Enhancing SOM-based data visualization by linking different data projections

Johan Himberg
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 2200, FIN-02015 HUT, Finland
email: Johan.Himberg@hut.fi

April 30, 1998

Abstract

The self-organizing map (SOM) is widely used as a data visualization method especially in various engineering applications. It performs a non-linear mapping from a high-dimensional data space to a lower dimensional visualization space. The SOM can be used for example in correlation detection and cluster visualization in explorative manner. In this paper two tools for refing the SOM-based visualization are presented. The first one brings out a sharper view to the correlation detection and the second one brings additional information to the input space distance visualization. Both tools are based on linking two different data projections using color coding. The tools are demonstrated using a real world data example from a queuing system.

1 Introduction

The self-organizing map (SOM) [3] is a neural network algorithm that has been widely used in various engineering applications [4]. The SOM is based on unsupervised learning and it performs a topology preserving mapping from the high-dimensional data space to map units. The map units are usually organized into a two-dimensional (2D), regular lattice. The mapping is nonlinear and it retains the topology of the high-dimensional data space in such a way that data points lying near each other in the input space are mapped to nearby locations on the map. The SOM may be used as a tool in investigating and visualizing complex dependencies between different process parameters and in visualizing the process state evolution [6, 7].

In this paper two enhancements for the SOM-based visualization are presented. The first one (Section 3) refines the SOM-based visualization of dependencies between parameters. The second one (Sec. 4) shows how to combine the SOM with some other data visualization method in order to get "a second opinion" on the relations in the data space to the same visualization. Both enhancements are based on linking different types of projections by a color coding method [2] (Sec. 2).

The motivation for these tools has emerged from the needs of process monitoring using the SOM. Thus, the examples and conclusions in this paper are made having the process monitoring task in mind. However, the enhancements proposed are not limited to this kind of data, but may be used in explorative data analysis in general.

2 Linking two 2D-projections by color coding

In this paper the concept "color linking" is often used and it will be described here briefly. The idea of color linking is adopted from Kaski et al. [2] who use it in SOM-based cluster structure visualization.

Consider two different 2D visual representations of the same set of objects. How to see the connection between representations, that is, how to easily identify the same object in the two representations?

A simple and visually appealing solution is to give a different color (or shape, orientation etc.) for each object. Now the connection is easily seen: the color links the two representations. In our problem the objects are high-dimensional vectors and the visual representations are their (possibly non-linear) projections in two dimensions.

In this paper two projections of the same vectors are linked in the following manner: One of the projections determines the colors; the coordinates of the projected points on the 2D plane are used directly to determine an RGB-coded color for the points. The points on the second projection are then colored according to their counterpart in the first one. The projection method can be selected among various alternatives: principal component analysis, projection pursuit, curvilinear component analysis [1], Sammon's projection [5], self-organizing map [3] etc. The point is to combine two different projections of high-dimensional data in order to get more information to the 2D visualization as the different projections bring up different relations in the data.

The selection of the color coding is essential for getting a visually clear linking between the projections. Evidently the colors for the points in the first projection should be selected so that the relation between the location and the color is visually clear. Here a straight-forward approach is used: the x-axis location gives the value of R in the RGB-coding, the y-axis location corresponds B and G is set to 1−B. The coordinates of the points are scaled to the interval $[0, 1]$. Now each point gets a color which is a function of its location on the plane. To put it simply: the points are plotted on a rectangular continuous color palette and each point picks the color under it.

Note that the idea of using a slice of the RGB-cube is purely a heuristic design; further work have to be done if a psycho-visually optimal color coding is desired.

