

Chapter 14

SOM in data mining

Esa Alhoniemi, Johan Himberg, Jaakko Hollmén, Sampsa Laine, Pasi Lehtimäki, Kimmo Raivio, Timo Similä, Olli Simula, Miki Sirola, Mika Sulkava, Jarkko Tikka, Juha Vesanto

14.1 Introduction

The Self-Organizing Map (SOM) has proven to be one of the most powerful algorithms in data visualization and exploration. Application areas include various fields of science and technology, e.g., complex industrial processes, telecommunications systems, document and image databases, and even financial applications. The SOM maps the high-dimensional input vectors onto a two-dimensional grid of prototype vectors and orders them. For a human interpreter, the ordered prototype vectors are easier to visualize and explore than the original data. The SOM has been widely implemented in various software tools and libraries, for example, the SOM Toolbox [1].

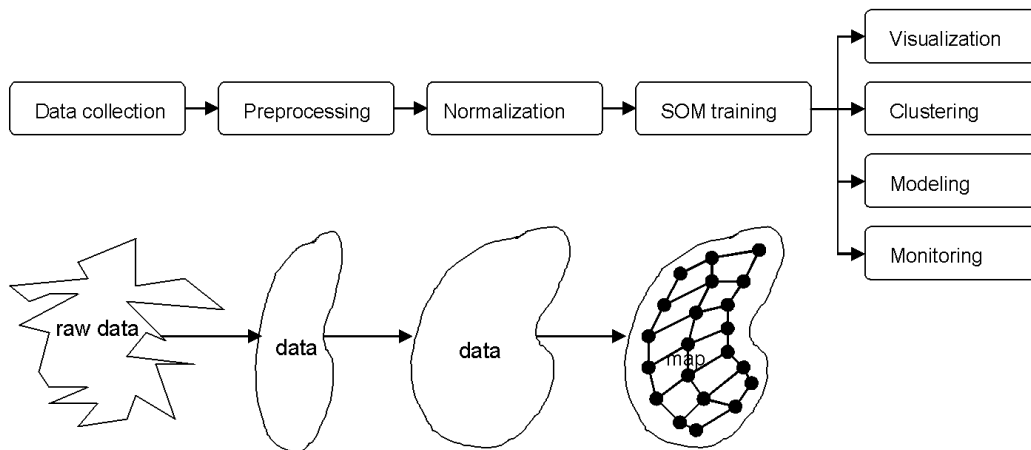


Figure 14.1: Applying the SOM in data mining. Post-processing the SOM extracts qualitative or quantitative information of the data. Visualization and clustering provide qualitative information, while modeling and monitoring give quantitative information resulting in deeper understanding of the system behavior.

The research work has been motivated by a number of practical data mining projects where SOM has been a central data analysis tool [2]. It has become apparent that while the SOM can be used to quickly create a qualitative overview of the data, turning this qualitative information to quantitative characterizations requires a great deal of expertise and manual work. There is no wide consensus or understanding of the methods needed for post-processing of the SOM-based data analysis (see Figure 14.1). The subsequent research has concentrated on devising such methods and on gaining a better understanding of the possibilities, strengths, and weaknesses of the SOM in data exploration.

References

- [1] Alhoniemi E., Himberg J. Parhankangas J., Vesanto J., The SOM-toolbox, 2000. Available from <http://www.cis.hut.fi/projects/somtoolbox/>.
- [2] Kohonen, T., Self-Organizing Maps, Series in Information Sciences, second edn. 1997, Springer, Heidelberg.

14.2 Clustering of the SOM

Clustering of data is one of the main applications of the Self-Organizing Map (SOM) [1]. U-matrix is one of the most commonly used methods to cluster the SOM visually. However, in order to be really useful, clustering needs to be an automated process. When clusters are identified visually the results may be different when performed by different people. There are several techniques which can be used to cluster the SOM autonomously, but the results they provide do not follow the results of U-matrix very well.

