XVII IMEKO World Congress
Metrology in the 3rd Millennium
June 22–27, 2003, Dubrovnik, Croatia

DATA MINING IN QUALITY ANALYSIS OF DIGITAL MOBILE TELECOMMUNICATIONS NETWORK

Pekko Vehviläinen[*], *Kimmo Hätönen*[**], *and Pekka Kumpulainen*[*]

*Measurement and Information Technology, Tampere University of Technology, Tampere, Finland
**Software Technology Laboratory, Nokia Research Center, Helsinki, Finland

Abstract – Quality analysis of a digital mobile telecommunications network can be a daunting task. For monitoring and operating purposes hundreds of measurements are constantly collected from a complex network. From the measurements the operating personnel of the network should be able to detect the changes that affect the service quality that the subscriber expects. Data mining methods can be used to ease the operating staff's task.

Keywords: data mining, quality measurement, telecommunications

1. INTRODUCTION

To answer the human needs of communication the development in telecommunications has led to cellular radio networks which enable the subscribers to communicate regardless of their location and move when connected. The first generation of such networks was analogue, but the benefits of digital transmission such as fewer errors in transmission and more efficient use of radio frequencies, have paved the way for digital mobile telecommunications networks. The most widely spread realizations of them are based on the Global System for Mobile communications standard (GSM) and are called second-generation networks (2G) [1]. The next generation, 3G network, is being built and it bases on the Universal Mobile Telecommunications System (UMTS) standard. UMTS network extends the user connectivity to a global scale and provides more bandwidth for the user, enabling multimedia transmissions.

Digital mobile telecommunication networks are highly complex systems and thus their planning, management and optimization are not trivial tasks. The complexity arises from the multitude of such elements as controllers, transceiver and receiver stations - to name just a few. Combined together, they form the base station subsystem of a network for mobile stations to connect. The basic units of the complex network are base transceiver stations, which have antennas pointed to give a radio coverage area, called a cell.

Subscribers of the network expect network availability, connection throughput and affordability. Moreover, their connection should not degrade or be lost abruptly as they move within the network area.

The user expectations constitute Quality of Service (QoS), which is specified as "the collective effect of service performances, which determine the degree of satisfaction of a user of a service" [2]. To gain a competitive edge to the other operators, the operating personnel have to measure the network in terms of QoS. By analysing the information in the measurements they can manage and improve the quality of service. Also, with the information they can optimize the parameters and the configuration of the network.

Sometimes it is not an easy task to find the essential information for QoS improvement from the complex data. At this point the recently developed data mining methods come of use.

2. QUALITY PERFORMANCE MEASUREMENTS

Several hundred measurements are recorded from the network [3]. However, it would be impractical for operating staff to monitor and analyze all the measurements. To reduce the amount of data the performance measurements are aggregated to Key Performance Indicators (KPI), which are further divided to capacity and quality related KPIs [4], [5].

The choice of which KPIs to monitor is up to operating staff. The data set selected for this paper has seven KPIs that are used for monitoring, two for the traffic channels, two for signal channels, two for handovers and one for dropped calls. The KPIs are from 30 cells of a GSM network and the measurements are 24 hour averages. The traffic KPIs relate to quality since calls and data are transmitted via traffic channels. Signaling channel is used for several purposes, the most important of them are data transmission (when a connection is established between a cell and user), and transmission of short messages. Handover is issued when there is a need to switch a connected user from a traffic channel to another. Typically this happens when a user moves from a cell to another. The average of dropped calls per hour is also recorded. Should any of these KPIs indicate problems, the users of the network are is likely to experience deterioration in QoS.

Personnel's expertise with KPIs and the problems occurring in the cells of the network varies widely, but at least they know what values of a KPI are desirable and what are unacceptable. Furthermore, the operators have their individual ways to judge if the KPIs observed from the cell

indicate that it is performing as desired and whether the state of the cell is normal. The decisions may be very simple rules such as "if any of the KPIs is unacceptable then the state of a cell is unacceptable." The rules are a portion of the *a priori* knowledge for the data. Four rules are used in this paper and they are summarized in the table below.

