Proceedings of the 23rd IASTED International Multi-Conference
DATABASES AND APPLICATIONS
February 14-16, 2005, Innsbruck, Austria

ASSESSING THE USE OF THE SOM TECHNIQUE IN DATA MINING

Dorina Marghescu
Turku Centre for Computer Science
Åbo Akademi University
Lemminkäisenkatu 14 A, 5th floor
FIN-20520 Turku, Finland
E-mail: dorina.marghescu@abo.fi

Mikko J. Rajanen
Dept. of Information Processing Science
University of Oulu
PO Box 3000
FIN-90014 Oulu,, Finland
E-mail: mikko.rajanen@oulu.fi

ABSTRACT
The Self-Organizing Maps method is a special type of neural network used in clustering, visualization and abstraction. In this paper, we evaluate to what extent users of the SOM technique are satisfied with this tool in visualizing large amounts of data. The contribution of the paper consists of identifying the factors that influence the quality of use of SOM tools at the three levels considered for analysis: visualization of data, interaction with the tool, and information obtained.

KEYWORDS
SOM, visualization, evaluation, quality of use

1. Introduction

There is an increasing demand to provide decision makers, from middle management upward, with correct information and at different levels of detail. Data mining applications support this need, being used for knowledge discovery, the process of searching data for unanticipated new knowledge.

Typically, data mining is used in problems such as prediction, identification, classification, and optimization. The ways in which knowledge discovery is represented during data mining process are: association rules, classification hierarchies, sequential patterns, patterns within time series, and categorization and segmentation [1]. One of the current challenges in data mining research is to find ways of representing data into an accessible and understandable format for end-users.

Therefore, the advances in information visualization gain more and more interest among the data-mining applications designers. This is because the information visualization techniques it is believed to contribute to the creation of visual interfaces for large-scale databases and document collections [2]. However, the problem of usability of these techniques is rarely addressed in the literature.

In this paper, we look at one particular visualization technique, the Self-Organizing Maps (SOM) method [3]. The SOM method is a special type of neural network used in clustering, visualization and abstraction. Several studies demonstrate the benefits of using the SOM in analyzing massive sets of data in finance [4,5,6], industrial processes [7], macroeconomics [8], medicine, biology, and other fields [9,10]. In this paper, we examine the attitude and opinions of potential users regarding the use of SOM in visualizing large volumes of data. We perform a measurement of quality of use of SOM tools based on data collected in a questionnaire survey in November 2003.

The aim of this study is to identify the factors contributing to the quality of use of SOM tools. In a previous paper [11] we showed that for evaluating the user satisfaction with a visual data-mining tool we need to consider three levels of assessment: visualization, interaction and information. In this article, we continue the previous study with exploratory factor analysis in order to identify the factors contributing to the quality of use of SOM tools, at each of the three levels of assessment. We illustrate how the SOM characteristics reflect on the user satisfaction, and discuss the user performance with the technique.

The paper is organized as follows. Section 2 describes briefly the SOM algorithm and the steps of applying it in data mining. Section 3 illustrates the use of the SOM with financial data. Section 4 presents the exploratory study conducted to evaluate the quality of use of the SOM tools. Section 5 presents the results of the study. The conclusions and future work are presented in Section 6.

2. Self-Organizing Maps

The SOM method is an effective technique for the visualization of high-dimensional data. Kohonen [3] developed the SOM algorithm in 1982, describing a nonlinear, ordered, and smooth mapping of high-dimensional input data vectors onto the elements of a regular, low-dimensional array. The way the mapping is implemented resembles the classical vector quantization.

To describe formally the self-organizing process, the set of input variables (ξ_j) is defined as a real vector $x = [\xi_1, \xi_2, ..., \xi_n]^T \in \Re^n$. The SOM array consists of i nodes (neurons), each node having associate a parametric real vector $m_i = [\mu_{i1}, \mu_{i2,...,} \mu_{in}]^T \in \Re^n$, which is called a *model* (or reference) vector.

Each input vector x has an image on the SOM array, which is defined as the array element m_c that matches best

with x. The matching is computed using a distance function, $d(x,m_i)$, which can be the Euclidian distance.

