analytics. In: Int. Conference on Autonomic Computing (ICAC'16), pp. 351-356, July 18-22, Würzburg, Germany (2016).
10. Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., Hovhannisyan, N.A., and Kushmerick, N.: On machine learning approaches for automated log management, *Journal of Universal Computer Science*, special issue on Collaborative Technologies and Data Science in Smart City Applications, vol. 25, issue 8, pp. 925-945 (2019).
11. Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., and Kushmerick, N.: Learning base-line models of log sources. In: *Collaborative Technologies and Data Science in Smart City Applications (CODASSCA)*, pp. 145-156, Sep. 12-15, Yerevan, Armenia (2018).
12. Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., Kushmerick, N., and Beybutyan, H.: Identifying changed or sick resources from logs. In: *IEEE Int. Workshop on Foundations and Applications of Self Systems*, pp. 86-91, Sep. 3-7, Trento, Italy (2018).
13. Bodik, P.: Automating data center operations using machine learning. PhD thesis <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/thesis.pdf>, University of California, Berkeley, (2010).
14. Cohen, I, Zhang, S., Goldszmidt, M., Symons, J., Kelly, T, and Fox, A.: Capturing, indexing, clustering, and retrieving system history. In Andrew Herbert and Kenneth P. Birman, editors, *Symposium on Operating Systems Principles (SOSP)*, ACM (2005).
15. Pettitt, A. N.: A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, pp. 126-135, 28(2), (1979).
16. Wasserman, L.: *All of Statistics: A Concise Course in Statistical Inference*. Springer, (2004).
17. Cover, T., Thomas, J.: *Elements of Information Theory*. Wiley, (1991-2006).
18. Contextual Inquiry: <http://www.usabilitybok.org/contextual-inquiry>.

Learning Data Center Incidents for Automated Root Cause Analysis

Arnak Poghosyan¹, Ashot Harutyunyan¹, Naira Grigoryan¹, Nicholas Kushmerick¹

¹ VMware, Inc.

{apoghosyan; aharutyunyan; ngrigoryan; nicholask}@vmware.com

Abstract. Identification of a problem fingerprint or incident in a data center is of crucial importance for the system administrators. Automated discovery of such important patterns in cloud environments is recently gaining a lot of popularity for effective and efficient root cause analysis of business-critical issues. A problem incident is a group of alerts with sufficient historical evidence in reoccurrence and similarity. Presumably, all known incidents should be stored in user's knowledge base together with available annotations regarding the problem description and its possible resolutions. The knowledge base of incidents with enough coverage of system's possible failures is an invaluable asset for any system administrator. It will help to detect and isolate a problem, accelerate its remediation. In many cases it will also help to anticipate upcoming performance degradations before they impact the system. This classical approach totally relies on authentic alert definitions which mostly require expert knowledge regarding the environment peculiarities which makes impossible automation of the root cause analysis of unknown and very complex systems. In this paper, we consider essentially different approach that bypasses the alert definition and management process. Application of rule-learning algorithms to system indicators defines appropriate incidents in terms of rules with enough statistical evidence. We consider different possibilities both with labeled and unlabeled metric spaces. The labels can be derived from users' feedbacks on system failures or performance degradations. In case of unlabeled metric space, outlier detection procedures can be applied for data labeling.

Keywords: root cause analysis, alert, incident, rules

1 Introduction

Management of modern cloud infrastructures and applications [1,2] heavily relies on monitoring of all available indicators in the form of time series, logs, traces, histograms and storing them for further analysis, visualization, and reporting. The most challenging vision of IT management is building of self-driving data centers capable for self-healing in case of service-critical issues. Root Cause Analysis (RCA) [3] is the important constituent driver of monitoring and management tools for problem detection, isolation, learning, and remediation. Accumulated historical experience while tackling similar

problems potentially allow proactive prediction of well-known issues or at least accelerated remediation.

There is a myth that the main goal of the RCA is finding the root (the main cause) of an issue which means in practice the problem identification, its assignment to the right administrator, and fast remediation [3,4]. Unfortunately, this ideal situation with the main root within our control rarely realizes in practice and many problems have different simultaneous causes. In many cases, the problem can be identified by controlling and managing the surrounding symptoms even without identified real roots. Possibly, the roots will be outside of our control and additional knowledge should be useless for a problem remediation. Classical example is the “fire triangle” [4]. Fire needs heat, fuel and oxygen. Actually, all three are simultaneous symptoms of the “fire” as a problem which can be resolved by managing only one of them without even knowing the next two ones and the real root cause.