Table 1 Classification rules of cell's state, based on the *a priori* knowledge.

Test Statement	State of a Cell
if any of the KPIs is "UNACCEPTABLE"	"UNACCEPTABLE"
if any of the KPIs is "BAD"	"BAD"
if all the KPIs are either "NORMAL" or "GOOD"	"NORMAL"
if signal and traffic KPIs are "NORMAL" and the other KPIs are "GOOD"	"GOOD"

The rules can be easily programmed for computer and used for preclassification. The classification can be useful by its own right, but even more interesting is to find which combinations of KPIs occur most often, which of them are the most relevant in the analysis and which of the cells have the least problems and which ones have the most. This kind of information does not show explicitly from the preclassified data set. Also, the data set has to be discretized first to match the semantic values ("unacceptable", "bad", "normal" and "good") for the preclassification to work.

To sort out the needed information the basic statistics and simultaneous visualization of the KPIs is not enough and so more advanced methods are needed. In the past decade Knowledge Discovery in Databases (KDD) and data mining approaches have been introduced to find out the valuable information hidden in the large data sets. Thus KDD should be suitable for the telecommunications domain as well.

3. DATA MINING PROCESS AND KNOWLEDGE DISCOVERY

The purpose of a knowledge discovery process is to find new knowledge from an application domain [6]. The process consists of many separate, consecutive tasks of which data mining phase produces the patterns and information to be analyzed.

Knowledge discovery deals with the data in a data warehouse or database. Several descriptions of a KDD and data mining processes have been published; solid representations are for example in publications [7] and [8]. For the purposes of knowledge discovery in the mobile telecommunications we suggest five main phases for KDD and further five within data mining,

1. knowledge requirement
2. data selection
3. data mining
    a. data reduction
    b. data mining method selection
    c. data preprocessing
    d. data preparation
    e. use of data mining methods
4. result analysis
5. knowledge incorporation

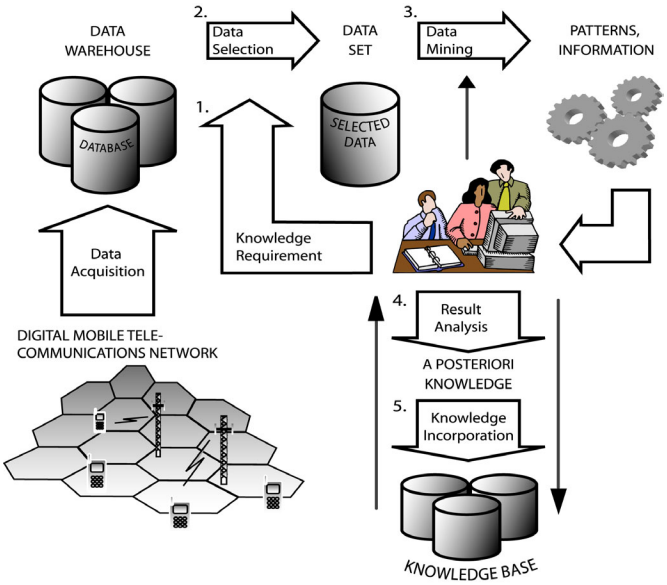

Figure 1 Knowledge discovery process.

The suggested processes are depicted in Figures 1 and 2, respectively.

The first step of knowledge discovery is to specify the knowledge requirement, that is, to define what it is that the analyst wants to find out from the application domain. To be able to specify the knowledge requirement the analyst has to have some *a priori* knowledge of the application domain. For our purposes the knowledge requirement is to find out problem related information from the data, more specifically, "what is the most common combination of KPIs when cell state is 'unacceptable', which KPIs indicate problems and what is the most common problem in the network?"