Applying the distance function to all input vectors, it is defined a mapping from the n-dimensional input data space to the two-dimensional SOM array. The SOM task is to define the m_i in such a way that the mapping is ordered and descriptive of the distribution of x.

In the learning process, when model vectors are updated, those nodes that are topographically close in the array to the current best-matching node will activate each other to learn the attributes' values from the input x. This will result in a local relaxation or smoothing effect on the model vectors in this neighborhood, which in continued learning leads to global ordering.

To illustrate the clustering of model vectors in the SOM, a *graphic display* called *U-matrix* has been developed [3], in which the average distances between the neighboring reference vectors are represented by shades in a gray scale. If the average distance of the neighboring neurons is small, then a light shade is used; and vice versa, dark shades represent large distances. A "cluster landscape" formed over the SOM then clearly visualizes the classification.

SOM algorithm is implemented in different software packages [3], and three of them will be considered for analysis in this paper. They are SOM_PAK [12] and SOM Toolbox for Matlab [13], developed by the Laboratory of Computer and Information Science of Helsinki University of Technology, and Nenet [14] developed by the Nenet Team of Helsinki University of Technology.

Regardless the tool used the main steps in applying SOM for data mining are depicted in Figure 1:

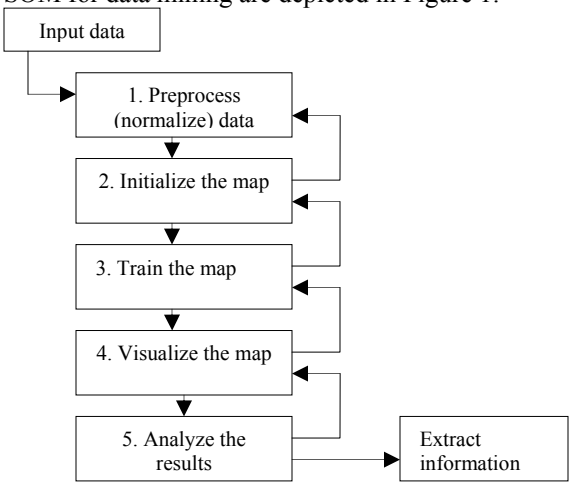


Figure 1 The steps in using Self-Organizing Maps for data mining

The data mining process is not always straightforward, and some of the steps may be repeated several times. At the initialization step, the following parameters have to be set: topology (rectangular or hexagonal – the latter is preferred), size (X and Y- dimensions), and initialization of the map units (random or linear). At the training step, the parameters are: neighborhood function type ("bubble" or Gaussian), neighborhood radius, learning type (batch or sequential), learning rate, and learning length.

3. The Use of the SOM Tools with Financial Data

To illustrate the capabilities of the SOM to find patterns in financial data, we use a set of data involving 77 companies from pulp and paper industry from different countries, observed during the years 1997 and 1998. The size of the data matrix is 184 samples x 5 variables. The countries averages are included in the data set. The five variables represent financial ratios that characterize the performance of companies. These data and ratios are selected according to [6]. The ratios represent profitability (Operating margin and ROTA), solvency (Equity to capital and Interest coverage), and efficiency (Receivables turnover). Assuming the input data available, the interest of a data analyst could be focused on the following problems:

- How many clusters can be found in the data and what are the characteristics of each cluster?
- Which cluster yields as the best performing one? Which cluster is shown as the poorest one?
- Which is the performance of certain companies of interest?
- What can be said about benchmarking a certain number of companies against each other?

To answer these questions, we must apply the algorithm several times conform to Figure 1, until we obtain a map which is satisfactory in terms of both accuracy and visual clarity. First, the data are normalized so that the variance of each variable becomes equal to 1. The map we selected for decision making has the size of 6 x 5 neurons. The initial neighborhood radius is 6 neurons, and it decreases in time down to 1. The neighborhood function type is bubble, and the map units are linearly initialized. The algorithm used for learning is the batch train algorithm.

Figure 2 illustrates the clustering of the input vectors' models by using U-matrix method for visualization. The borders between neurons represent how similar in terms of a distance measure are two neurons. Dark shades of the borders account for large distances between data mapped into the corresponding two neurons, and light shades signify similarities between the data. Looking at the border shades we can distinguish the clusters existing in the data.

