An approach to automated interpretation of SOM

Markus Siponen[1,2], Juha Vesanto[1], Olli Simula[1] and Petri Vasara[2]

(1) Neural Networks Research Centre,
Helsinki University of Technology,
P.O.Box 5400, 02015 HUT, Finland
{Juha.Vesanto,Olli.Simula}@hut.fi
(2) Jaakko Pöyry Consulting
P.O.Box 4, 01621, Finland
Petri.Vasara@poyry.fi

Abstract The objective of this work was to develop automatic tools for post-processing of SOMs, especially in the context of hierarchical data — data where each higher level object consists of a varying number of lower level objects. Both low and high level data is available and needs to be utilized. The information from lower levels is transferred to higher level using data histograms of lower level clusters. The clusters are formed and interpreted automatically so as to summarize the information given by the SOM, and to produce meaningful indicators that are useful also to problem domain experts. The results show that the approach works well at least in the case study of pulp and paper mills technology data.

Keywords: interpretation, self-organizing map, clustering, data mining

1 Introduction

The self-organizing map (SOM) [6] is frequently being utilized in data analysis, including analysis of hierarchical data [13]. An example of such data is forest industry data used as a case study in this project [10][1]. The data consists of paper mills where each mill contains one or more pulp and paper machines. On a higher level, pulp and paper companies own varying number of mills. An earlier study concerned a very similar data set [13]. In that study, at least, one of the major workphases was interpretation of the resulting SOM: the data clearly consisted of a number of clusters, but quantifying their significant properties required a great deal of manual work.

In this paper, an approach is proposed which automates much of the whole analysis process. An automatic clustering procedure is utilized, and the resulting clusters are analysed with post-processing methods which facilitate interpretation. In hierarchical context, data histograms based on the lower level clusters are used as features on the upper level. Because each cluster has an interpretation, the new features are meaningful also to problem domain expers.

[1] This work has been carried out in the "Enterity" project in Helsinki University of Technology. The financing of TEKES and Jaakko Pöyry Consulting is gratefully acknowledged.

2 Methodology

The data analysis approach described in this section consists of four phases. The first phase is training a SOM based on the data. This involves several choices, such as normalization method and training parameters. However, these are not discussed in this paper. The second phase is clustering of the SOM (Section 2.1), and the third phase is interpretation of the clusters (Section 2.2). If the data is hierarchical and the lower level information needs to be transferred to a higher level, an additional fourth phase is needed to construct the frequency components for each higher level object (Section 2.3).

2.1 Clustering

There are several ways to perform clustering of the SOM [12]. Frequently it has been based on visual cues and manual assignment of map units to clusters. In order to automate the analysis process this approach is not applicable. In this paper, the clustering procedure is based on k-means algorithm and Davies-Bouldin index [4,2,12]. Several different partitions of the SOM prototypes are made with k-means algorithm using a range of k-values. The best of these is the one that minimizes Davies-Bouldin index: $I_{DB} = \frac{1}{C}\sum_{i=1}^{C} \max_j(\frac{S_i+S_j}{M_{ij}})$, where C is the number of clusters, S_i is the dispersion (e.g. mean squared distance from center) of cluster i, and M_{ij} is the distance between centers of clusters i and j. This clustering procedure tries to find spherical clusters which are internally compact but widely separated.

Note that the clustering method could be applied directly to the data instead of the SOM. The SOM is used as an intermediate phase because this reduces the computational complexity of clustering, and because the trained SOM can be efficiently used for visualization, both important properties in practice [12].

2.2 Interpretation of clusters

Methods which aid in interpretation of SOMs is a topic which has been gaining interest recently with the use of the SOM for data mining [3]. The aim of interpretation is to create an understanding of which components are important with respect to each cluster; both within the cluster and with respect to other clusters.

A recently favoured approach is to derive some measure of significance and rank the components using it [7,9,5]. The data in the case study consisted mostly of sparse variables where a big value is much more significant than a small value. The measures take this into account by comparing how big the relative values of the components are in each cluster:

$$s_v(i,k) = \frac{m_{ik} - \min_k}{\max_k - \min_k}, \quad (1)$$

where m_{ik} is the mean value for component k in cluster i, and $\min_k$, $\max_k$ are minimum and maximum values of component k. This measures the relative importance of the components within each cluster, but it does not take other clusters into account at all. For this purpose, a second measure can be used:

$$s_{\hat{v}}(i,k) = \frac{s_v(i,k)}{\frac{1}{C-1}\sum_{j \neq i} s_v(j,k)}, \tag{2}$$

which measures how big value component k has with respect to its value in other clusters. Both s_v and $s_{\hat{v}}$ are maximized to find the most significant component. The measures are very closely related to those defined in [7]: their product $s_v s_{\hat{v}}$ corresponds to the G^2 measure for keyword selection for areas in document maps.