As a result, many RCA vendors like VMware, BigPanda, and Moogsoft follow this practical guidance describing a problem by its surrounding symptoms, events, alerts within time and topology proximity named as an incident. According to Moogsoft an incident is a Probable Root Cause of a problem meaning that the surrounding symptoms can be helpful for the remediation. Then, it remains to carefully collect those historical incidents, annotate them and utilize for accelerated remediation of similar issues.

The main sophistication of this approach is understanding of the cause-consequence relationship of alerts or events supported with evidence [5]. In many cases, the cause-effect relationship is based on historical (statistical) evidence and user feedback. We see that the entire responsibility of the RCA was laid on the predefined events and alerts (see [5-9]). Hence, the power of the RCA is directly connected with the completeness and reliability of the entire alert space. Taking into account the variety of systems and users, the latest is always connected with some expert knowledge

In this paper (see also patent [10]), we aim to automate the entire process by skipping the alert management process and directly proceeding with incident detection via ML approaches. Rather than defining alerts and combining them into incidents, we find incidents via solution of some supervised or unsupervised problems. Those incidents can be applied to the alert definition problem if needed. We show some results of experiments when the time stamps of failures are unknown. Naturally, more accurate results are expected when information about the system failures is available.

2 The Approach and Related Work

Gartner anticipates [11] that “by 2023, 40% of DevOps teams will augment application and infrastructure monitoring tools with artificial intelligence for IT operations (AIOps) platform capabilities”. The most important capabilities of the AIOps platform are smart solutions for thresholding, anomaly detection, correlations, and causality with the final goal of fully automated RCA.

Many AIOps platform vendors like IBM, Facebook, Google, BigPanda, and Moogsoft have almost complete vision and solution for the domain-centric RCA based on

system KPI behaviors and related alerts [12-16] (see Fig. 1). They all follow the classical straightforward approach starting with alert definition, alert management, and incident detection. Incidents should be stored in a knowledge base for further utilization. Annotations can be assigned to known incidents that will help to identify unknown runtime problems. It will support RCA by accelerating a problem identification and resolution.

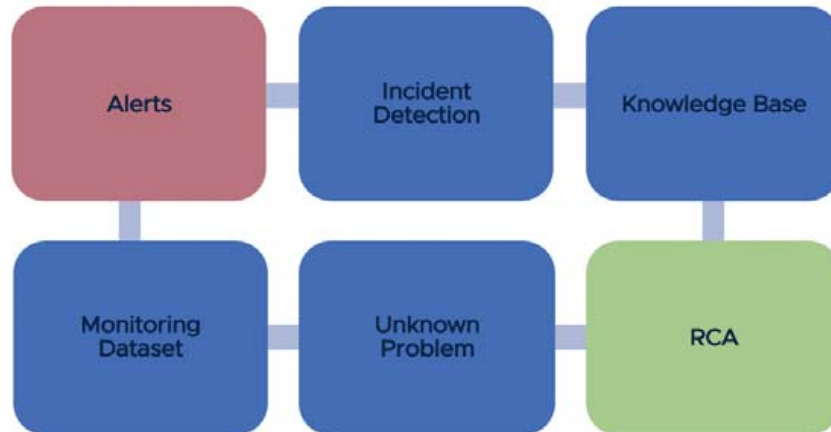


Fig. 1. Domain centric RCA.

Solutions that are based on alerts are domain centric as their definitions require expert knowledge of environmental indicators and their behaviors. Prior related art in this domain was mainly focused on similarity analysis of data center incidents in their recurrence ([17-19]).

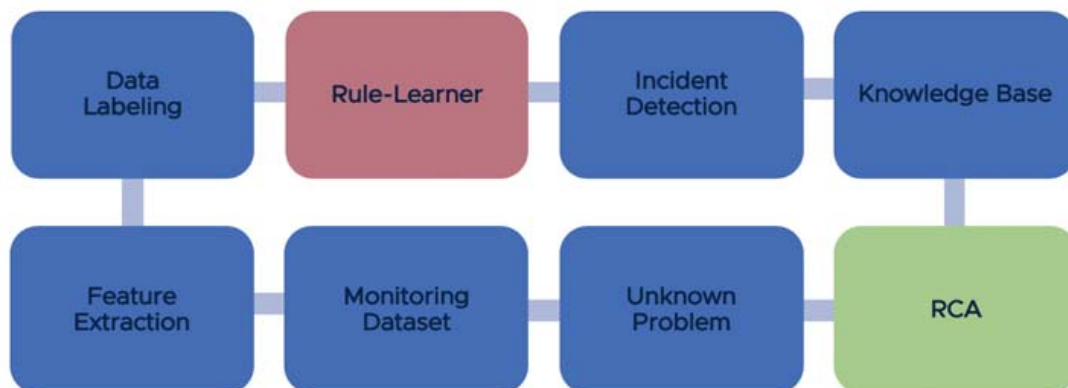


Fig. 2. Domain agnostic RCA.