3 Refining the SOM-based "correlation hunting"

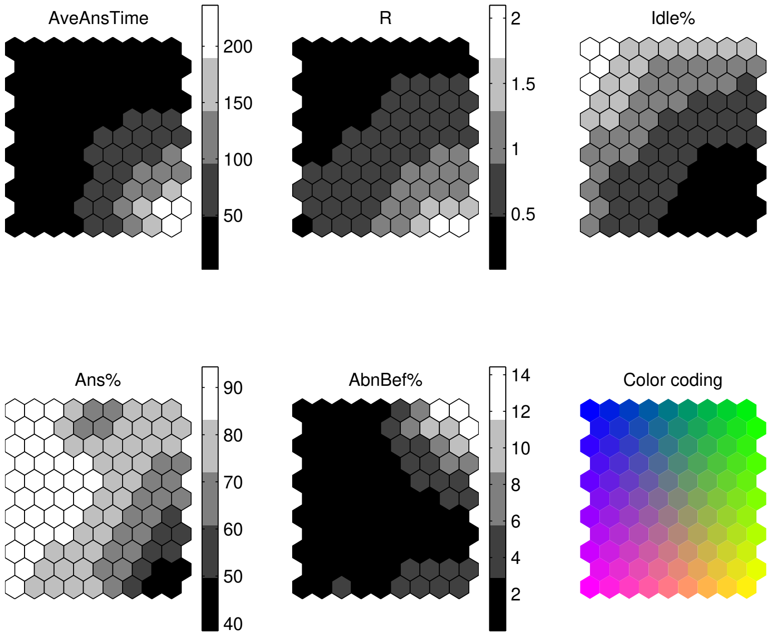
A conventional method to get an overall picture on the pairwise dependencies between two data parametres is, of course, the simple 2D scatter plot. However, the number of displays (pairwise scatter plots) grows quadratically as the number of data parameters increases. A 2D visualization of a SOM can be used to avoid this problem.

To do the comparision between parameters a two-dimensional SOM is often "sliced" to component planes [3]. Each plane represents the value of one parameter (component) in each node of the SOM using gray-level or color representation (see Fig. 1(a)). By comparing these planes correlating parameters may be detected as the correlating component planes resembles each other. Even "piecewise" (local) correlations may be found: then two parameter planes resembles each other in some regions. Now the number of displays grows linearly, and furthermore, the SOM brings an additional benefit for the visualization: it removes some amount of noise by doing a vector quantization.

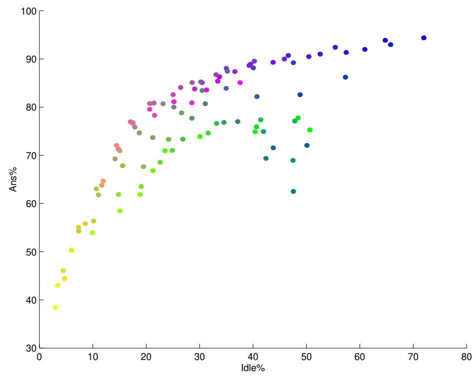
However, an accurate visual, quantitative comparision of color coded values on the planes is difficult. To solve this, we might now plot the detected interesting parameter pairs as a scatter plot. For this the values from the SOM can be used — as well as the original data. The word "correlation hunting" emphasizes that our approach is *explorative* — not quantitative correlation analysis.

Let us assume that we have a SOM which is taught with a multidimensional data consisting of vectors $\mathbf{x}$ and we detect a clear dependency between components k and l using the SOM component plane representation. We take the values x_k and x_l from every map unit vector and plot the points (x_{ki}, x_{li}). Let us denote by A,B and C some regions of the data space, and further, let us assume that we see a dependency of form $x_k = f_1(x_l)|\, \mathbf{x} \in A$, *and* some outliers $x_k = f(x_l) + \epsilon|\, \mathbf{x} \in B$ or another clear dependency $x_k = f_2(x_l)|\, \mathbf{x} \in C$ in the plot.

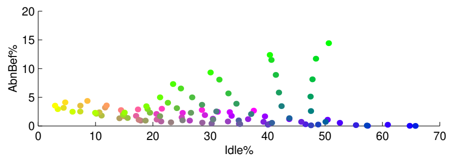
The question is now under what conditions $\mathbf{x} \in B, C$ do these anomalities occur, that is, what are the values of other parameters $x_1, \ldots, x_n$ then.



(a) Some component planes of a SOM: **AveAnsTime**: average succesfull waiting time in queue, **R**: phone call traffic intensity, **Idle%**: Ratio of idle agents, **Ans%**: ratio of calls answered, **AbnBef%**: ratio of calls which abandoned before a target time. Last subfigure is the color coding.



(b) Ratio of idle agents vs. answering percentage



(c) Ratio of idle agents vs. quickly abandoned calls

Figure 1: Figure (a) shows the component plane view of a SOM and the color coding which links the grid in (a) and the plots in (b) and (c). Figures (b) and (c) presents two pairs of parametres (vector components of the SOM in (a)) as scatter plots. The data is from a help desk queuing system. It can be seen in (b) that there is a clear dependency between the idle ratio and the answering ratio and some outliers (green points). Fig. (c) shows another dependency where it can be seen that the idle ratio vs. quickly abandoning (impatience) is differing from its normal behaviour on the same region (green points). Using the color linking even the values of other parameters can be visually investigated on the counterparts of the outliers on all the component planes.