The second step is to select, possibly from several sources, the right data to support the knowledge requirement. Again, the skills and *a priori* knowledge of the analyst are needed. He should state what measurements, cells, and time periods are of interest. The necessary measurements are then fetched from a database, aggregated to KPIs and organized to a data table. For this paper, we were given a KPI data set from a real operator. The data set consists of 53 measurements, covering three months period of 30 cells totaling in 3069 observations.

The third step, data mining - a process by itself – is now ready to start from the reduction of the selected data set.

Result analysis and knowledge incorporation are considered steps of the KDD process and are not the scope of this paper.

3.a Data Reduction

The raw data from a base controller database are commonly aggregated to KPIs according to predetermined equations. The author of the equations can either be the manufacturer of the network management system or the network operating staff can make the KPIs by themselves.

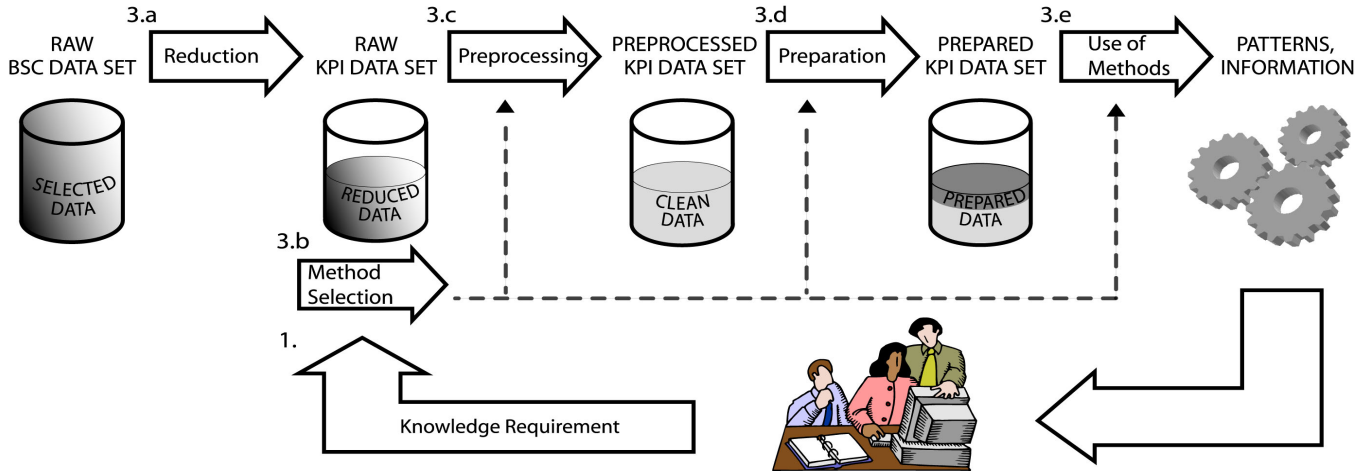

Figure 2 Data mining process.

We did the aggregation to our data and gained the seven KPIs described in section 2. The aggregation reduced the amount of the data over seven times as seven KPIs are used instead of the original 53 measurements.

Sometimes data reduction is regarded as a part of preprocessing; see section 3.c.

3.b Data Mining Method Selection

The selection of a analysis method is dependent on the knowledge requirement and the preference of the analyst. An important requirement was the method's ability to retain the original KPI name in the results. It was also desired that results should be in the same rule format as the *a priori* knowledge. Additionally, visual representations were needed to quickly address the analysis to the data structures. On these basis rough sets, Classification And Regression Trees (CART) and Self-Organizing Map (SOM) were selected. Rough sets and CART are capable of producing rules whereas both CART and SOM are used to visualize the structures in the data.