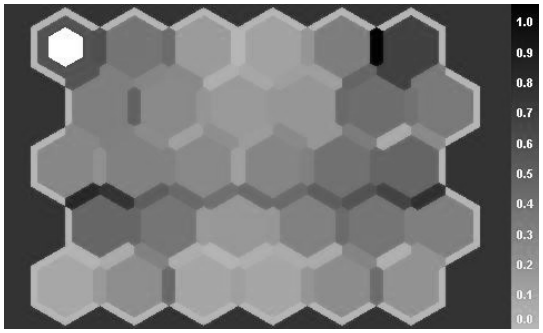


Figure 2 Screen shot from Nenet that displays SOM using the U-matrix method

In Figure 3, we have strengthened the borders with a thick dark line and we have obtained six clusters. The neuron with a white spot is the "active" node, having the values of each attribute listed in the lower-left corner. The upper-left corner box shows the companies matched to the selected neuron.

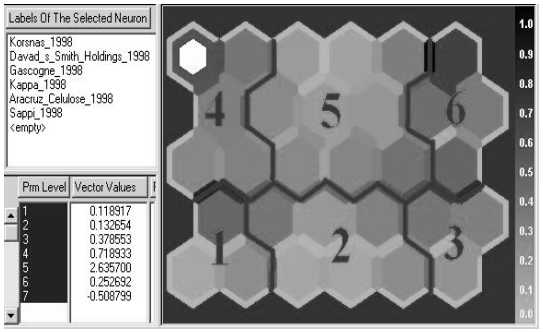
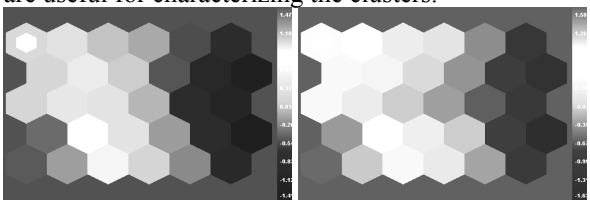
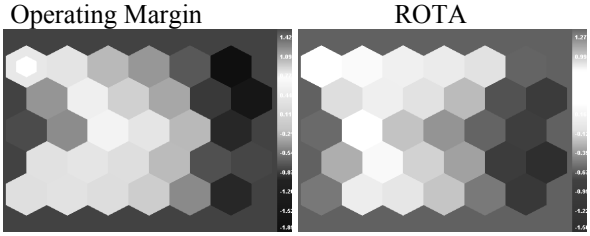


Figure 3 Clusters and clusters' identifiers

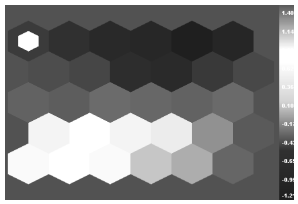
Along with the U-matrix, the SOM tools provide the visualization of the maps (feature planes) associated to each input variable (Figure 4). These feature planes show the distribution of input data for each variable and they are useful for characterizing the clusters.



Operating Margin ROTA



Equity to capital Interest Coverage



Receivables turnover

Figure 4 Feature planes corresponding to each of the input data dimension

According to the grayscale bars assigned to each feature plane, very dark tones similar to black represent very low values of the ratios, bright tones of gray represent medium and above medium values, while the gray tone corresponding to the upper part of the scale shows high values[1]. By examining the feature planes, one obtains description of the data clusters as follows. Cluster 1 shows excellent profitability and solvency, and very good efficiency. Cluster 2 contains very efficient companies but with medium values for profitability and solvency. Cluster 3 reveals very low profitability and solvency, but very high efficiency. Cluster 4 is characterized by very good solvency, good or medium profitability, and very low efficiency. Cluster 5 is characterized by medium profitability and solvency, and poor efficiency. Cluster 6 is the worst cluster, containing the poorest companies in all respects: profitability, solvency, and efficiency.

We have illustrated above how the SOM technique can be used to visualize multidimensional financial data, and facilitate the decision maker to rate and benchmark different companies of interest according to certain criteria or importance of the ratios.

