The SOM in data mining: analysis of world pulp and paper technology

Juha Vesanto
Juha.Vesanto@hut.fi
Helsinki University of Technology,
Laboratory of Computer and Information Science,
P.O.Box 2200, FIN-02015 HUT, Finland

3rd June 1997

Abstract

The Self-Organizing Map (SOM) is a powerful neural network method for the analysis and visualisation of high-dimensional data. In the Entire project, a data mining tool using the SOM was implemented and used to analyse world pulp and paper technology.

1 Introduction

In 1996 the joint Jaakko Pöyry Consulting/HUT project ENTIRE, funded by Jaakko Pöyry Consulting and the Technology Development Center of Finland (TEKES) set out to solve the problem of integrating environmental, technological and financial data of world pulp and paper industry. The goal was to create a software tool to help forming a comprehensive, combined view of the whole industry. Both the high dimensionality of the data and the need for visualisation suggested the use of the Self-Organising Map (SOM) [4] as the analysis tool. The extensive databanks of Jaakko Pöyry Consulting formed the foundation on which to build the analysis.

The data sets that were utilized in the case study included information on world's over 4000 pulp and paper mills, and the over 11000 paper machines and pulp lines in them. Similar studies regarding the world pulp and paper industry have never been reported. Instead the economical aspects of the industry were investigated with the SOM by Vanharanta *et al* [7]. Their results showed that the SOM is a promising tool for financial benchmarking. There are also other recent studies where the SOM has been used in financial diagnosis, for example by Serrano-Cinca who proposed a decision support system for financial diagnosis based on the SOM [6] and by Kiviluoto *et al.* who used the SOM in analysing financial statements [1]. Other studies in which the SOM has been used in comparing organizations, countries or companies based on socioeconomical figures have been made by Varfis *et al.* [8] and Kaski *et al* [3].

This report constitutes of three parts. The SOM is very shortly introduced in section 2. A data mining tool, Entire, based on the SOM is shortly presented is section 3 and then used in a data mining case study in section 4. This report is partially based on [10] and can be found in a much complete form from [9].

2 The Self-Organizing Map

The SOM, developed by academy professor Kohonen, is one of the most popular neural network models [4]. The SOM algorithm is based on unsupervised learning, which means that the training is entirely data-driven and that little or no *a priori* information about the input data is required. These and other properties of the SOM make it a valuable tool in data mining [2, 9].

A SOM consists of neurons which are organized on a regular, typically 2-dimensional grid. Each neuron of the SOM is represented by an n-dimensional weight vector, where n is equal to the dimension of the input vectors. The number of neurons may vary from a few dozen upto several thousand. The SOM algorithm performs a topology preserving mapping from the high-dimensional input space to the map units so that the relative distances between data points are preserved. Data points lying near each other in the input space are mapped to nearby map units of the SOM. The SOM can, thus, serve as a clustering tool of high-dimensional data. The SOM also has a capability to generalize, i.e. the network can interpolate between previously encountered inputs.

Each neuron of the SOM is represented by an n-dimensional weight, or model vector, $\mathbf{m} = [m_1, \ldots, m_n]$, where n is equal to the dimension of the input vectors. The neurons of the map are connected to adjacent neurons by a neighborhood relation, which dictates the topology, or the structure, of the map. The neighborhood of a neuron constitutes of its immediate neighbors.

The SOM is trained iteratively. In each training step, one sample vector $\mathbf{x}$ from the input data set is chosen randomly and the distance between it and all the weight vectors of the SOM is calculated using some distance measure, e.g Euclidian distance. The Best-Matching Unit (BMU) is the neuron whose weight vector is closest to the input vector $\mathbf{x}$. After finding the BMU, the weight vectors of the SOM are updated by moving the BMU closer to the input vector in the input space. The topological neighbors of the BMU are also treated in a similar way. This adaptation procedure stretches the BMU and its topological neighbors towards the sample vector.

The SOM forms an elastic net that during learning folds onto the "cloud" formed by the input data. The net tends to approximate the probability density function of the input data. This means that the weight vectors tend to drift there where the data is dense, while there are only a few weight vectors where data is sparsely located. A more throughout introduction to the theory and usage of SOM can be found professor Kohonen's book, "The Self-Organizing Map".