In [2], a clustering approach based on distance matrices was introduced which produces results which are very similar to the U-matrix. It was compared to other SOM-based clustering approaches and found to produce more reliable results. The automated clustering algorithm has also been applied to study the relations of nutrient concentrations in tree needles [3] (see also Section 15.1).

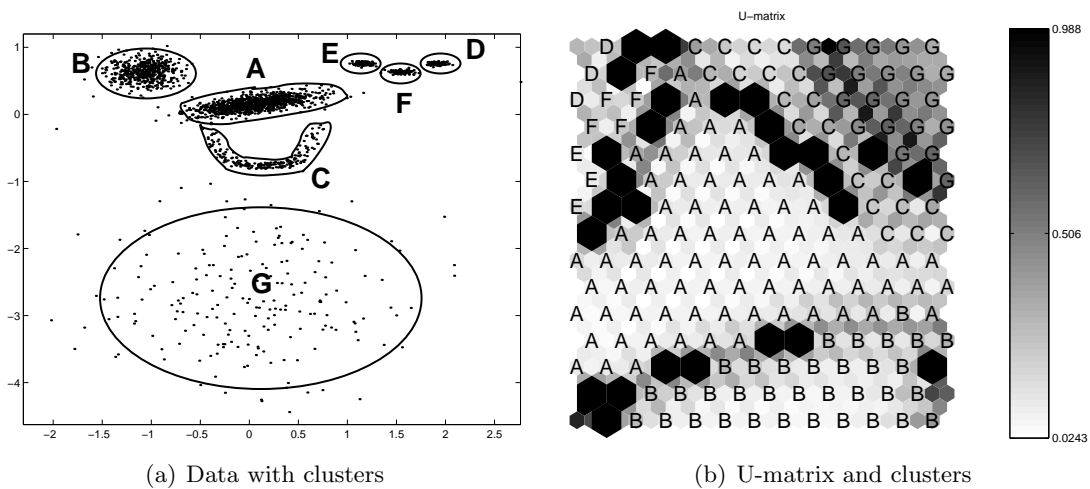


Figure 14.2: (a) Data set with true clusters indicated with encircled areas and the letters. (b) U-matrix of the data with empty map units shown as black, and clustering result with the letters.

References

- [1] Juha Vesanto and Esa Alhoniemi. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11(2):586–600, March 2000.
- [2] Juha Vesanto and Mika Sulkava. Distance matrix based clustering of the self-organizing map. In José R. Dorronsoro, editor, *Artificial Neural Networks - ICANN 2002*, volume 2415 of *Lecture Notes in Computer Science*, pages 951–956, Madrid, Spain, August 2002. Springer.
- [3] Mika Sulkava and Jaakko Hollmén. Finding profiles of forest nutrition by clustering of the self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 243–248, Hibikino, Kitakyushu, Japan, September 2003.

14.3 Use of operator maps to analyze mobile access network

One of the most flexible extensions of the SOM suitable for the analysis of switching time-series data is the operator map [1]. In operator maps, the map units (or operators) are generalized to be parametric models that are able to describe the interesting behavioral patterns in different parts of the data.

In [2, 3], operator maps were applied in the analysis of downlink traffic performance in a macrocellular network scenario. In the analysis, the relationships between the number of users, the downlink average transmission power and the downlink frame error rate were studied. Several operator maps consisting of 16 local operators in a rectangular lattice of size $[4 \times 4]$ were trained using different types of map operators. In Figure 14.3, an example of an operator map with 16 neuro-fuzzy operators is shown. Each neuro-fuzzy operator provides a linguistic description for the input variable condition in which the quality problems occur, enabling easy human analysis of the dependencies in the data.