In the next section, we evaluate the extent to which potential users are satisfied with the SOM technique.

4. Exploratory Study: Evaluating Quality of Use of the SOM Tools

Quality of use represents the totality of features and characteristics of the tool that account for its ability to satisfy the users' needs. It reflects the user satisfaction with all the features of the tool. In order to evaluate the quality of use of the SOM tools, we conducted an exploratory study in which the data collection process consisted of three phases. First, participants were trained to use the SOM technique. Second, they were asked to solve a task and report their findings (qualitative data). Third, the participants filled a questionnaire (providing quantitative data for the study).

4.1 Participants

The participants were 26 students, enrolled for an Information Systems course, in a public university.

4.2 Task

The task required the students to train for several times different SOMs with the input data provided, until they obtain a map on which to visualize the data and identify the clusters. Students were asked to answer five questions and document their solutions in a report. The questions were: 1: How many clusters do you identify and what are the characteristics of each cluster? 2: Which is the cluster that contains the best performers in the market? 3: Which is the cluster that contains the worst performers in the market? 4: Discuss the performance of three specific companies based on their positions on the map and compare the results with the real data from the file provided. 5: Benchmark five specific companies one against the other, based on their positions on the map.

[1] The presentation of the feature planes and U-matrix is in grayscale due to printing requirements for these Proceedings. The Nenet provides colored maps, for a better visualization on the computer screen.

4.3 Materials

We have used three software packages, which have implemented the SOM algorithm, all being available online for downloading. These were SOM_PAK, SOM Toolbox for Matlab, and Nenet. Nenet was definitely preferred by all students, for visualizing the maps, while different students used either SOM_PAK or SOM Toolbox to train the maps. We have used Binomial, and Chi-square tests to find whether there are differences in attitudes or opinions of the SOM_PAK and SOM Toolbox users, but no differences were found significant.

4.4 The quality attributes

For evaluating the quality of use of SOM tools, we considered three levels of assessment [11]: visualization, interaction, and information. For each level, we selected and included in the questionnaire a number of attributes to be assessed (Tables 1, 2 and 3).

Table 1 Attributes of visualization

Initial settings	- Requirements on input data format - Adequacy of normalized data - Easy to understand parameters - Easy to use parameters
Data display	- Data structure: Data clusters, trends, attribute values, correlations between attributes - Data content: Exploration and description of data - Data variation - Data comparison - Tabulation of data - Decoration of data - Labeling of data - Dimensionality and size of the graphic
Reporting functions	- Thinking about what is seen: - Substance of the data - Design elements - Computational issues - Easy to integrate the resulting maps within other software applications

Table 2 Attributes of interaction with the tool

Ease of use	- Too many steps required - Easy to use tool
Learnability	- Easy to learn tool - Satisfaction with learnability
Efficiency	- Time needed to obtain a good map - Provides the information needed
Accuracy	- Satisfaction with the accuracy of the system

Table 3 Attributes of information

Richness	Reliable, complete, interesting, needed, useful
Accuracy	Accurate, precise, correct
Clarity	Clear and understandable, Easy to interpret
Novelty	New

All questions used a 5-point scale, e.g. *very good; good; medium; poor; very poor*, but in the analysis we have mapped the answers on a 3-point scale as follows: *very good + good; medium; poor + very poor.*

4.5. Data analysis

We performed exploratory factor analysis on the quantitative data collected in the questionnaire survey. We used the Principal Axis Factoring technique with Varimax Rotation method in SPSS 11.5, and we obtained the factors that account for the variance in each of the three levels of quality: visualization, interaction, and information. The variables used in the analysis and the factors obtained are listed in Table 4. The variables are displayed in the table so that it is clear to which factor they correspond. The factors loadings *(f)*, considered for identifying the factors, are showed in the same table.

Partly due to the small number of participants, the KMO measures [15] computed for assessing the adequacy of data for factor analysis were relatively low, but still acceptable. The Bartlett test [15], which assesses whether the correlation matrix is appropriate for factoring, shows good values for the significance levels (i.e. $p \rightarrow 0$).

