On Clustering Properties of Hierarchical Self-Organizing Maps

Jouko Lampinen and Erkki Oja

Lappeenranta University of Technology
Department of Information Technology
PO Box 20, SF-53851 Lappeenranta, Finland

E-mail: `Jouko.Lampinen@lut.fi`, `Erkki.Oja@lut.fi`

Tel: +358 53 571 3448 (Lampinen)
+358 53 571 2821 (Oja)
Fax: +358 53 571 2899

Topic category: Pattern Recognition

1 Novelty of the paper

2 Last paper on the same topic

The overview of the neural object recognition system being developed by the authors is presented in [3]. The feature extraction block of the system contains a hierarchical SOM clustering. The focus of the paper is on the classification part; the clustering properties of the HSOM are not analyzed.

3 Summary of the paper

Number of layers in feedforward neural network has a major effect on the forms of decision boundaries that the network can implement. A one layer network can only divide the space by hyperplanes, while a two layer network can produce any arbitrarily complex boundaries. The same principle is shown to apply to self-organizing maps also, so that one layered map divides the space into convex polyhedral regions so as to minimize the squared distance of data samples from centers of the regions (represented by the reference vectors on the map). A two layered map, where output values of the map are quantized by another map, can produce any arbitrarily complex forms in the space. The tessellation of a 2-layer map minimizes quantization error in the case that random errors are introduced to the quantized codes prior to decoding [4].

In the paper, the properties of 2-layered self-organizing maps are studied in the application of clustering sensory information. The basic function of the hierarchical map is as follows. A convex region of the first layer map (or hidden map) is mapped to each unit on the second layer map. Thus, each cluster contains all the convex regions mapped to any of the hidden map units inside the cluster. The cluster is now represented by several reference vectors instead of just one as in the one layer map. The units in the hidden map $\mathbf{m_r}$ are bound together so that they minimize the cost function

$$D(\mathbf{m}) = \sum_{\mathbf{q}} \sum_{\mathbf{r}} h(\mathbf{r}-\mathbf{q}) \sum_{\mathbf{x} \in A_{\mathbf{q}}} p(\mathbf{x}) \|\mathbf{x}-\mathbf{m_r}\|^2 \qquad (1)$$

where h is the neighborhood function, $p(\mathbf{x})$ is the probability density of samples $\mathbf{x}$ and $A_{\mathbf{r}}$ is the set of points $\mathbf{x}$ that are mapped to unit $\mathbf{r}$. $A_{\mathbf{r}}$ is a convex volume limited by normal hyperplanes of nearest units in the space. The equation means, that the cost introduced of a data sample $\mathbf{x}$ is a weighted distance of the sample from the nearest units along the lattice. In clustering terminology, this is the distance measure used to define distance of a sample from a cluster, and the 2-layer map roughly minimizes sizes of the formed clusters. Since the connections in the hidden map reach to adjacent clusters, the global intercluster size is not exactly minimized. The hidden map connection could be cut from cluster boundaries. However, keeping the connections ensures that the global similarity relations of the data samples are reflected on the hidden and output maps. The distance measure is adaptive to the local statistics of the data, since the weighting function h is defined in map coordinates. Because the units in the hidden map are almost equiprobable, the distance function takes into account roughly the same amount of probability mass in every point in the space.

Clustering sensory information

The special feature of sensory information is that the signal space is very high dimensional, but the actually occurring signals are usually rather low dimensional, since the primary feature detectors tend to be orthogonal or otherwise mutually exclusive. As an example, the space spanned by a set of orientation and frequency specific Gabor-filters is studied and shown to contain long low dimensional clusters of data samples, corresponding to edges in certain orientations, regular textures etc. Thus the clustering implemented by HSOM should be efficient in projecting the essential structure of such signal space onto low dimensional cluster space.

3.1 Experimental Results

In an experiment with artificial data we generated random elongated Gaussian classes and measured how the samples from different classes distributed among the developed clusters. We measured classification errors (by labeling the units first) and the entropies of the distributions. The SOM performed similarly as the well known k-means clustering (as expected, since the k-means is a batch training version of the SOM without neighborhood). HSOM was found to outperform the both methods and also the much more laborious ISODATA clustering algorithm both in classification error and entropy measures. The main problem with the powerful ISODATA is the large variations is class variances, while the self-organizing maps adapt to the local variances easily without any tedious tuning of the system parameters.

As a practical example we Gabor-filtered the test images with filters in 8 different orientation and 2 scales one octave apart. Then the 16 D vectors in each pixel were mapped with 100-unit SOM to feature values. Another set of features was obtained by first mapping the both resolutions by a 100 unit map and then clustering the map outputs by a second 100 unit map.

Distortions in the imaging (i.e. different lighting, contrast, viewing angle, etc.) make the Gabor filter responses to move in the 16 D space in regular way. If the clustering of the space reflects that structure, the same detail in the image should map to the same cluster even in somewhat different imaging conditions.