3 Data mining tool ENTIRE

As part of the Entire project, a data mining tool was implemented. The current version of ENTIRE is a prototype program, meant to be used as a platform for the use of the SOM in data mining. It concentrates on providing the basic properties and visualization tools of the SOM, along with some essential data pre- and postprocessing abilities. ENTIRE was built on top of the SOM_PAK program package [5]. The GUI was built using the XForms graphical user interface library for X windowing environment [11]. Below some of the most important properties of ENTIRE are listed.

Component scaling: The most important preprocessing method is component scaling which is a linear transformation performed for each vector component separately. It can be used to control the relative importance of different data components in training, determined by the variance of the component. An important property of component scaling is that it is reversible. To the user the scaled values of the data components hardly say as much as the original values. ENTIRE keeps track of the scaling factors and offsets applied to vector components and can thus reverse the preprocessing and report the original, unscaled data values to the user. When using non-linear preprocessing methods this is not always possible.

Component names: In early phases of development it was noticed that giving names to the data components helped considerably the data mining task. In ENTIRE, instead of having either to memorize the vector component order or to check the component name from a separate list, the names of the vector components are always displayed along with the vector. Another aspect which helps the user in data mining is the possibility to divide the vector components into groups. When a data vector is shown, the vector components are displayed in blocks so that all components belonging to a certain group are in one block.

Map training: ENTIRE offers a set of basic training options. The map structure is always a 1- or 2-dimensional regular lattice, based either on a hexagonal or a rectangular grid. For initialization ENTIRE offers two basic algorithms: random initialization with small random values and linear initialization along the subspace spanned by the two principal eigenvectors of the data set. In the training phase, the neighborhood function can be either bubble or gaussian and the learning coefficient can be selected to decrease either linearly or proportional to inverse of time. To evaluate the quality of the maps, several quality measures have been implemented. These include average quantization error, topographic error and topological quantization error.

Map visualization: Perhaps the most important features of ENTIRE are its visualization capabilities. The visualization is based on two different kinds of presentations of the map: the component planes and the u-matrix. Both are visualized as 2-dimensional colour images, where the colour of a map unit corresponds to its value. On top of these images different kinds of markings can be drawn including labels and operation point trajectories. The ability to show several component planes at the same time is an important property, because it allows direct visual inspection of correlations between components. In addition to component planes and u-matrix, the user can calculate and visualize the Sammon's projection of the SOM.

Colour maps: The basic colour map used by ENTIRE is a gray scale of 100 shades ranging from pure white to pure black. The colour map can be changed, however. Using the colour controls the user can select the number of colours used, the range of values they correspond to and the specific colours themselves. Figure 1 shows four different kinds of views of a component plane. In the first image (top left) the standard set of 100 shades of gray are used. In the second image (top right) only 5 colours are used. Also in the third image (bottom left) 5 colours are used, but their positioning is logarithmic rather than equidistant. In the last image (bottom right) only two colours are used. Using a smaller number of colours is sometimes beneficial because it visualizes better the categories of values that can be found from the map.

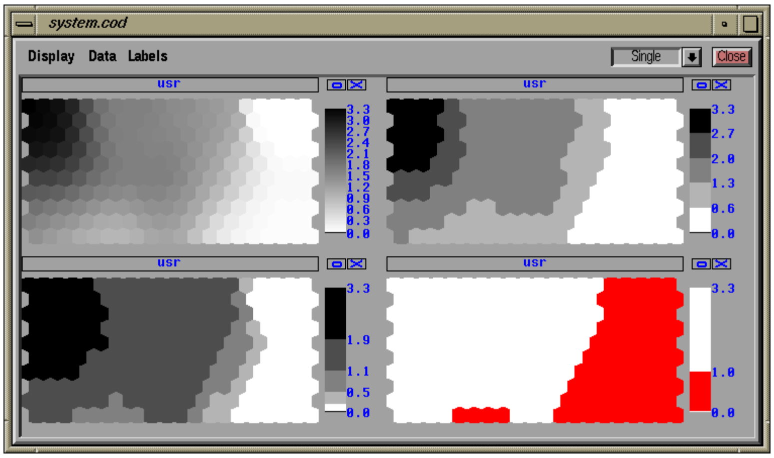


Figure 1: *An example of a component plane visualization with four different kinds of colour maps.*