Table 4 Factors and their variables

Level	Factors	Variables and factor loadings
Visualization quality *KMO = .550* *Bartlett test: p=.006* *Cronbach's alpha= .72*	- Description of data: *(α=.65)*	- Data description *(f=.736)* - Data comparison *(f=.608)*
	- Explanation of data: *(α=.64)*	- Labeling of data *(f=.863)* - Attributes values *(f=.462)* - Decoration of data (use of colors, lines, etc.) *(f=.605)*
	- Exploration of data: *(α=.69)*	- Exploration of data *(f=.160)* - Tabulation of data *(f=.822)* - Dimensionality of data *(f=.94)*
	- Graphic elements: *(α=.57)*	- Data variation *(f=.706)* - No problems in reading the graphic content *(f=.565)*
Interaction quality *KMO = .55* *Bartlett test: p=.04* *Cronbach's alpha=.65*	- Preparation for usage*: (α=.64*)	- Easy to meet requirements for the data *(f=.506)* - Easy to understand parameters *(f=.536)* - Ease to learn tool *(f=.842)*
	- Ease of use: *(α=.67)*	- Easy to use parameters *(f=.709)* - Easy to use tool *(f=.662)* - Number of steps required to get a good map *(f=.312)* - Satisfaction with the time *(f=.697)*
Information quality *KMO = .555* *Bartlett test: p=.000* *Cronbach's alpha=.78*	- Ease of interpreting: *(α=.80)*	- Clear and understandable *(f=.51)* - Easy to interpret *(f=.658)* - Precise *(f=.771)* - Correct *(f=.663)*
	- Usefulness: *(α=.68)*	- Satisfaction with the information content *(f=.607)* - Useful *(f=.796)*
	- Reliability: *(α=.68)*	- Reliable *(f=.738)* - Complete *(f=.813)*
	- Accuracy: *(α=.56)*	- Satisfaction with information correctness *(f=.8)* - Accurate information *(f=.52)*
	- Novelty*: (α=.55)*	- New *(f=.641)* - Interesting *(f=.263)* - Needed for the task *(f=.771)*

To check the consistency of the constructs, we examined the Cronbach's alpha value for each of the levels and latent factors *(α)* and the results are showed in Table 4. The overall Cronbach's alpha value is 0.82. A

rule of thumb says that values of Cronbach's alpha over 0.7 are acceptable. The smaller values of Cronbach's alpha corresponding to the latent factors may be due to the small number of observations available in the study. The selection of the variables of each scale has been based on the factor analysis results. We eliminated from the initial list of attributes those which by including them would have decreased the KMO or Cronbach's alpha values.

5. Results

5.1. Quantitative data

The quality of visualization is presented in Figure 5. According to Table 4, the factors, which influence the quality of visualization, can be formulated in terms of description of data, explanation of data, exploration of data, and graphic elements. Each factor explains some of the relationships between the variables that are observed. These factors accounts for 58.25% of the variance in the quality of visualization data. Description of data accounts mostly for how well the data comparison, description and attributes values are represented in the visualization. In the case of SOM, students found good visualization of comparable data (85% of the students agreed), and good depiction of attributes values (58% agreed, and 23% were neutral). Explanation of data is associated with labeling of data (which was good for 46%) and decoration of data – the use of colors and lines to describe the data (good for 31% and medium for 46%). Exploration of data is found good by 61% of the respondents. It is observed that dimensionality of data is considered adequate only by 38%, while tabulation of data is appropriate only by 27%. Students found well represented the variation of data (65%) – i.e. different graphic elements display different pieces of information, and about 38% of the users did not have problems in interpreting the graphic content, but 42% of them did have problems.

Quality of interaction is depicted in Figure 6. The two factors identified to account for the 55.59% of the variance are: preparation for usage and ease of use. About 69% of the students found the tools easy to learn, though for 27% of the students the parameters were not easy to understand. Data requirements were considered easy to meet by 73% of the users. The tools were found also easy to use (by 65% of the respondents), the parameters easy to use/change (by 69% of the respondents), but the time required to get a good map was satisfactory only for 27%. This is perhaps due to the fact that many trials are required until an appropriate map is obtained. Therefore, most of the people (58%) considered that there are too many steps needed to get a good map.