Our approach differs in nature. It is domain agnostic. We learn important combinations (rules) of anomaly states of data center flows subject to performance degradations or system outlying behavior. The main purpose is automation of alert definitions in cloud management products and explainability of the situations. Let us discuss the variant of this approach when the information on system performance degradations is not available (see Fig. 2). We apply outlier detection for data labeling and rule induction

algorithms to the labeled data for learning incidents that correspond to outlying behaviors of the system.

Data monitoring overload is the major problem in autonomous management solutions where all kind of indicators from the infrastructure and application are measured for a full visibility into the cloud environment. Information overload leads to high level noise in data management systems making it unfeasible to rely on their recommendations. Wide range of procedures are known in the literature (see [20] with references therein) for noise or dimensionality reduction.

We apply principal component analysis PCA [21] to the initial monitoring space pursuing two important milestones. Firstly, principal components can substantially decrease the number of metrics by preserving the total variability of the space of initial metrics. Secondly, the principal components are uncorrelated and their group-wise violation from the historical baseline is highly improbable and indicate possible problems in the object. It means that application of outlier detection procedures to the principal component space is more logical rather than to the initial monitoring space.

As mentioned before, we discuss self-supervised learning problem where data labeling is performed via outlier detection. We successfully applied Local Outlier Factor (LOF) [22] and modified K-means [23] (we will call it as k-mod) which is more robust to outliers. Both approaches provide data time stamps with outlying behavior which we can use for data labeling both in the feature space and in the original metric space. We proceed with the modified K-means algorithm applied to the feature space composed of the principal components.

Further, we apply rule-learning systems like decision trees [24] or RIPPER [25] to the labeled data. Rule-learners have a very important property to show individual metric participations with the corresponding thresholds at the time stamps with outlier behaviour. Those groups of metrics with the corresponding threshold-rules comprise an incident accompanying a specific problem.

If classification is applied to the principal component space, the root-cause identification will be implicit via feature-space metrics. If classification is applied to the initial metric space, the root-cause will be explicit via original-space metrics. We can learn to resolve different incidents via historical analysis and apply it to run-time mode to accelerate the process of remediation. Further, in our experiments, we discuss only the procedure applied to the principal components.

3 Results and Discussions

We performed experiments for an application server with real customer workloads. The corresponding data-frame contained 1290 time series metrics all with the same lengths and time stamps. Each metric had 44392 data points with 1-minute monitoring interval. Hence, each time series had 1-month duration.

The dataset contains a lot of redundant information. Some of the metrics are almost constant and many are correlated or anti-correlated. Fig. 3 shows the correlogram of initially monitored time series data where “Vx” are the names of initially monitored metrics.