Data projections: ENTIRE can be used to investigate local data sets. The local data sets are obtained by taking a set of input vectors and assigning each vector to the local data set of its BMU. The number of samples in each map unit gives the distribution of the data set on the map. In ENTIRE the data histograms can be visualized by showing the number of samples in the set, or as a square the sidelength of which is proportional to the number of samples in the set. The local data sets can also be investigated in detail simply by selecting the corresponding map unit. For time-series data analysis ENTIRE offers a trajectory tool. The trajectory of a data set is formed of the consecutive BMUs of the data vectors in the data set.

ENTIRE can also be used to project arbitrary user-provided vectors on the map. The tool can be used to find multiple BMUs. Searching multiple BMUs visualizes well the ordering of the map and the familiarity of the sample vector: if the found BMUs are neighboring units, it can be assumed that the map is well organized. On the other hand BMUs far from each other mean a folding in the map topology: either the map is poorly trained, or the sample vector is atypical.

Postprocessing: Data postprocessing is an important part of data analysis. The primary postprocessing tool in ENTIRE is labeling. The map can be labeled either automatically or manually. ENTIRE offers four labeling methods: autolabeling, labeling by unit selection, labeling by BMU search and labeling by component value range. The autolabeling method is based on having a labeled set of sample vectors. The BMUs of these vectors are searched and the units are labeled with the labels of the vectors. This can be done also the other way around: a data set can be autolabeled using a labeled SOM. Other labeling methods are interactive. In unit selection method the user manually selects a set of units and gives them a common label. In BMU search method a number of BMUs of a sample vector are searched and the units are given a common label. In component value range method the user selects a range of values for a certain component and gives all units having the weight vector component value in that range a common label.

Another postprocessing approach is to form new data sets from the data projections. For example a data set can be constructed of the BMU coordinates and quantization errors of each data vector in a data set. The new data sets can then be further analysed using some other method. This postprocessing method also forms the basis of operation of the hierarchical maps.

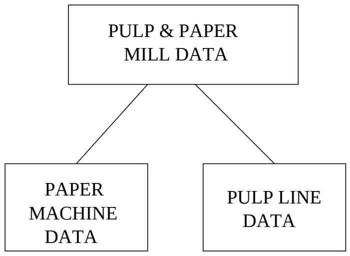
Futher development: In the future some important features are added to ENTIRE. One such feature is visualization using 3D-graphics. While 2D-images are adequate in most cases, they sometimes pale when compared with the visualization power of 3D-presentation. Other future features include:

- tools for easy construction of hierarchical architectures of SOMs
- other map topologies besides the rectangular: toroid and cylinder
- automatic clustering and interpretation tools possibly using fuzzy algorithms.

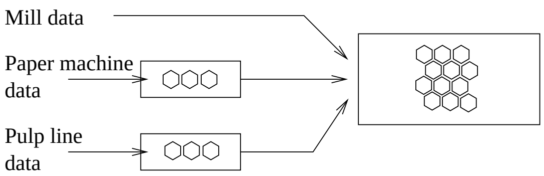
4 Pulp and paper technology

In the study of world pulp and paper technology three data sets were utilized. The first one included information on the production capacities of the mills, the second included information on the technology of the paper machines and the third on the technology of the pulp lines, see figure 2a.

Combining the information provided by the data sets was not trivial. Since each mill could contain several paper machines and pulp lines, a simple concatenation of data vectors from each set could not be used. As a solution, a hierarchical structure of maps was used as shown in figure 2b. The two low level maps were constructed from the paper machine and pulp line data sets. These maps provided a clustering of the different machine types. The actual technology-map was trained using the mill-specific info in the mill data set and the data histograms from the two low-level maps.