Figure 7 depicts the quality of information obtained. The factors responsible for 66.24% of the variance in the data are: ease of interpreting, usefulness, reliability, accuracy, and novelty. 73% of the students agreed that the format of the information is clear and understandable, but a less number found it correct (58%), and easy to interpret (54%). Preciseness seemed to be the weakest feature, because only 35% of the respondents found the information precise, while 23% found it imprecise. The usefulness and satisfaction with the information content are quite high (65% of the users gave positive answers), and even much higher is the confidence in the information obtained (81% of the students found it reliable). Satisfaction with correctness is relatively good (46% agreed, 46% neutral). Very encouraging were the ratings recorded for the novelty. The information was found needed by 88%, interesting by 81%, and new by 58% of the students.

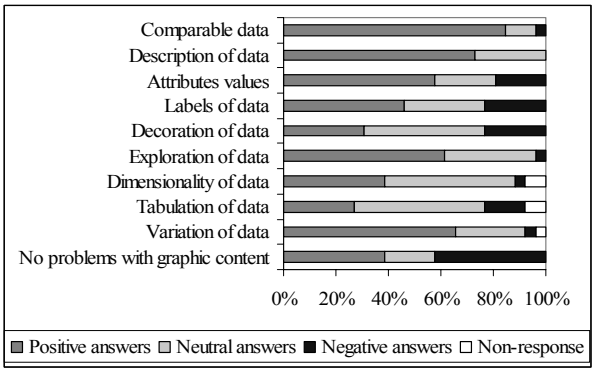


Figure 5 Quality of visualization

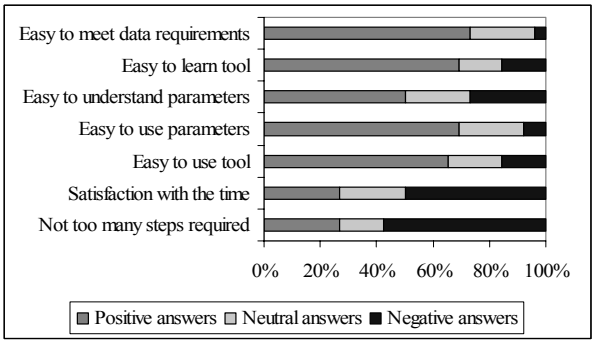


Figure 6 Quality of interaction

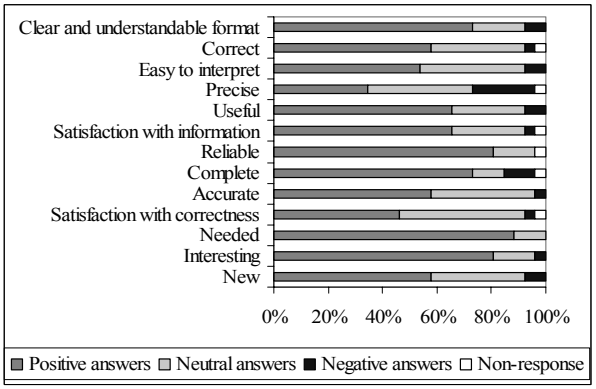


Figure 7 Quality of information

Regarding the usefulness of the technique in data mining, the SOM visualization was found helpful in finding clusters (92.3% agreed), comparing data (84.6% agreed), finding trends in data (73,1% agreed), and correlations between attributes (50% agreed).

5.2. Qualitative data

Examining the solutions to the task given, we observed that 65.3 per cent of the students identified and described satisfactory the clusters found. The SOM tools are very flexible, in the sense that they allow the user to have the control over the parameters so that a suitable map is obtained. Consequently, different users obtain different visualizations and the criteria for distinguishing between different clusters belong to the analyst. The number of clusters found depends on the interest of the user, too. In the case of financial benchmarking, for different analysts some financial ratios may have different importance, depending on the purpose of the study. Therefore the comparison between all solutions obtained is subjective. In our case, the correctness and acceptability of the 65.3% of the solutions account for both quality of the clusters (whether they group similar companies) and quality of description of each cluster.