3.c Data Preprocessing

The purpose of data preprocessing is to ensure that the analysis methods are able to extract the required information from the data. The most important phase of preprocessing is feature extraction that is affected by the overall goals of the analysis, method selection and *a priori* knowledge. We view preprocessing as an iterative process, for more details see [9].

In feature extraction step, data integration, reduction, cleaning and transformation tasks can be recognized [8], but the tasks overlap somewhat.

For our data set, the integration was not needed as the data was acquired from a single source. The calculation of KPIs (section 3.a) is a data reduction step, but at this point it is possible for the analyst to further reduce the amount of data by omitting some of KPIs.

Data cleaning was done by omission of the measurements that had missing values. Visual inspection of the data was done to sort out the obvious outliers from the data.

Several transforming operations were done, first of those was the attachment of the preclassification result to the data set as a new variable. Two other transformations, discretization and normalization were needed because rough sets are only usable for discrete data and SOM is usually trained with normalized values.

The method specific transformations are regarded as the preparation of data.

3.d Data Preparation

Because rough sets are only usable for discrete data, the preprocessed data set had to be transformed accordingly. This is called discretization and it is used to transfer all the continuous values of the selected data set to just few, discrete values. The limits for semantic values ("unacceptable", "bad", "normal", "good") were given by an expert and they were further modified by visual inspection of the distributions. An example is given in Figure 3, where dropped calls are given discrete limits of 0,20, 0,65 and 2,0 to discretize the variable to the respective "good", "normal", "bad" and "unacceptable" zones.

SOM algorithm is based on the measurement of Euclidian distances between the measurements. Different scales of variables can greatly distort the results if the variables are supposed to have equal importance [10]. As this was the case for our data, it was decided to give equal importance to all the KPIs and each KPI was scaled to the range of [0 … 1,0].

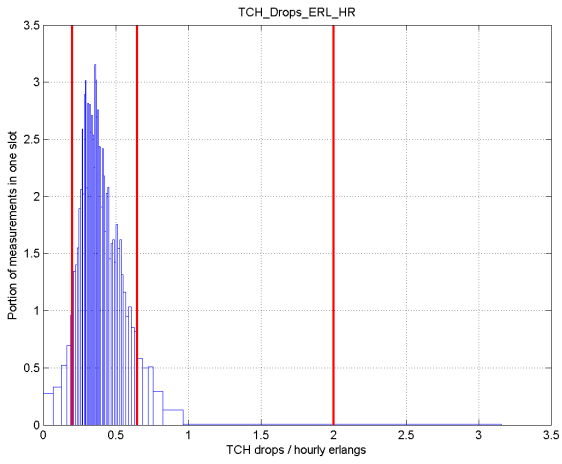

Figure 3 Discretization limits of the dropped calls KPI.

3.e Use of Data Mining Methods

After the preprocessing and preparation the data is ready to be processed with the selected methods.

4. METHODS

4.1. Rough Sets

The rough set theory was first introduced by Pawlak in 1982 [11]. The theory has since then had many applications in data analysis, such as finding relationships in imprecise data, removing redundancies from data and generating decision rules [12].

Only discrete data can be used with rough sets, so its discretization is necessary. The discretization can be done either manually or automatically. The discretization for our data is based both on the expert opinion and visual inspection of the KPIs distributions and was done manually. Also an automatic discretization process was developed for the data, see [9]. If decision rules are to be created with rough sets, preclassification of the data is necessary. The values of preclassification were added to the data as a decision variable. The observations that contained missing values were left out from the analysis, as rough sets are not suitable for such data.

Discretized rows of measurement are recognized by discernibility matrix [13]. The use of discernibility matrix reveals only those KPIs that are necessary to separate the measurement rows in the data matrix. These KPIs are the reduct of the original data set. Reducts were sought separately for each cell's measurement data, and it was found out that in most cases not all the KPIs were needed in explaining the variation within the data. Most notably, while at least one of the signaling channel, handover and dropped call related KPIs were present in every reduct, the traffic channel KPIs contributed less to them.