While definitions of s_v and $s_{\hat{v}}$ can be easily changed to accomodate the situation where small values are significant, they fail in more general cases. A simple refinement is to use standard deviation σ_k instead of the component value. This approach is closely related to the LabelSOM method [9].

Earlier, the characterization of SOM clusters has been done using rules [8,11]. Also this approach was applied: for each cluster, a set of rules of type r `is true, if` $x_{lk} > \alpha_k$ was generated. The rules were evaluated using the product of two measures: $P(r|i) = n_{r\&i}/n_i$ which measures the confidence of rule r being true in cluster i (internal significance), and $P(i|r) = n_{r\&i}/n_r$ which measures confidence in a data sample being in the cluster i if rule r is true (external significance). Their product measures the significance of the rule:

$$S_r(i,r) = P(i|r)P(r|i) = n_{r\&i}^2/(n_r n_i), \tag{3}$$

where n_r is the number of samples for which the rule r is true, n_i is the number of samples in cluster i, and $n_{r\&i}$ is the number of samples in cluster i for which the rule r is true. The measure $S_r(i,r)$ reaches its maximum value 1 if the rule and cluster correspond to each other perfectly.

2.3 Frequency components

After lower level maps have been clustered and interpreted, the information can be transferred to an upper level. This is accomplished through the use of frequency components, each of which corresponds to one cluster on the lower level map.

To each lower level vector $\mathbf{x}_l$, a vector $\mathbf{p}_l$ is associated which indicates which cluster $\mathbf{x}_l$ belongs to. At most rudimentary level, this can be a binary vector $\mathbf{p}_l = [p_{l1}, ..., p_{lC}]$, with $p_{li} = 1$ if the best-matching unit of $\mathbf{x}_l$ belongs to cluster i and zero otherwise. In this work, a more sophisticated version was used based on a gaussian mixture model estimated on top of the SOM [1]. The probabilities of each data vector to belong to each map unit were calculated based on the mixture model, and these probabilities were averaged over each cluster, giving a non-binary vector $\mathbf{p}_l$ for each low level data vector.

For each higher level object (such as a forest industry company) a set of new features was constructed by aggregating the $\mathbf{p}_l$ vectors of all lower level objects (such as pulp and paper mills) associated with it. In the case study, the aggregation was done as an average of the vectors.

3 Case study

A database of pulp and paper mills was investigated [10]. The database contained data about various technical aspects of pulp and paper mills around the world. Analysis presented here contains two levels, low level map of the mills ($n = 4205$) and a higher level map of the forest industry companies ($n = 279$).

The mill data consisted of 47 variables most of which were production capacities of various pulp and paper types. All variables were scaled to have unit ([0,1]) range. A map with 22 x 15 units was trained (using default training parameters as defined in SOM Toolbox[2]). The map was then clustered using k-means (see previous section) resulting in 15 clusters, and analysed using s_v as the measure of significance.

Figure 1 shows U-matrix of the map, the 15 automatically determined clusters, and the most signicant component (and its value) in each cluster. In addition, some rules were constructed for each of the clusters and tested using the data belonging to that cluster, see Table 1.

Each of the 15 clusters formed one feature used in the higher level company map. The feature vectors for each company were formed with the procedure described in the previous section, such that each feature indicated the relative frequency of mills of that particular type in the company. A 16th component was added with the total number of mills in each company. No scaling was performed on the frequency components, but the 16th component was scaled between [0,1]. The same training and interpretation procedure as for the lower level map was applied, except for the fact that $s_{\hat{v}}$ was used as the significancy measure instead of s_v.

The clusters, most significant components and one component plane ("Integrated news") of the company level map are shown in Figure 2. The clusters are not quite as clear as in the mill map, but they still give a coherent picture of the forest industry companies.

4 Conclusions

In this paper, a framework has been presented for interpretation of the cluster structure and contents of SOMs, and for generation of new meaningful components for higher level data in hierarchical data sets. In the case study, the methodology worked well, and the domain experts were very satisfied with the results. It was not investigated whether the achieved clusters and interpretations were optimal in some sense, but they gave the experts a sensible and coherent picture of the data.

On the company level map, the cluster based components offered a vast improvement in interpretability over the projection (ie. SOM unit-)coordinates utilized in earlier work [13]. Apart from selecting the significancy measures, giving actual names to the frequency components is the only part of the procedure

[2] A SOM library for Matlab: `http://www.cis.hut.fi/projects/somtoolbox/`

which makes it less than totally automatic. The list of most significant components gives a starting point for the naming, but the data analyst must apply his or her own insight to crystallize the automatically produced information.