According to our experiments the feature set produced by HSOM is indeed clearly more distortion tolerant than that of direct SOM. An example of the tests is in table 2. We compared the features by compiling normalized histograms of the features over each image and computing inner products of the histograms. (The most simple case of the subspace classifier.) One of the test image from each class was selected as the class prototype and the other images were compared to the prototypes.

From the results in table 2. it is clear that the features of the HSOM perform much better. With SOM features all the test images cannot be even separated correctly. In larger experiments we have been able to classify 19 out of 20 similar faces using only the histograms of the 100 hierarchical features and the subspace classifier.

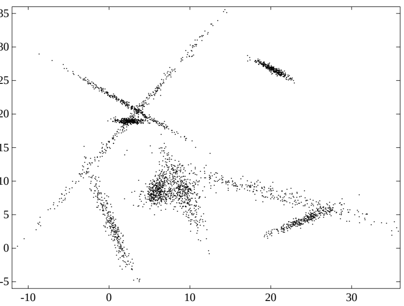

Figure. 1. Example test set of 10 clusters

Method	Errors %
k-means	30.3
SOM	30.7
ISODATA	28.0
HSOM	26.3

Table 1. Classifcation errors in the clustering tests

4 Other relevant works

In [4],[5], [6] the theory of hierarchical maps (quantization of map outputs with other map) is derived from principals of coding theory. There it is shown that the hierarchical map minimizes the decoding squared error if the training neighborhood in SOM equals the probability distribution of errors in the codes. In hierarchical VQ the higher order map quantization error is the source for the lower order map code errors. In this context the hierarchical SOM is optimal in quantization. However, relation of the method to clustering, and the shapes of regions formed is not discussed. Also, the work does not relate the method to pattern recognition, and distortion tolerance.

With regard to classification there are very few works with hierarchical maps. In [1] the LVQ method is pruned by adding successive LVQ classifiers that take all the outputs of the previous classifier as the input signal. Experimental results indicated some advantage over one layer LVQ.

Hierarchical SOM			
	T	J	P
T1	1.000	0.212	0.212
T2	0.550	0.295	0.508
J1	0.212	1.000	0.730
J2	0.291	0.814	0.717
P1	0.212	0.730	1.000
P2	0.231	0.703	0.860

SOM			
	T	J	P
T1	1.000	0.515	0.627
T2	0.790	0.504	0.698
J1	0.515	1.000	0.933
J2	0.572	0.984	0.980
P1	0.627	0.933	1.000
P2	0.568	0.985	0.971

Table 2. Comparison of SOM and HSOM features (see text for details)

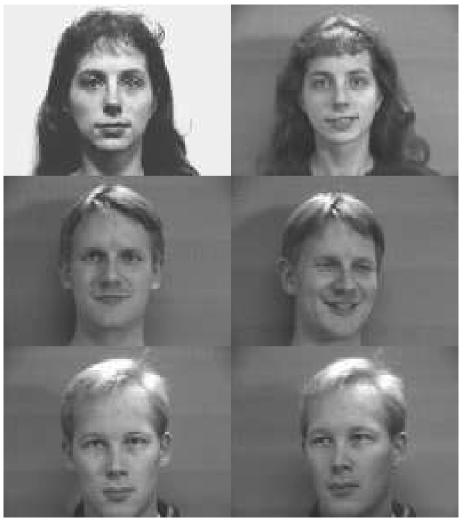

Figure. 2. Test images of Table.2.

In [2] a distantly related configuration is proposed for representing time sequencies on the SOM. Response of a SOM is integrated over time and the outputs of all the units are quantized by a higher order map. Here, as in the previous reference, input for the higher order map is a very high dimensional vector containing the activation values of all the units, raising the problem of dimension expansion when inputs from several data sources are converged to one representation, as noted in [2].

References

[1] Gyorgy Barna, Unpublished talk in Finnish-Hungarian Workshop on Neural Networks and Signal Processing, Tihany, Hungary, Oct. 1989.

[2] Jari Kangas, Representation of Time-Dependent Signals in Self-Organizing Maps, Lic. thesis, Espoo, 1990.

[3] Lampinen J. Distortion Tolerant Pattern Recognition Using Invariant Transformations And Hierarchical SOFM Clustering, Artificial Neural Networks, *Proc. ICANN-91*, Vol. 1., pp. 99-104, North Holland, 1991.

[4] S.P. Luttrell, Self-Organisation: a derivation from first principles of a class of learning algorithms, Proc. IEEE IJCNN-89 Washington DC, Vol.2. June 18-22, 1989, pp.495-498.

[5] S.P. Luttrell, Hierarchical vector quantisation, IEE Proc. Vol. 136, Pt. 1, No. 6, Dec 1989, pp. 405-413.

[6] S.P. Luttrell, Image compression using a multilayer neural network, Pattern Recognition letters 10, July 1989, pp. 1-7.