(a) Data sets



(b) Hierarchical map structure

Figure 2: *(a) There were three technological data sets: one of mill production capacities, one of paper machines and one of pulp lines. Each mill could contain several paper machines and pulp lines. (b) The hierarchical map structure. Data histrograms from the two smaller maps were utilized in the training of the third map. The arrows show which data sets were used in training the maps.*

4.1 Paper machines and pulp lines

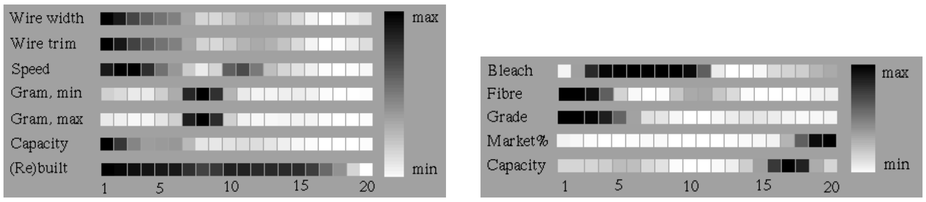
Using paper machine and pulp line data, two 20x1 sized maps were trained. The vector component values in the map units can be seen from the component plane representations in figure 3. By visual inspection of the paper machine map the three first components have a strong correlation, as do the next two components. In the pulp line map none such global correlations can be seen.

The interpretation of the maps was based on visual inspection of the u-matrixes supplemented by the knowledge of the distribution of component values. Both paper machines and pulp lines can be divided into three different types. The three main types of paper machines in the world are:

1. The new machines with highest capacity and speed and widest wire.
2. The paper machines with big paper weight.
3. The majority of the paper machines belong to the third type the capacity and size of which decrease steadily with increasing age.

The three main types of pulp lines are:

1. Those using waste paper for pulp.
2. Those making unbleached pulp, further divided into three subtypes:
   (a) wood fibre and semichemical pulp grade
   (b) wood fibre and either chemical or mechanical pulp grade
   (c) other virgin fibers with chemical pulp grade
3. Those using chemical pulp grade to produce mainly bleached pulp from wood fibre. This type could further be divided into three subtypes: those with small capacity, those with high capacity and those with a high market ratio of production.



(a) Paper machine data components

(b) Pulp line data components

Figure 3: *The component planes of the paper machine (a) and pulp line (b) maps. The name of the component on the right and the corresponding values in the map units from 1 to 20 in the middle. All components are presented with a grayscale with black representing the maximum value and white the minimum.*

4.2 Mill techonology

The mill data set consisted of the total pulp and paper production capacities, the number of paper machines, the number of coaters and the percentage of total production of different pulp and paper types for each mill. This information was augmented by the data histograms from the paper machine and pulp line maps. For each mill the two data histograms were constructed using the paper machines and pulp lines in that particular mill. All in all the technology data set consisted of 4205 75-dimensional vectors each corresponding to one mill. A 40x25-sized SOM was trained with this data.

Several correlations between components could be seen from the component plane representation of the technology map, such as that the mills with high paper production capacity usually also have a high pulp production and that newsprint production corresponded to high thermomechanical pulp production. Of special interest were correlations between different pulp and paper types and the distribution of values in the components corresponding to the low level maps (pm1–20, pulp1–20). By observing the component planes it could also be seen that for some components the high values were divided between clearly different types of mills. For example there were two types of mills that produce chemimechanical pulp: those that also produce wood containing paper and those that concentrate purely on producing pulp.

An analysis of mill-types was performed by first dividing the map into 31 initial clusters. After this all components were separated to three ranges and it was investigated which clusters had component values in the two highest thirds. Finally each cluster was examined to see what kind of mill-type it corresponded to. During this procedure some of the initial clusters were combined. For example the SmallInd cluster was combined from several initial clusters corresponding to mills producing various industrial papers. The only truely separating factor between the initial clusters was that they received data from different units of the paper machine map: units 13 – 20 corresponding to old and small paper machines. The cluster analysis resulted in the 20 different mill-types described in table 1 and shown in figure 4.