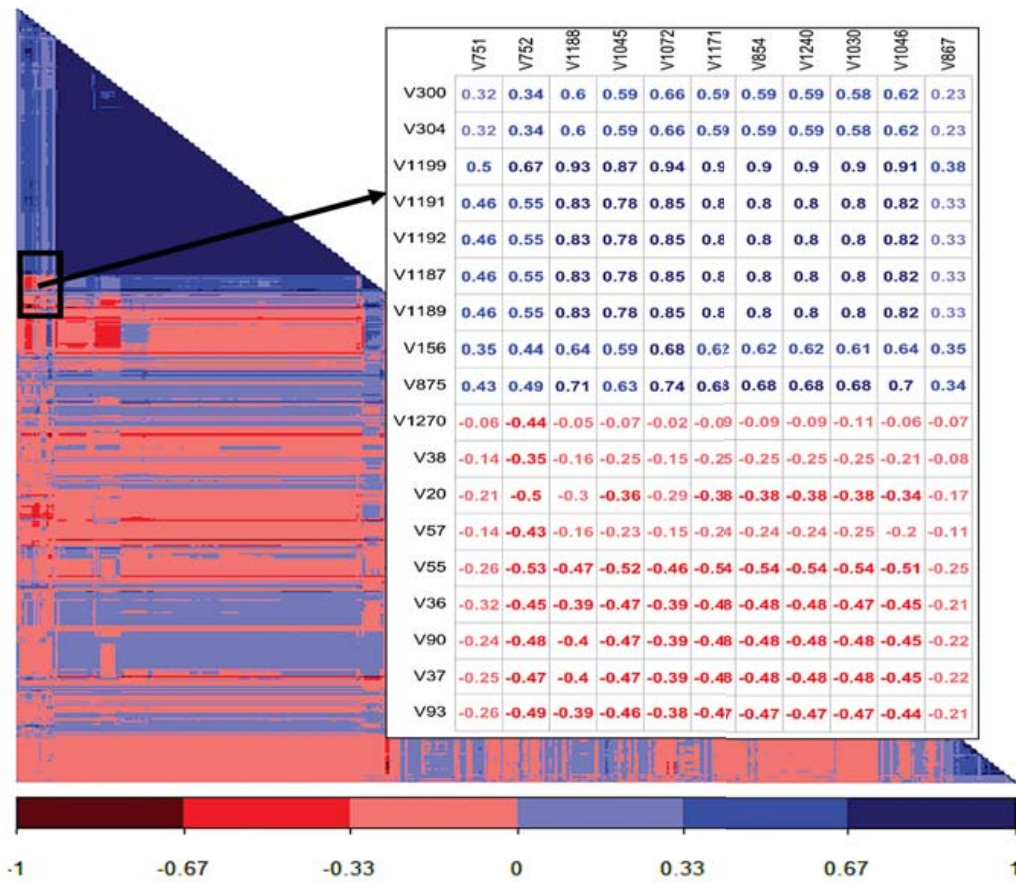


Fig. 3. Correlogram of initially monitored time series data.

For convenience, the correlogram rows and columns are reordered via hierarchical clustering with “complete” linkage and correlation coefficient as a distance measure. The resultant one (see Fig.3) is visually more suitable for estimating the percentage of correlated and anti-correlated metrics in the environment. In the figure, dark blue color corresponds to correlated, and dark red to anti-correlated metrics.

Fig. 3 shows that application of PCA for dimensionality reduction will be rather effective. As a result, PCA required 115 components (see Fig.4) to explain 90% of the overall variance of the initial metric space. The resultant compression rate is 91%. Worth noting that the PCA was applied to the normalized dataset with 0 mean and 1 standard deviation as the scales of the initially metrics can be drastically different.

Fig. 5 shows the result of application of the modified K-means (k-mod) algorithm, where the red points (asterisks) correspond to outliers (system “Abnormal” state) and the black points (dots) to non-outliers (system “Normal” state). Black data points correspond to the average of the principal components (the first 115 ones). One of the important parameters that k-mod requires is the percentage of outliers for detection. General recommendation is to use maximum 5%. We used 4%. Another important parameter is the number of clusters. We used 4 clusters. Further, those parameters should be tuned subject to maximization of the classification accuracy. For comparison, the green line (dashed) in Fig. 5 shows threshold $-mean(data) + stand.dev.(data)$.