1 if nUsr(n) is very high nUsr(n-1) is very high dlTxp(n) is very med dlTxp(n-1) is very med then dlFer is 0.70	5 if nUsr(n) is very med nUsr(n-1) is very high dlTxp(n) is very high dlTxp(n-1) is very med then dlFer is 0.72	9 if nUsr(n) is very high nUsr(n-1) is very high dlTxp(n) is very low dlTxp(n-1) is very med then dlFer is 0.74	13 if nUsr(n) is very high nUsr(n-1) is very high dlTxp(n) is very high dlTxp(n-1) is very med then dlFer is 0.73
2 if nUsr(n) is very med nUsr(n-1) is very high dlTxp(n) is very med dlTxp(n-1) is very med then dlFer is 0.70	6 if nUsr(n) is very low nUsr(n-1) is very high dlTxp(n) is very high dlTxp(n-1) is very med then dlFer is 0.76	10 if nUsr(n) is very high nUsr(n-1) is very ² high dlTxp(n) is very high dlTxp(n-1) is very med then dlFer is 0.72	14 if nUsr(n) is very med nUsr(n-1) is very high dlTxp(n) is very ² low dlTxp(n-1) is very high then dlFer is 0.73
3 if nUsr(n) is very high nUsr(n-1) is very ² high dlTxp(n) is very med dlTxp(n-1) is very med then dlFer is 0.72	7 if nUsr(n) is very low nUsr(n-1) is very High dlTxp(n) is very med dlTxp(n-1) is very High then dlFer is 0.70	11 if nUsr(n) is very low nUsr(n-1) is very high dlTxp(n) is very med dlTxp(n-1) is very High then dlFer is 0.74	15 if nUsr(n) is very high nUsr(n-1) is very med dlTxp(n) is very high dlTxp(n-1) is very low then dlFer is 0.77
4 if nUsr(n) is very med nUsr(n-1) is very med dlTxp(n) is very med dlTxp(n-1) is very ² high then dlFer is 0.72	8 if nUsr(n) is very med nUsr(n-1) is very high dlTxp(n) is very high dlTxp(n-1) is very med then dlFer is 0.71	12 if nUsr(n) is very high nUsr(n-1) is very low dlTxp(n) is very med dlTxp(n-1) is very med then dlFer is 0.72	16 if nUsr(n) is very med nUsr(n-1) is very high dlTxp(n) is very med dlTxp(n-1) is very med then dlFer is 0.73

Figure 14.3: Rule-based descriptions of the operators.

References

- [1] Teuvo Kohonen. *Self-Organizing Maps*, 3rd edition. Springer, 2001.
- [2] Pasi Lehtimäki. *Self-Organizing Operator Maps in Complex System Analysis*. Master's Thesis, Department of Computer Science and Engineering, Helsinki University of Technology, 2002.
- [3] Pasi Lehtimäki, Kimmo Raivio and Olli Simula. *Self-Organizing Operator Maps in Complex System Analysis*. In *Proceedings of Joint International Conference on Artificial Neural Networks and Neural Information Processing*, pages 622–629, Istanbul, Turkey, June 26 - 29 2003.

14.4 Use of LogSig-scaling to incorporate expert knowledge to SOM-based visualization of GSM-network data

Normalization of data is an important step of data analysis. Since most analysis methods measure distances between data points, a variable having higher variation will dominate the results. A common way to perform normalization is done by subtracting mean and scaling to unit variance each of the variables. Outliers, or equivalently uninteresting parts of the data distribution, reduce weight of interesting parts of the distribution when such normalization is performed. This causes analysis methods to concentrate on wrong issues.

Data collected from operation of GSM-network is studied in order to compare effects of two different normalization methods on information content of SOM trained with normalized data [1]. Process experts of the GSM-network provided value ranges of importance for each of the variables. The proposed normalization method transforms the data by sigmoidal function whose shape is fixed based on auxiliary information from the experts. By normalizing the data with the proposed method, the SOM visualizes better overall behavior of the GSM-network, whereas the reference method performing unit variance normalization causes SOM neurons to stretch toward extreme parts of the data distribution. These extreme parts represent severe, but rare problems in network operation. Sammon mappings [2] in Figure 14.4 visualize relative positions of neurons in the two multidimensional SOMs.