A couple of important notices are in place here. Firstly, the set of 20 mill-types is by no means definitive. Depending on the desired precision also several other clusters could be extracted from the map. For example the Pulp cluster can be divided into three subtypes based on the pulp type and the production capacity. Secondly, the clustering is heavily influenced by the choice and scaling of the components. The mill-types achieved in this study reflect only the information present in the used data components. By changing part of the components, or even scaling them differently, a different set of mill-types could be achieved. Thirdly, the analysis made so far details the different clusters only in very sketchy terms. To get a better view of the different types, performing a more detailed in-cluster or in-type analysis would be beneficial.

4.3 Geographical areas

It is interesting to note that some mill-types are typical for certain geographical regions. For the analysis of different geographical areas the data was separated to 11 sets each consisting of pulp and paper mills in a certain area. The data sets were projected on the map and based on the resulting histograms some conclusions can be drawn for each region. For example Scandinavia has mostly new, high-capacity mills the majority of which produce printing/writing papers and pulp. North America has in addition a large number of old and small industrial paper mills. China is a special case: their mills have many machines and the mills produce both industrial and printing/writing papers.

5 Conclusion

The Self-Organizing Map (SOM) is a versatile tool for exploring data sets. It is an effective clustering method and it has excellent visualization capabilities including techniques which use the weight vectors of the SOM to give an informative picture of the data space, and techniques which use data projections to compare data vectors or whole data sets with each other. The visualization capabilities of the SOM make it a valuable tool in data summarization and in consolidating the discovered knowledge. The SOM can also be used for regression and modeling or as a preprocessing stage for other methods [9].

The SOM was used to analyse the technology of world pulp and paper industry with encouraging results. However, much work is still needed regarding the post-

Table 1: *Pulp and paper mill types. The clusters refer to the areas marked on the u-matrix of the mill-technology map in figure 4.*

Type	Description
UncWF	Uncoated woodfree paper, either new big machines or older and smaller.
SmallInd	Various industrial papers, machines old and small.
Tissue	Tissue paper, average wirelength and speed.
DeWa	Some tissue paper, but especially high deinked waste paper usage.
CoWF	Many coaters, coated woodfree paper, machines of average capacity and speed.
PMs	Many paper machines including some high capacity paper machines, uncoated woodfree but also various industrial papers, unbleached and semibleached sulphite pulp.
DiWa	Cartonboard, linerboard and fluting papers, disperged waste paper for pulp.
BlWF	Uncoated woodfree paper, bleached chemical pulp from wood fibre.
Big	High capacity, many machines and coaters, woodfree paper or linerboard, pulp is chemical (sulphate), machines are big.
WrLi	Wrapping paper and linerboard, unbleached sulphate pulp, big paper machines.
Mech	Various industrial papers, pulp is unbleached and mechanical: groundwood or rmp.
AveInd	Cartonboard, linerboard and fluting paper, average capacity machines.
Flut	Fluting paper, semichemical unbleached pulp, average capacity paper machines.
WoodC	Wood containing paper, chemimechanical or mechanical pulp from wood fibre, average capacity paper machines.
News	High paper capacity: newsprint, thermomechanical pulp, high capacity paper machines and pulp lines.
Pulp	No paper production, but large pulp production, high capacity pulp lines, big market percentage.
Carton	Cartonboard and other papers, big weight in paper machines.
LinFl1	Linerboard and fluting paper, small to average sized machines.
LinFl2	Linerboard and fluting paper, high capacity machines.
BigInd	Cartonboard, wrapping, tissue and other papers, high capacity machines.

processing stage and the interpretation of results. The development and automated usage of algorithms that cluster the units of the SOM will be an essential part of future work. All in all, the many abilities of the SOM together with its robustness and flexibility are a combination which makes the SOM a prime tool in knowledge discovery and data mining.

6 Acknowledgments

This work has been carried out in the technology program "Adaptive and Intelligent Systems Applications". The work has been financed by Jaakko Pöyry Consulting and the Technology Development Center of Finland (TEKES).

References

[1] P. Bergius K. Kiviluoto. Analyzing financial statements with the self-organizing map. In *Proc. of the Workshop on Self-Organizing Map*, 1997. in press.

[2] S. Kaski. *Data Exploration Using Self-Organizing Maps*. PhD thesis, Helsinki University of Technology, 1997.