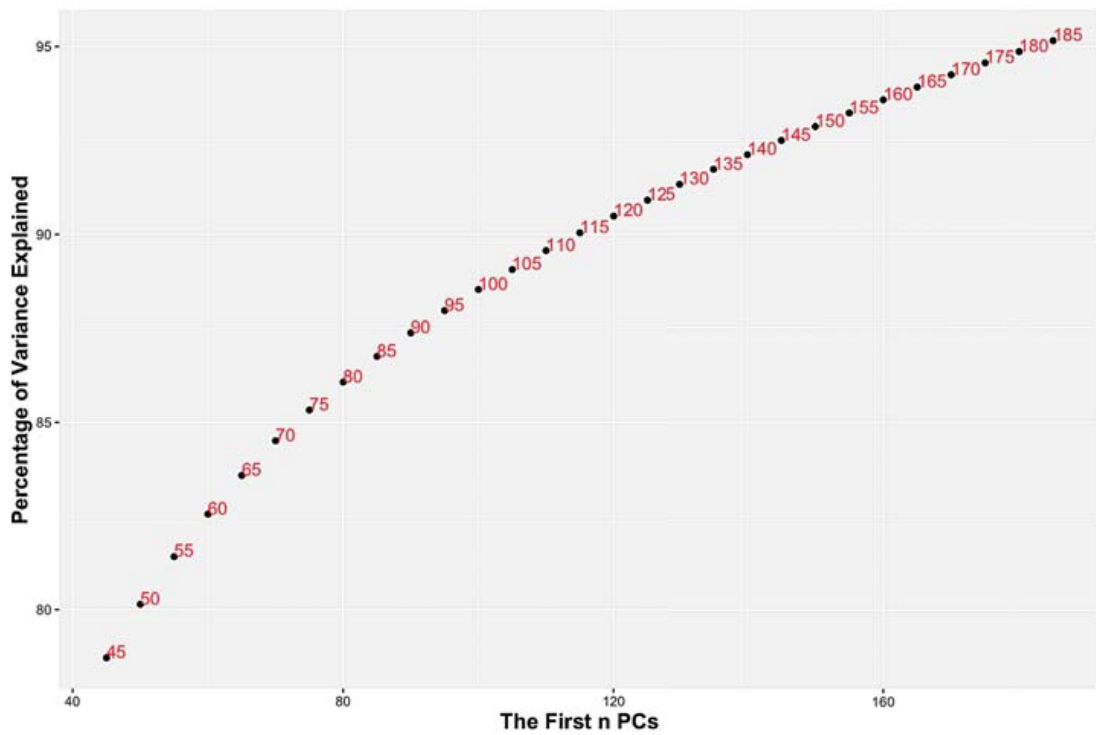


Fig. 4. Percentage of variance explained by the first n principal components.

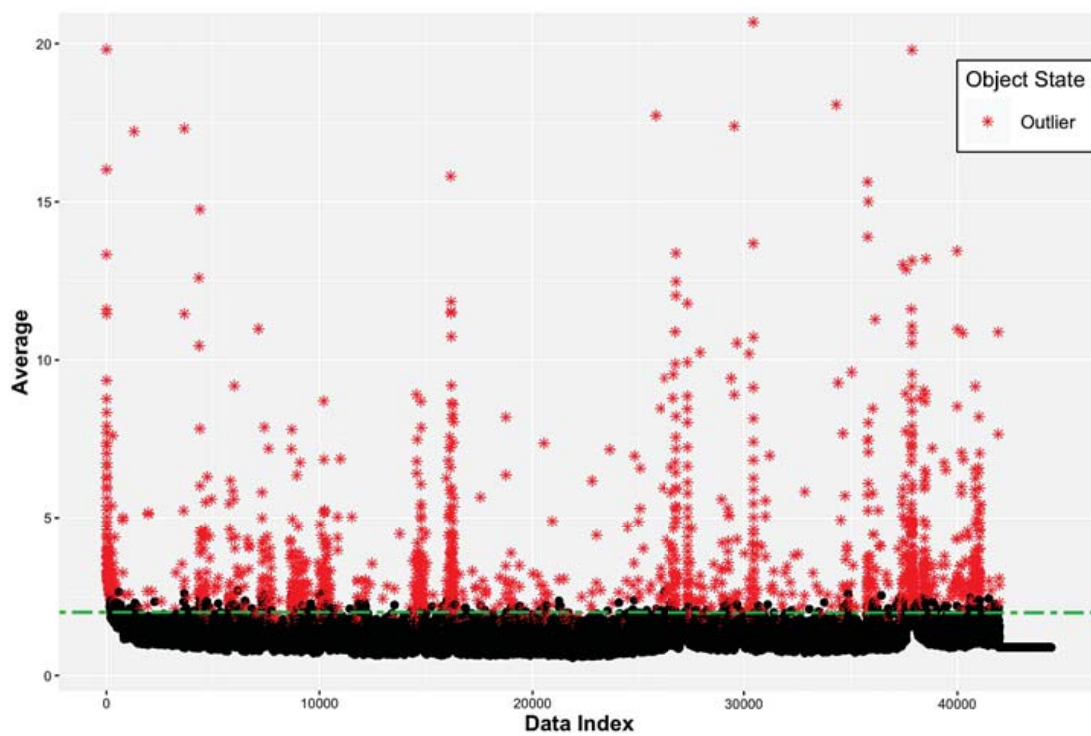


Fig. 5. The result of data labeling.