Besides of reducts, another application of rough sets use is the generation of decision rules from the data, initially suggested by Skowron in 1993 [12]. For our data, decision rules were made for all the cells separately and to the data set as a whole. The rules explain the variation by revealing the combinations of the KPIs in the data. For example, based on the rules, the most typical combination that causes an "unacceptable" state can be found from a cell. The revealed combinations in the format of decision rules are easy for the human operator to interpret, thus facilitating his work in finding the causes for problems in the network.

4.2. Classification Tree

Classification tree is a classification model gained by series of data splitting rules. A model structure is acquired by the automatic use of splitting criteria. That can be varied, the Gini index of diversity and twoing rule were used in the ground laying publication "Classification and Regression Trees (CART) [14]. The Gini index was used for our data because of computational simplicity.

The splits divide the data into subsets that can be represented in a tree-shaped structure, thus the name of the algorithm. From the structure it is possible for the analyst to review the properties of the data.

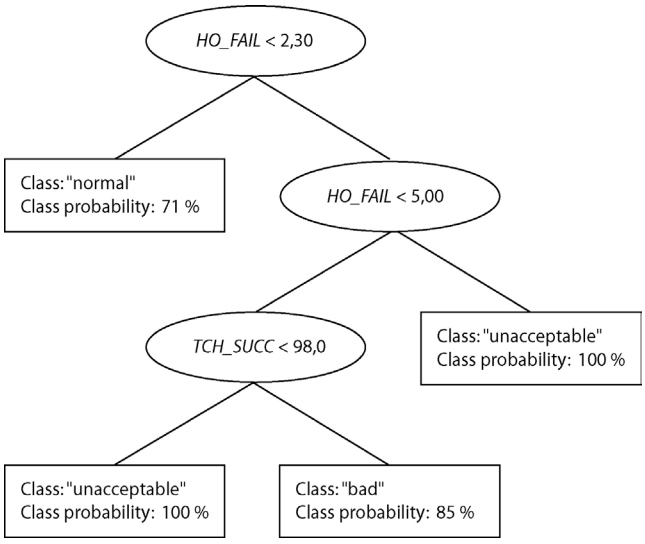

Figure 4 Simple classification tree of four levels.

When applied to KPI measurement data set, CART does not need the discretization of the measurements, but the algorithm needs a decision value to work. Although CART can handle missing values in the data, the observations with missing data were intentionally left out. That ensured the comparability of the CART model to rough sets and self-organizing map.

Figure 4 represents a tree after the CART algorithm (with Gini splitting rule) has been applied to the data set. The tree has been pruned to four levels. The rectangular shapes in the tree are called leaves, and are the subsets of the data. The oval shapes are called nodes and represent the splitting rules after the CART training. Subsets agreeing with the imposed split criteria fall on the left hand's side of the tree, disagreeing subsets on the right.

Note the tendency of Gini split criteria to find the purest possible subsets in terms of class probability (two subsets in the figure have 100% class probability). An interesting piece of knowledge for the operators could be the fact that failures in handovers indicate also other problems in the cells. The first found split suggests that if more than 2,30% of handovers from a cell to another fail, there are other problems as well. If, on the other hand, the failure percentage is better than 2,30, the probability of a normal state is 71 %. For more accurate classification the pruning level can be altered.

In accordance to the rough sets, the tree model can also be represented as a set of decision rules.

4.2. Self-Organizing Map

Self-Organizing Map (SOM), first introduced by Kohonen in 1982, is a neural network model that organizes the data set being analyzed to a low dimensional grid. The most typical use of SOM is a two-dimensional map of neurons; one or three dimensional SOM mappings are rarer. The data is mapped onto this map by using the SOM algorithm, for more details see [15].

