Chapter 15

Methods for efficient utilization of SOM in data mining

Olli Simula, Jaakko Hollmén, Esa Alhoniemi, Juha Vesanto, Johan Himberg, Markus Siponen, Jussi Ahola, Juha Parhankangas, Mika Sulkava

The Self-Organizing Map (SOM) algorithm has been widely implemented in various software tools and libraries, for example, the SOM Toolbox [5]. However, for a common practitioner a difficulty with applying the SOM in data mining has been that there is no wide consensus or understanding of the methods needed for post-processing the SOM. Thus, methods for the two most important application areas of the SOM — visualization and cluster analysis — have been investigated and developed.

The goal of this work has been to enhance the data exploration process based on the SOM, see Figure 15.1. The 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. 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.

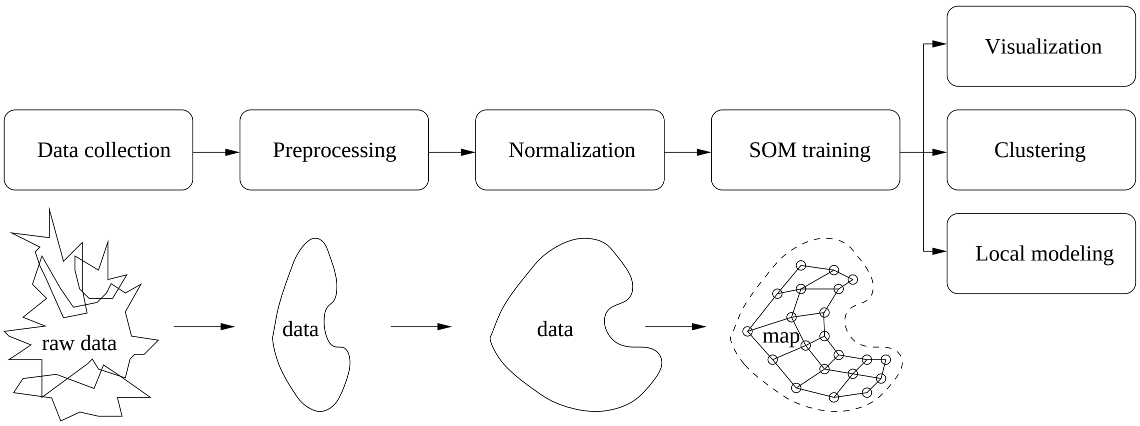

Figure 15.1: Applying the SOM in data mining. After data collection, the data is preprocessed, normalized, and a SOM is trained. In this work, we have concentrated on visualization and clustering in the post-processing of the SOM.

In [4], clustering of data using SOM is investigated. This is essentially a two-phase approach which reduces the computational load of clustering considerably, making clustering of large data sets feasible. Also, a SOM-based cluster visualization and its application to false coloring is described in a closely related paper [1].

Methods for interpretation of SOM clusters and a framework for (semi)automated analysis of hierarchical data are presented in [3]. Clusters are derived automatically, and then characterized by ranking the variables, and by constructing characterizing rules for the variables. The rules are formed by maximizing a novel measure of significance. In case of hierarchical data, the clusters form new variables for the upper level data, and the characterizations allow one to give them meaningful names.

An outline of a system for automatically generating data survey reports of table-format numerical data is described in [6]. Algorithms for constructing a cluster hierarchy and describing the clusters are proposed. Different methods and representations are combined to produce a unified and comprehensive report of the properties of the data manifold.

References

[1] J. Himberg. A SOM Based Cluster Visualization and its Application for False Coloring. In *Proceedings of International Joint Conference on Neural Networks (IJCNN2000)*, volume 3, pages 587–592, 2000.

[2] J. Himberg, J. Ahola, E. Alhoniemi, J. Vesanto, and O. Simula. Pattern recognition in soft computing paradigm. Volume 2 of FLSI soft computing series, ed. Nikhil R. Pal, chapter The Self-Organizing Map as a Tool in Knowledge Engineering. World Scientific, 2001.

[3] M. Siponen, J. Vesanto, O. Simula, and P. Vasara. An approach to automated interpretation of SOM. In N. Allinson, H. Yin, L. Allinson, J. Slack, editors, *Proceedings of Workshop on Self-Organizing Map 2001*, pages 89–94. Springer, June 2001.

[4] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

[5] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. SOM Toolbox for Matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, April 2000.

[6] J. Vesanto and J. Hollmén. An Automated Report Generation Tool for the Data Understanding Phase. In A. Abraham, M. Koeppen, editors, *Hybrid Intelligent Systems*, Advances in Soft Computing, Heidelberg, 2002. Physica Verlag. In print.