Then, we assigned “Y” labels to the outliers and “N” labels to the other time stamps and proceeded with the application of CART [24] (classification and regression trees) for rule detection. Fig. 6 shows the corresponding tree. For comparison, we applied also other classification approaches like SVM and C5.0 [24]. For each method, we calculated the corresponding precision-recall curves on a test data (20%). The highest performance had C5.0 algorithm. Less accurate was SVM and the worst was the CART. Experiments showed that the linear kernel is the best for the SVM algorithm. For example, C5.0 had sensitivity=0.91 and precision=0.95 for the positive class “Y”, while decision tree had sensitivity=0.84 and precision=0.83. Although decision trees had the worst predictive power, their outputs in the form of a tree (see Fig. 6) provided better interpretability of the results. The tree shows the groups of principal components with the corresponding thresholds that comprise a specific anomaly. Each sub-tree ending up with “Y” class comprises an incident with the corresponding set of principal components within the sub-tree and thresholds which simultaneous violations lead to an outlying behavior.

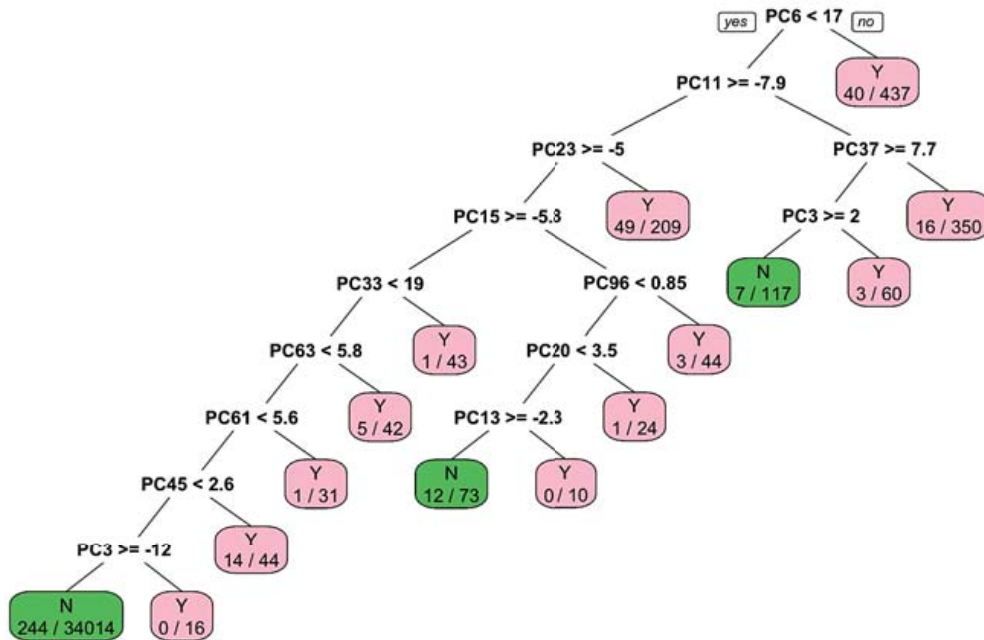


Fig. 6. Decision tree for a rule detection.

Table 1 shows 4 examples of such rules composed of principal components with some thresholds corresponding to Fig. 6.

Table 1. Incidents in terms of the principal components.

Incidents	Rules
1	$PC_6 \geq 17$
2	$(PC_6 < 17) \& (PC_{11} < -7.9) \& (PC_{37} < 7.7)$
3	$(PC_6 < 17) \& (PC_{11} \geq -7.9) \& (PC_3 < -5)$
4	$(PC_6 < 17) \& (PC_{11} < -7.9) \& (PC_{37} \geq 7.7) \& (PC_3 < 2)$

In case of the decision trees, the incidents are mutually exclusive, and an anomaly state can be identified uniquely. The drawback of this approach is that all incidents are starting with the same impactful metric (PC_6). Fig. 7 shows some of the incidents across the time axis from Table 1. Grey graph corresponds to the average of the initially monitored metrics while the incidents are composed of the principal components.