The SOM algorithm reveals the structure of each KPI of the data set by mapping them onto component planes. However, the clustering structure of the whole data set is not

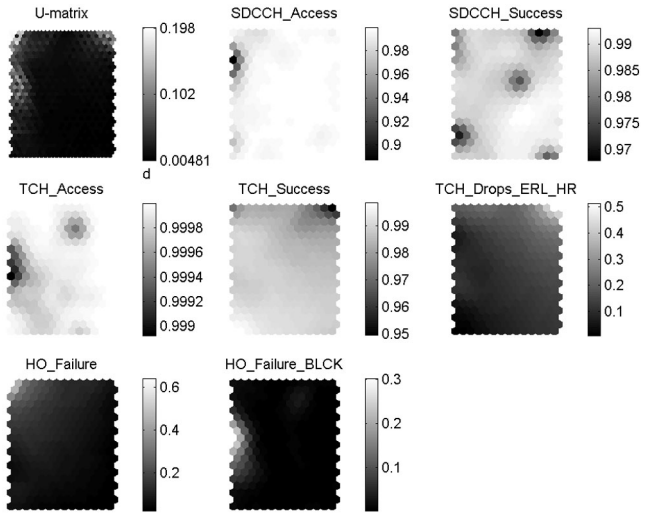

Figure 5 Visualization of SOM with the U-matrix and seven component planes of KPIs.

clearly visible from them. A unified distance matrix (U-matrix) was developed for clustering purposes, and it visualizes the geometrically correct vector distributions of the SOM [16].

Figure 5 shows a SOM trained from the data set. On the top left corner of the figure is the U-matrix; followed by the component planes of the seven KPIs. From left to right are signaling channel access and success measurements, and then traffic channel access and success. The rightmost component plane in the middle is the average of dropped calls per hour in a cell. The bottom row has component planes for failed handovers KPI and failed handovers due to blocking KPI.

The picture shows significant connection between dropped calls and traffic channel success (top right corner of the corresponding component planes) and handover failures and traffic channel success (top left corner of the corresponding component planes). Also notable is the small variation of the traffic channel access and signaling channel success KPIs.

In the U-matrix, where the dark color is dominant, distances between the neurons are shown. Dark tone means that the neurons are close to each other in those areas, whereas brighter tones mean greater distances between neurons. The variation in the data is shown on the both top corners and also on the left side of the U-matrix. Comparison of the U-matrix to the component planes suggests that the traffic channel and handover related KPIs best explain the distances.

The operating staff can visualize the data structure with the SOM and use it to find dependencies between the KPIs. Also, it could be useful, especially from the optimization point of view, if clusters based on the U-matrix are found from the data.

5. CONCLUSIONS

From a complex digital network it is possible to find essential quality related information if data mining process is applied properly. The prerequisite is that the selection of data, pre-processing, the use of *a priori* information and choice of data mining methods are understood equally important steps of a knowledge discovery process. If the preliminary steps for the data mining methods fail the results of the KDD process can be misleading.

The knowledge requirement should be known before data mining process can be applied. The requirement affects the choice of the data set and the selection of the data mining method. The validity of the results gained from the use of method is dependent on the preprocessing, so the choice of method should also take the preprocessing phase into account.

If the knowledge requirement is not satisfied after the analysis, the KDD process should continue, starting either from the preprocessing, selection of the method, selection of the data or even the redefinition of the knowledge requirement.

The methods for this paper were chosen to help the operating staff to find the essential information from the network quality performance measurements. The personnel may not be experienced with the theory of the methods, rather they should be able to interpret the results themselves.

Rough sets and CART are usable for the analyst as the found information can be expressed as acquired rules with the preserved KPI names. CART and SOM also provide effective visual means for interpreting the data set. This type of information can be understood with a little effort from the operator, and the acquired knowledge can be stored for further use.

Further research concentrates on the preprocessing; the aim is to ensure that the methods produce robust and understandable results. Also the role of the metadata and information required in the phases of the process is an interesting topic of further study. While the research concentrates on the methods presented in this paper, other methods such as fuzzy clustering could be useful for this type of data.