[3] S. Kaski and T. Kohonen. Structures of welfare and poverty in the world discovered by the self-organizing map. Technical Report A24, Helsinki University of Technology, Laboratory of Computer and Information Science, February 1995.

[4] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995.

[5] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. SOM_PAK: The self-organizing map program package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, 1996. Program package available via WWW at URL http://nucleus.hut.fi/nnrc/som_pak/.

[6] C. Serrano-Cinca. Self-organizing neural networks for financial diagnosis. *Decision Support Systems*, 1996. to be published.

[7] H. Vanharanta, K. Sere, M. Irjala, and B. Back. Competitive financial benchmarking using self-organizing maps. *Paper and Timber*, 79(1):42 – 49, 1997.

[8] A. Varfis and C. Versino. Clustering of socio-economic data with kohonen maps. *Neural Network World*, 2:813 – 833, 1992.

[9] J. Vesanto. Data mining techniques based on the self-organizing map. Master's thesis, Helsinki University of Technology, May 1997.

[10] J. Vesanto, P. Vasara, R-R. Helminen, and O. Simula. Integrating environmental, technological and financial data inforest industry analysis. In *Proc. of SNN'97*. Stichting Neurale Netwerken, Netherlands, 1997.

[11] T.C. Zhao and M. Overmars. Forms library: A graphical user interface toolkit for x. http://bragg.phys.uwm.edu/xforms, 1996.

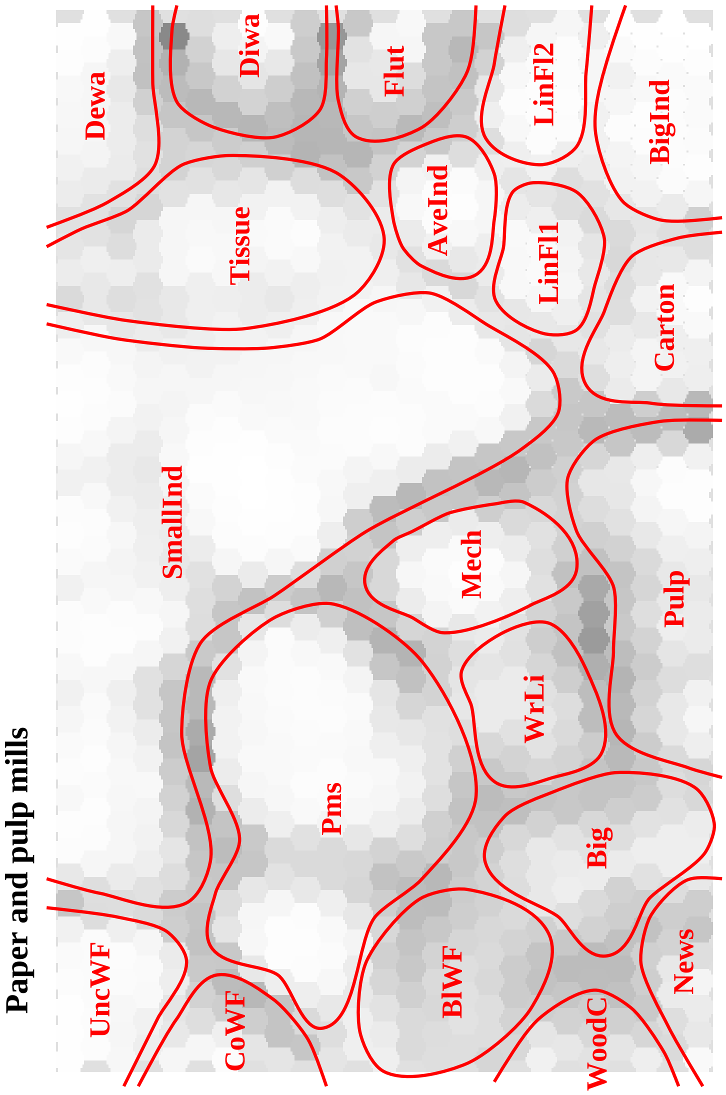


Figure 4: *The u-matrix of the mill-technology map. Black corresponds to high value of the u-matrix, white to small value. The regions corresponding to different mill types have been marked on the map and labeled as in table 1.*