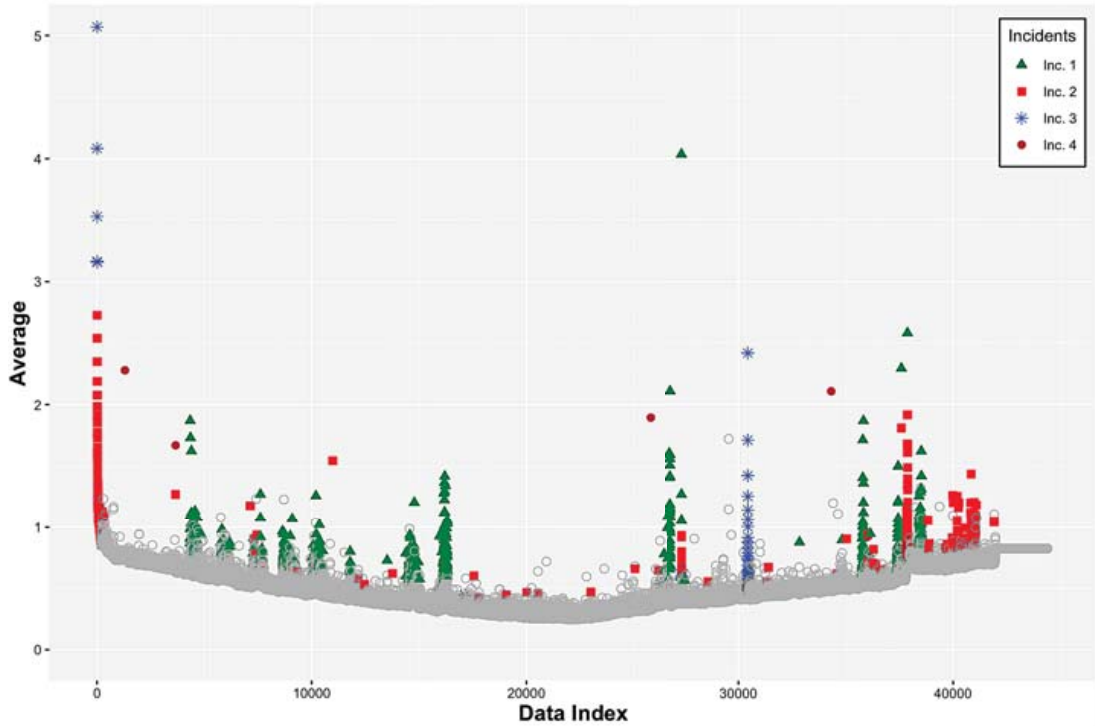


Fig. 7. Four different incidents (see Table 1) in different colors across the time axis.

4 Conclusion and Future Work

We saw that different outlying states (maybe also system anomaly states) were characterized by different rules (incidents) that described different behavioral peculiarities of system indicators. A system administrator can assign appropriate annotations to different incidents and apply them to run-time anomaly issues for problem identification and accelerated remediation. We discussed the case when information about system historical anomaly issues is not available and proceeded with data labeling via outlier detection procedures. We expect more accurate results in case of additional information on system historical performance degradations.

The core idea of the approach is application of rule-learning ML algorithms to incident detection problem. It assumes domain agnostic approach as no preliminary domain centric alerts are needed. Huge number of different rule-learners are available in the literature. We tried decision trees, not the most powerful ones, but the best in the sense of interpretability. In future, we will try to apply C5.0 and Ripper. The latest is the state-of-the-art of classification rule-learners.

In this paper, we used PCA for data labeling. Those labels can be applied to the feature space composed of principal components or to the initial monitoring space composed of system indicators. We tried the first scenario and got the rules in terms of the principal components. In future, we will try to get the rules composed of the system indicators. This approach will increase the entire interpretability of the method. Also, we will try other outlier detection approaches for data labeling.

References

1. IT Infrastructure Monitoring Tools, <https://www.gartner.com/reviews/market/it-infrastructure-monitoring-tools>, last accessed 2020/06/27.
2. Application Performance Monitoring Tools, <https://www.gartner.com/reviews/market/application-performance-monitoring>, last accessed 2020/06/27.
3. Schnepf, R., Vidal, R., Hawley, C.: Incident Management for Operations. First Edition, O'Reilly Media, 2017.
4. Galley, M.: What is root cause analysis? <https://blog.thinkreliability.com/what-is-root-cause-analysis>, last accessed 2020/06/27.
5. Marvasti, M.A., Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M.: Method and Apparatus for Root Cause and Critical Pattern Prediction using Virtual Directed Graphs. Filed Oct 12, 2011. Application No: 13/271,554. Published: Apr 18, 2013. Publication No: US 2013/0097463 A1. Granted: Jun 10, 2014. Patent No: US 8,751,867 B2.
6. Marvasti, M.A., Harutyunyan, A.N., Grigoryan, N.M., Poghosyan, A.V.: Methods and Systems to Manage Big Data in Cloud-Computing Infrastructures. Filed: Apr 30, 2015. Application No: 14/701066. Published: Nov 3, 2016. Publication No: US 2016/0323157 A1. Granted: Apr 17, 2018. Patent No: US 9,948,528 B2.
7. Marvasti, M.A., Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M.: Methods and Systems for Abnormality Analysis of Streamed Log Data. Filed Aug 6, 2013. Application No: 13/960,611. Published: Feb 20, 2014. Publication No: US 2014/0053025 A1. Granted: Mar 29, 2016. Patent No: US 9,298,538 B2.
8. Marvasti, M.A., Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M.: An Enterprise Dynamic Thresholding System. USENIX Int. Conf. on Autonomic Computing (ICAC), Philadelphia, US, June 18-20, 129-135 (2014).
9. Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Marvasti, M.A.: Data-Agnostic Anomaly Detection. Filed: Mar 29, 2013. Application No: 13/853,321. Published: Oct 2, 2014. Publication No: US 2014/0298098 A1. Granted: Mar 26, 2019. Patent No: US 10,241,887 B2.
10. Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Kushmerick, N.: Methods and Systems that Detect and Classify Incidents and Anomalous Behavior using Metric-Data Observations. Filed: Dec 10, 2018. Application No: 16/214,272. Published: Jun 11, 2020. Application No.: US2020/0183769 A1.
11. Gartner: Market Guide for AIOps Platforms. Published 7 Nov 2019 – ID G00378587. Available online: <https://www.gartner.com/doc/reprints?id=1-1XS12Z80&ct=191118&st=sb>, last accessed 2020/06/27.
12. Moogsoft Alternatives & Competitors, <https://www.g2.com/products/moogsoft/competitors/alternatives>, last accessed 2020/06/27.
13. Khanna, S.: A Journey Through IT Incident Management. Available online: <https://www.moogsoft.com/blog/aiops/journey-through-incident-management/>, last accessed 2020/06/27.