The data set used in this paper is measured from a GSM network. The methods that we have used can undoubtedly be used for the UMTS network as well, because similar KPIs are used for 3G network measurements. The real operator data is not yet available in quantity, so the initial tests will be based on simulated data.

REFERENCES

[1] GSM Association. (n.d.). *GSM World - the Website of the GSM Association.* Available from: - http://www.gsmworld.com/index.shtml [Accessed 28th Nov. 2002].

[2] ETSI, TS 100 615 V8.0.0. (2001-02). *GSM Technical Specification 12.04: Digital Cellular Telecommunications System (Phase 2+); Performance Data Measurements.*

[3] NOKIA INC. (1999). *Nokia NMS/2000, Database Description for BSC Measurements. Reference Guide.*

[4] SUUTARINEN, J. (1994). *Performance Measurements of GSM Base Station System*. Thesis (Lic.Tech.) Tampere University of Technology.

[5] LEMPILÄINEN, J. MANNINEN, M. (2001). *Radio Interface System Planning for GSM/GPRS/UMTS*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

[6] KLÖSGEN, W., ZYTKOW, J. (1996). Knowledge Discovery in Databases Terminology. *In*: FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R. eds. *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: The MIT Press, 573-592.

[7] FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. (1996). From Data Mining to Knowledge Discovery. An Overview. *In*: FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R. eds. *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: The MIT Press, 1-34.

[8] HAN, J., KAMBER, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers.

[9] KUMPULAINEN, P., HÄTÖNEN, K., VEHVILÄINEN P. (2003). Automatic Discretization in Preprocessing for Data Analysis in Mobile Network. *In press: XVII IMEKO World Congress Metrology in the 3rd Millennium, Cavtat-Dubrovnik, Croatia, June 22-27, 2003*

[10] VESANTO, J., HIMBERG, J., ALHONIEMI, E., PARHANKANGAS, J. (2000). *SOM Toolbox for Matlab 5*. Helsinki, Finland: Helsinki University of Technology.

[11] PAWLAK, Z. (1982). *Rough Sets.* International Journal of Computer and Information Sciences, **11** 5, 341-356.

[12] CIOS, K., PEDRYCZ, W., SWINIARSKI, R. (1999). *Data Mining Methods for Knowledge Discovery*. 2nd edition. Upper Saddle River, NJ: Prentice Hall Inc.

[13] SKOWRON, A., STEPANIUK, J. (1995). Decision Rules based on Discernibility Matrices and Decision Matrices. *In:* T.Y. LIN, ed. *Conference Proceedings (RSSC'94) The Third International Workshop on Rough Sets and Soft Computing.* San Jose, CA: San Jose State University, 602-609.

[14] BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J. (1984). *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC Press LLC.

[15] KOHONEN, T. (1995). *Self-Organizing Maps.* Berlin: Springer-Verlag.

[16] ULTSCH, A., SIEMON, H.P. (1990). Kohonen's self organizing feature maps for exploratory data analysis. *In Proceedings of International Neural Network Conference, INNC'90*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 305-308.

Authors:

Pekko Vehviläinen, Measurement and Information Technology, Tampere University of Technology, P.O. Box 692, FIN-33101, Tampere, Finland, tel. +388-3-365-3572, fax +358-3-2171, pekko.vehvilainen@tut.fi.

Kimmo Hätönen, Software Technology Laboratory, Nokia Research Center, P.O. Box 407, FIN-00045 Nokia Group, Finland. Tel. +358-50-483-7278, kimmo.hatonen@nokia.com

Pekka Kumpulainen, Measurement and Information Technology, Tampere University of Technology, P.O. Box 692, FIN-33101, Tampere, Finland, tel. +388-3-365-2458, fax +358-3-2171, pekka.kumpulainan@tut.fi.