References

- [1] K. Hätonen, S. Laine, T. Similä. Using the LogSig-function to integrate expert knowledge to Self-Organizing Map (SOM) based analysis. In *Proceedings of the 2003 IEEE International Workshop on Soft Computing in Industrial Applications*, pages 145-150, Binghamton, NY, USA, June 23-25, 2003.
- [2] J. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers* C-18(5):401-409, 1969.

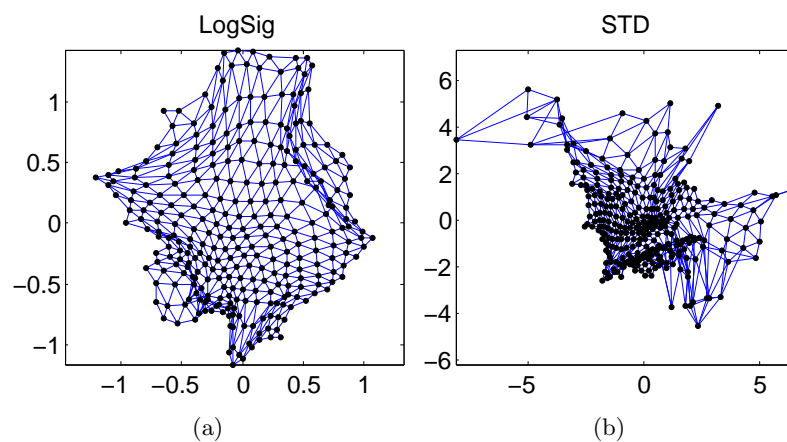


Figure 14.4: Sammon mapping of SOM trained with (a) sigmoidal normalized and (b) unit variance normalized data.

14.5 Analysis of mobile access network

Both 3G and GSM networks produce a huge amount of data. In this project, the Self-Organizing Map has been used to analyse mobile data [1][2]. 3G network data has been generated using a wideband code division multiple access (WCDMA) radio network simulator. The GSM data has been collected from real network. The goal is to develop efficient adaptive methods for monitoring the network behavior and performance. Special interest is on fault detection and on finding clusters of mobile cells. Cells of one cluster can be configured using similar parameters.

The method utilizes the SOM algorithm twice when clustering mobile cells. The Self-Organizing Map is used together with some clustering algorithm to cluster data vectors of single mobile cell and to cluster the mobile cell features. This two phase clustering algorithm [3] begins with training a SOM with the data vectors. The codebook vectors of the SOM are clustered using K-means or some hierarchical clustering method with a validity index so that exact number of clusters can be defined. The input data vectors are classified using labeled SOM codebook vectors.

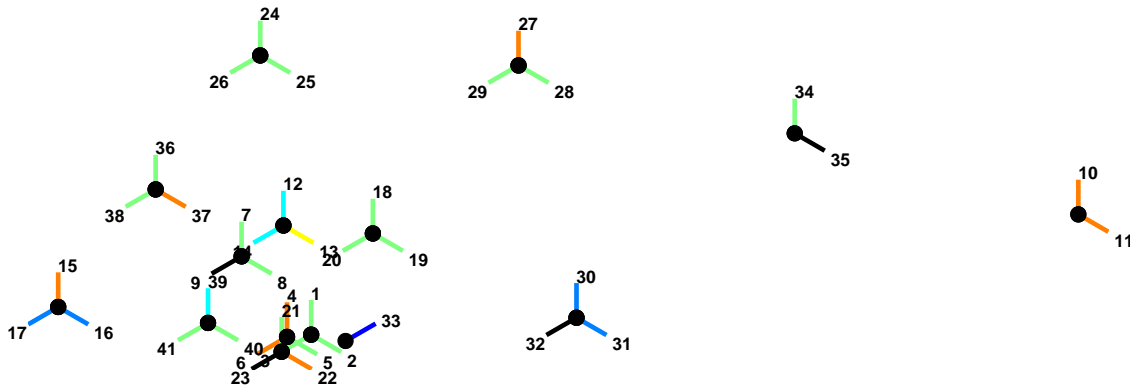


Figure 14.5: Classified GSM cells

For each mobile cell a histogram is computed. The histogram describes how the data from one cell fall into the data clusters. These histograms are used as profiles in cell classification. The profiles are fed into second SOM, which is clustered to find the classes of cell profiles. The classified mobile cells and their locations are presented in Figure 14.5.

