51 Self-Organizing Map for Data Mining in MATLAB: the SOM Toolbox

Juha Vesanto, Esa Alhoniemi, Johan Himberg,
Kimmo Kiviluoto, and Jukka Parviainen

The SOM Toolbox (`http://www.cis.hut.fi/projects/somtoolbox`) is a free function library for MATLAB 5 implementing the Self-Organizing Map (SOM) algorithm which is a neural network algorithm based on unsupervised learning [1]. Basically it performs a vector quantization and simultaneously organizes the quantized vectors on a regular low-dimensional grid. The SOM has proven to be a valuable tool in data mining because it is readily explainable, simple and easy to visualize. It has been successfully applied in various engineering applications in pattern recognition, image analysis, process monitoring and fault diagnosis [2, 3].

Thus far, the most useful implementation of the SOM and related tools has been the SOM_PAK (`http://www.cis.hut.fi/nnrc/nnrc-programs.html`). It is a public domain software package developed in the Neural Networks Research Centre of the Helsinki University of Technology, written in C language for UNIX and PC environments. However, the Mathwork Inc.'s MATLAB has been steadily gaining popularity as the "language of scientific computing". Moreover, MATLAB is much better-suited for fast prototyping and customizing than the C language used in SOM_PAK, as MATLAB employs a high-level programming language with strong support for graphics and visualization. All of these properties are extremely important in data mining. SOM Toolbox is an attempt to take full advantage of these strengths and provide an efficient, customizable and easy-to-use implementation of the SOM.

While closely related to SOM_PAK, SOM Toolbox is, however, a new set of programs. Both program packages have their relative strengths and weaknesses. The advantages of SOM_PAK are that it is written in ANSI C and thus runs in virtually any environment. It is an order of magnitude faster than SOM Toolbox in training. The advantages of SOM Toolbox are mainly in user friendliness and visualization capabilities. If desired, the SOM_PAK files can be accessed with the Toolbox: it is possible to first train the SOM with the SOM_PAK and then use the Toolbox for visualization.

SOM Toolbox utilizes MATLAB structures and the functions are constructed in a modular manner, which makes it convenient to tailor the code for each users' specific needs. The use of structs allows the Toolbox to keep track of many kinds of information that greatly facilitate the data mining process: labels associated with individual data vectors, variable names, data normalization information and training log.

The basic usage of the SOM Toolbox consists of three steps: SOM initialization, training and visualization. To make things easier to the user, the high-level functions require minimum number of parameters. For example, SOM size and training parameters are, unless specified, determined automatically based on the training data.

```
» sM=som_init(data); %initialization
» sM=som_train(sM,data); %training
» som_show(sM); %visualization, see Figure 108
```

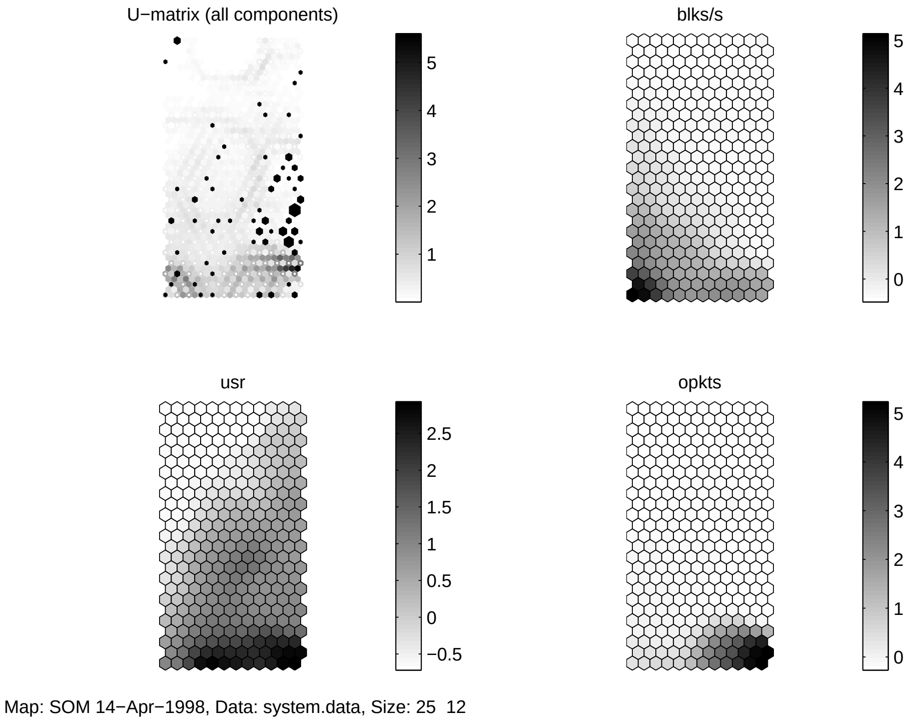

Figure 108: U-matrix and three components planes visualized by the SOM Toolbox. Hits from a small data set has been added on top of the U-matrix.

All this can also done through a graphical user interface. Around these three basic steps, SOM Toolbox has a large number of functions that can be used for preprocessing of the data and post-processing/analyzing the SOM.

We have found that the SOM Toolbox has greatly facilitated our research work. Implementation in MATLAB allows fast prototyping and powerful visualization. Building application specific tools on top of the Toolbox has proven to be easy.

Currently we are working on version 2 of the Toolbox. The major differences to the old version will be in visualization, which will utilize the newest research results in the field [4]. In addition, the package will include a larger set of supplementary algorithms and tools. Version 2 should be available during 1999.

References

[1] T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, Heidelberg, 1995.

[2] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas. Engineering applications of the self-organizing map. *Proceedings of the IEEE*, 84(10):27 pages, October 1996.

[3] O. Simula and J. Kangas. *Neural Networks for Chemical Engineers*, volume 6 of *Computer-Aided Chemical Engineering*, chapter 14, Process monitoring and visualization using self-organizing maps. Elsevier, Amsterdam, 1995.

[4] J. Vesanto. SOM-Based Data Visualization Methods. *Intelligent Data Analysis*, April 1999 (to appear).