For the second and third questions (identifying the clusters containing the best, respectively worst companies in terms of financial performance) we received quite extreme answers. 76.9% of the students gave good and acceptable answers to the second question, while only 30.7% of the students gave acceptable answers to the third question. However, it must be noticed that in the score assigned to each solution, we took into account whether the user considered in the analysis both years for which the data was available. Many students did not bring into discussion the years of observation, and one reason of relatively low quality solutions is actually the incompleteness of the reports, rather than the misunderstanding or poor quality of the maps.

When asked to characterize specific companies from the map and to benchmark certain companies against each other, all the students gave better solutions, more precise and correctly argued.

As a final remark, we notice that participants in the study felt challenged to learn the SOM technique, but also interested to use a new method for analyzing data.

6. Conclusions

In this paper, we have illustrated the potential of the SOM technique in finding patterns in financial data. We have also evaluated to what extent users of the technique were satisfied with the visualization, interaction, and information obtained. Overall, the user satisfaction was high, and users showed themselves interested in using and learning the SOM technique. The user performance was relatively good, and perhaps by enhancing the usability and functionality of the SOM tools, the level of performance and usage of this data mining technique will increase.

We have identified the factors contributing to the user satisfaction with all the three features of the tool, visualization, interaction and information. The low values computed for Kaiser-Meyer-Olkin (KMO) measure indicates that the data sampling adequacy is relatively low but still acceptable. One reason for this may be the sample size which is very small. It would be interesting to extend the analysis to a larger number of users of SOM and compare the results. Moreover, to test whether the factors identified are common to other visualization techniques, extensive evaluation of other data mining tools is needed.

References:

[1] R. Elmasri, & S. Navathe, *Fundamentals of database systems*, *third edition* (Addison–Wesley, 2000).

[2] S. Card, J. Mackinlay, & B. Shneiderman, *Readings in information visualization: Using vision to think* (Morgan Kaufmann, San Francisco, 1999).

[3] T. Kohonen, *Self-Organizing Maps*, *third edition* (Springer, 2001).

[4] B. Back, K. Öström, K. Sere, & H. Vanharanta, Analyzing company performance using internet data, *Proc. 11th Meeting of the Euro Working Group on DSS*, Ed. by Zaraté, Toulouse, France, 2000, 52-56.

[5] A. Costea, & T. Eklund, A two-level approach to classifying countries/companies economic/financial performance, *Proc. 36th Annual Hawai'i International Conf. on System Sciences (HICSS),* 2003, Decision Technologies for Management track.

[6] T. Eklund, B. Back, H. Vanharanta, & A. Visa, Assessing the feasibility of Self-Organizing Maps for data mining financial information, *Proc. 10th European Conf. on Information Systems*, Gdańsk, Poland, 2002.

[7] E. Alhoniemi, Analysis of pulping data using the self-organizing map, *Tappi Journal, 83*(7), 2000, 66.

[8] S. Kaski, & T. Kohonen, Exploratory data analysis by the Self-Organizing Map: Structures of welfare and poverty in the world, *Proc. 3rd International Conf. on Neural Networks in the Capital Markets*, London, England, 11-13 October, 1995, Publisher World Scientific, 1996, 498-507.

[9] S. Kaski, J. Kangas, & T. Kohonen, Bibliography of Self-Organizing Map (SOM) Papers 1981-1997, *Neural Computing Surveys*, *1*(3&4), 1998, 1: 176.

[10] E. Oja, S. Kaski & T. Kohonen, Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum, *Neural Computing Surveys, 3*, 2003, 1-156.

[11] D. Marghescu, M. Rajanen, & B. Back, Evaluating the quality of use of visual data-mining tools, *Proc. 11th European Conference on IT Evaluation*, Amsterdam, 2004, 239-250.

[12] SOM_PAK http://www.cis.hut.fi/research/som_lvq_pak.shtml, 1990, last accessed June 2nd 2004.

[13] SOM Toolbox for Matlab, http://www.cis.hut.fi/projects/somtoolbox/, 1996, last accessed June 2nd 2004.

[14] Nenet http://koti.mbnet.fi/~phodju/nenet/Nenet/General.html, 1997, last accessed June 2nd 2004.

[15] S. Sharma, *Applied multivariate techniques*, (Jon Wiley and Sons, 1996).