14. An AI-powered IT Incident Resolution Application, Fueled by Your Own Data, https://www.ibm.com/watson/assets/duo/pdf/WDDE814_IBM_Watson_AIOps_Web.pdf, last accessed 2020/06/27.
15. Arbisman, M.: How Google & Facebook Approach IT Incident Management at Scale, <https://www.moogsoft.com/blog/aiops/google-facebook-incident-management-scale-in-sights-srecon-2016/>, last accessed 2020/06/27.
16. Arbisman, M.: The Data is the Model: The Future of IT RCA & Event Correlation, <https://www.moogsoft.com/blog/aiops/data-model-future-root-cause-analysis-correlation/>, last accessed 2020/06/27.
17. Bodik, P: Automating Data Center Operations using Machine Learning. PhD Thesis, University of California, Berkeley (2010).
18. Cohen, I., Zhang, S., Goldszmidt, M., Symons, J., Kelly, T., Fox, A.: Capturing, Indexing, Clustering, and Retrieving System History. In Andrew Herbert and Kenneth P. Birman, editors, Symposium on Operating Systems Principles (SOSP), ACM (2005).
19. Rozsnyai, S., Slominski, A., Lakshmanan, G.T.: Discovering Event Correlation Rules for Semi-Structured Business Processes. Proceedings of the Fifth ACM International Conference on Distributed Event-Based Systems, DEBS 2011, New York, NY, USA, July 11-15.
20. Wang, G., Yang, Ji.Y., Li, R.: UFKLDA: An Unsupervised Feature Extraction Algorithm for Anomaly Detection under Cloud Environment, <https://onlinelibrary.wiley.com/doi/10.4218/etrij.2018-0475>, last accessed 2020/06/27.
21. Jolliffe, I.T.: Principal Component Analysis. Series: 2nd ed., Springer Series in Statistics, Springer, NY, 2002, XXIX.
22. Breunig M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: LOF: Identifying Density-Based Local Outliers. SIGMOD Rec. 29.2, 93–104 (2000).
23. Chawla, S., Gionis, A.: k-means--: a unified approach to clustering and outlier detection. Proceedings of the 2013 SIAM International conference on Datamining. <https://epubs.siam.org/doi/pdf/10.1137/1.9781611972832.21>.
24. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, 2013.
25. Cohen, W.W.: Fast Effective Rule Induction. Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, July 9–12, 115-123 (1995).

In September 2020, researchers from Armenia, Chile, Germany and Japan met at the American University of Armenia for a virtual conference to discuss technologies with applications in smart cities, data science and information theory approaches for intelligent systems, technical challenges for intelligent environments, smart human centered computing, artificial neural networks, and deep learning. This book presents their contributions to the 2nd CODASSCA workshop on collaborative technologies and data science in artificial intelligence applications, a highly topical issue in today's computer science.

AUA American University
of Armenia

UNIVERSITÄT
DUISBURG
ESSEN

Open-Minded



Logos Verlag Berlin

ISBN 978-3-8325-5141-4