References

1. Esa Alhoniemi, Johan Himberg, and Juha Vesanto. Probabilistic Measures for Responses of Self-Organizing Map Units. In H. Bothe, E. Oja, E. Massad, and C. Haefke, editors, *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA'99)*, pages 286–290. ICSC Academic Press, 1999.
2. David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979.
3. Guido Deboeck and Teuvo Kohonen, editors. *Visual explorations in Finance using Self-Organizing Maps*. Springer-Verlag, London, 1998.
4. Robert M. Gray. Vector quantization. *IEEE ASSP Magazine*, pages 4–29, April 1984.
5. Samuel Kaski, Janne Nikkilä, and Teuvo Kohonen. Methods for interpreting a self-organized map in data analysis. In Michel Verleysen, editor, *Proceedings of ESANN'98, 6th European Symposium on Artificial Neural Networks, Bruges, April 22-24*, pages 185–190. D-Facto, Brussels, Belgium, 1998.
6. Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995.
7. Krista Lagus and Samuel Kaski. Keyword selection method for characterizing text document maps. In *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, volume 1, pages 371–376. IEE, London, 1999.
8. W. Pedrycz and H. C. Card. Linguistic interpretation of self-organizing maps. In *Proceedings of International Conference on Fuzzy Systems '92*, pages 371 – 378, 1992.
9. Andreas Rauber and Dieter Merkl. Automatic labeling of self-organizing maps: Making a treasure-map reveal its secrets. In *Proceedings of the 3rd Pasific-Area Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, 1999.
10. Markus Siponen. Automaattisia jälkitulkintamenetelmiä hierarkisen tietoaineiston tutkimiseen itseorganisoivan kartan avulla. Master's thesis, Helsinki University of Technology, 2000. In Finnish.
11. A. Ultsch. Knowledge extraction from self-organizing neural networks. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 301–306. Springer Verlag, 1993.
12. Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.
13. Juha Vesanto, Petri Vasara, Riina-Riitta Helminen, and Olli Simula. Integrating environmental, technological and financial data in forest industry analysis. In Bert Kappen and Stan Gielen, editors, *Proceedings of 1997 Stichting Neurale Netwerke Conference*, pages 153–156, Amsterdam, the Netherlands, May 1997. World Scientific.

Table 1. Rules generated for cluster 1 ($n_i = 101$) of the mill map. The first 4 rows are conditions based on the 4 most significant components of the cluster, and the 3 last rows are their combinations such that all listed conditions must be true. The rules worked suprisingly well. The maximum significance was often achieved with 2-3 conditions and it was between 0.5–0.7. In this case, the most signicant rule contained 3 components and the significance was 0.91 (bolded).

Rule	$n_{r\&i}$	n_r	$P(i\|r)$	$P(r\|i)$	$S_r(i,r)$
(1) Price_vol $>$ 0.75	101	274	0.37	1.00	0.37
(2) Tot_chem $>$ 0.74	93	400	0.23	0.92	0.21
(3) Tot_sa $>$ 0.67	93	372	0.25	0.92	0.23
(4) Bl_sa $>$ 0.64	88	292	0.30	0.87	0.26
(1,2)	93	101	0.92	0.92	0.85
(1,2,3)	93	94	0.99	0.92	**0.91**
(1,2,3,4)	81	82	0.99	0.80	0.79

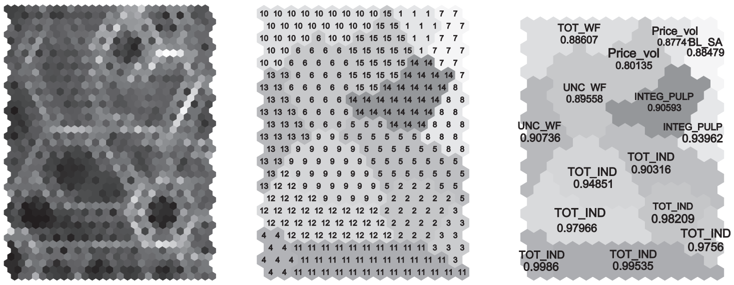

Figure 1. The U-matrix (on left, white denotes large distance from neighbors), automatically produced clusters (middle) and most significant component according to s_v (on right). On the whole lower half of the map, the most significant component is the industrial paper production capacity. For them (and other similar cases) differences between clusters only become apparent after investigation of the other significant components.

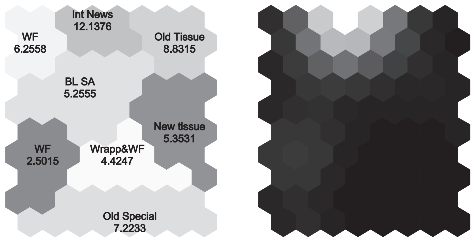

Figure 2. The clusters and most significant components (left) and the “Integrated News” component plane (right, white denotes high value), which clearly corresponds to the cluster on top middle of the SOM.