In this method, two level clustering procedure has been used because long term cell profiles are desired. At least, in 3G systems this is necessary due to high sampling rates, but also in GSM systems they give us more reliable classification results.

References

- [1] Pasi Lehtimäki, Kimmo Raivio, and Olli Simula. Mobile radio access network monitoring using the self-organizing map. In *Proceedings of European Symposium on Artificial Neural Networks*, pages 231–236, Bruges, Belgium, April 24 - 26 2002.
- [2] Kimmo Raivio, Olli Simula, Jaana Laiho, and Pasi Lehtimäki. Analysis of mobile radio access network using the self-organizing map. In *Proceedings of the International Symposium on Integrated Network Management*, pages 439–451, Colorado Springs, Colorado, USA, March 24 - 28 2003.
- [3] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

14.6 Impact of R&D on growth quantiles in manufacturing firms

Impact of research and development (R&D) on growth in Finnish manufacturing firms is studied. Growth of a firm is treated as random variable whose distribution is conditional on regressors, namely size of the firm, R&D intensity of the firm, R&D intensity of the industry of the firm and growth of the industry of the firm. Local linear quantile regression model is built in order to capture variation in firm growth given values of the regressors. Since the model is nonparametric, its parameters vary both by quantiles and values of the four regressors. SOM is used in visualization of the quantile regression model. Novelty of this choice is capability to track shapes of the conditional quantiles of firm growth distribution and perform sensitivity analysis for them as function of the regressors. Figure 14.6 shows an illustrative example of SOM in quantile regression visualization.

Results of the study suggest that there exists a relationship between the conditional quantiles of firm growth and the regressors. Smallest R&D doing firms gain the highest, but also the lowest benefits from firm-level R&D investments independent of other factors. Sensitivity analysis show that only already growing firms gain benefit from increase in firm-level R&D investments in medium and high technology industries, but in low technology industries even non-growing firms may gain benefit. Firms in low technology industries benefit more from increase in industry-level R&D intensity than firms in medium or high technology industries. More detailed results will be published later in a Master's thesis.

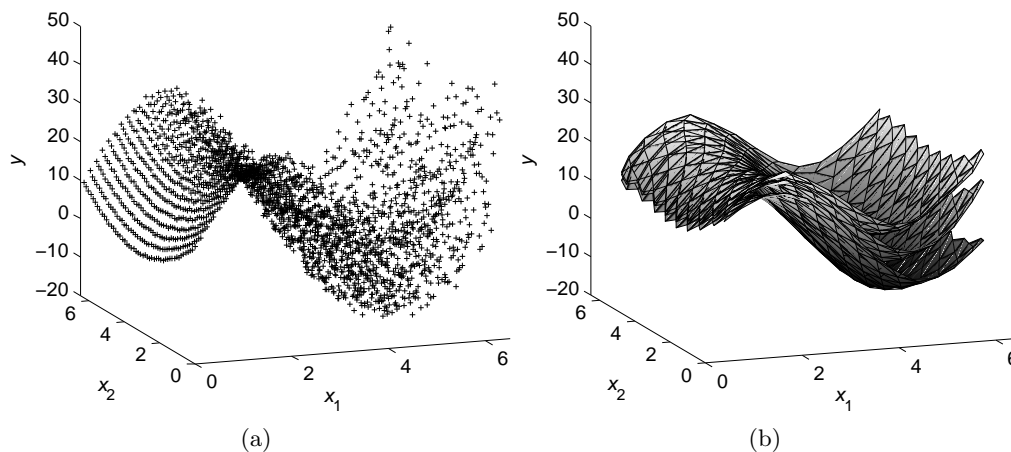


Figure 14.6: (a) Scatter plot of simulated data and (b) three-layered SOM arranged along quantiles 15%, 50% and 85% of random variable $y|\{x_1, x_2\}$.