To solve this problem we use the topology preserving property of the SOM to explore the relation between the points (x_k, x_l) on the scatter plot and the rest of the variables. The “color linking” described in Sec. 2 is applied here: the continuous color palette is associated with the SOM grid. Each unit gets a color, nearby units get similar colors. Two parameters are selected and plotted using their values in the map model vectors, that is in the map units. Each point gets the same color as the unit has on the grid. Now, an individual scatter plot may be compared with the overall system parameter behaviour as the grid coloring shows the connection between the scatter plots and the parameter plane representation as well as the connection between two different scatter plots. Obviously, this is most suitable for inspecting the anomalities in the dependencies between two parameters. An example is presented in Fig. 1.

It is evident that this approach fails, if the dependencies between variables are generally more complex, i.e. of form $x_l = f(x_1, x_2, \ldots, x_n)$. Another drawback is that the 2D SOM grid structure may cause false correlations to the plots due to map folding and interpolating units. The interpolating units have none or only few data samples on their Voronoi's region, so they can simply be left out from the plot. The folding problem is more difficult but it may be diminished using the original data in the plot instead of the model vectors — though then the generalization capability of the net is lost.

It is possible to use some other coloring for the SOM grid: one reasonable choice would be to color the grid according to its cluster structure (e.g. [2]) or to use a coloring based on a projection of the map model vectors (see next section). Another interesting approach would be an interactive tool for selecting groups of points from the scatter plots. One may then manually investigate the relations between interesting phenomena in pairwise dependencies and the rest of the data using the SOM.

4 Improving the data space distance visualization

The SOM grid visualization describes the topological relations of the data space in reference to the map grid, but it doesn't show the variable distances between adjacent units in the original data space. The unified distance matrix [8] is used to represent distances between adjacent grid units but it is bound to the grid topology as well.

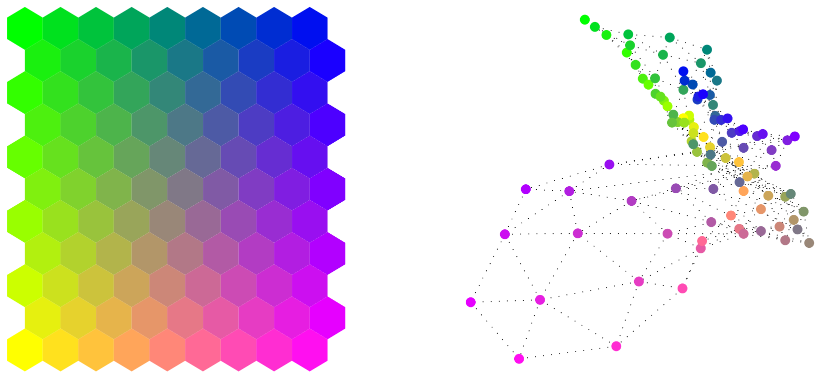
The data space distances in the SOM are often visualized using some method that projects the points in the n-dimensional data space to a 2D (or 3D) space trying to preserve the mutual distances of the projected points [3]. Sammon's projection [5] is a familiar algorithm of this type. The projection of the map describes the approximative distance between units in the input space and it may even reveal the folding of the 2D map in the high-dimensional input space. It is a practical visualization problem, though, that the connection between the projection and the grid is difficult to see, especially if the map is big and the projection folded.

Using the color linking described in Sec. 2 the topological grid and the projection of the model vectors are easily associated, see Fig. 2(a). The linking may be done to the opposite direction, too. Now, the grid coloring is done according to the location of the projected model vectors. First we project the model vectors using the selected method (i.e. principal component analysis or Sammon's projection) each point is colored according to its xy-location in the same manner as the SOM grid was colored. This color is transferred to the corresponding grid unit. Now we get a "second opinion" on the relations between map model vectors according to the projection method selected, see Fig. 2(b).

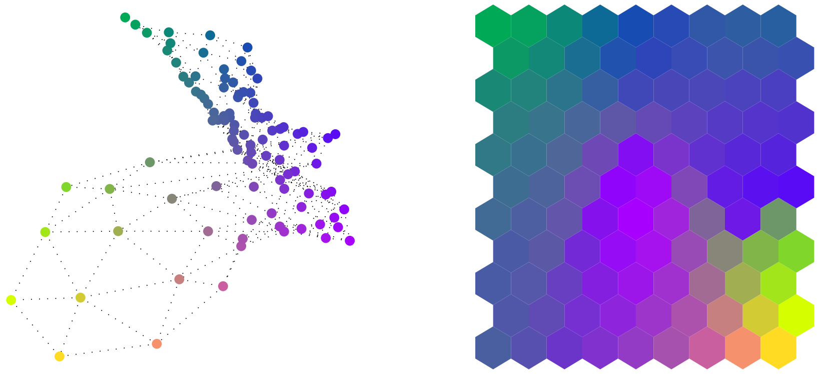
There is evidently a trade-off between color resolution and the shape of the original projection. If the projection that determines the colors is not square some of the potential color resolution is lost. To solve this, a 2D histogram equalization could be done for the projection before coloring, but then the input space distance visualization would be distorted.

A SOM which is taught using a process state data, may be used to track the process state transitions using a trajectory (see eg. [7]) which is a projection of the time series on the SOM grid visualization. The problem involved with this method is that the map folding may cause some jumps to the trajectory as the neighbourhood in the input space is split on the map.

This tool is thus useful for the process monitoring purposes. The combination of the SOM and a projection of its model vectors may reveal areas that are distant in the map, but near each other in the input space. The process state as a function of the time may be visualized as a trajectory. If the trajectory suddenly jumps, it may be caused by a sudden state transition, but in some cases it is due to a small change in an area in which is split due to the map folding.



(a) SOM grid $\mapsto$ Sammon's projection



(b) Sammon's projection $\mapsto$ SOM grid

Figure 2: Figure shows the connection between the SOM grid and a Sammon's projection of the map. Note that some resolution is lost in (b). As the original mapping was long and narrow it was first rotated so that the main axes are along the coordinate axes and then both axes were scaled between $[0, 1]$. This was done in order to get more color resolution. The map (and data) is the same as in Fig. 1.

5 Conclusions

In this paper the idea of "color linking" between two visual representations adopted from [2] is used to improve the SOM-based visualization of data.

Two tools were presented. The first one improves the SOM visualization for correlation detection and the second one gives a better view to relations between the SOM model vectors in the input space. These enhancements are based on combining information from two different data projections using a continuous color palette which provides a coding from spatial location to RGB-colors. The two tools provide extra information for explorative data-analysis, eg. for the needs of process monitoring.

The technics used are somewhat heuristic and bound by the fact that no method can provide a perfect projection of a high-dimensional space to two dimensions. There are two main drawbacks in the tools proposed. The "ghost correlations" due to the map grid, quantization and folding are a problem in the correlation detection in Sec. 3. A psycho-visually better color mapping would be crucial for getting an right impression of the data space distances in Sec. 4.

6 Acknowledgements

This work has been carried out in the technology program "Adaptive and Intelligent Systems Applications" financed by the Technology Development Center of Finland (TEKES). The co-operation of Leonia Bank, especially Mr. Kari Peltonen, is gratefully acknowledged. I thank also my colleagues Aapo Hyvärinen and Samuel Kaski for their valuable comments.

References

[1] Pierre Demartines and Jeanny Hérault. CCA: "curvilinear component analysis. In *Proc. of 15th workshop GRETSI. Juan-Les-Pins France*, 1995.

[2] Samuel Kaski, Teuvo Kohonen, and Jarkko Venna. Tips for SOM Processing and Colorcoding of Maps. In G. Deboeck and T. Kohonen, editors, *Visual explorations in Finance using Self-Organizing Maps*. Springer-Verlag, London, 1998. in press.

[3] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1995.

[4] Teuvo Kohonen, Erkki Oja, Olli Simula, Ari Visa, and Jari Kangas. Engineering applications of the self-organizing map. *Proceedings of the IEEE*, 84(10):27 pages, October 1996.

[5] John W. Sammon, Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409, 1969.

[6] Olli Simula, Esa Alhoniemi, Jaakko Hollmén, and Juha Vesanto. Monitoring and modeling of complex processes using hierarchical self-organizing maps. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'96)*, volume Supplement, pages 73–76, 1996.

[7] Viktor Tryba and Karl Goser. Self-Organizing Feature Maps for process control in chemistry. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, pages 847–852, Amsterdam, Netherlands, 1991. North-Holland.

[8] Alfred Ultsch. Self organized feature maps for monitoring and knowledge acquisition of a chemical process. In Stan Gielen and Bert Kappen, editors, *Proc. ICANN'93, Int. Conf. on Artificial Neural Networks*, pages 864–867, London, UK, 1993. Springer.