Chapter 17

Data-driven analysis of telecommunication systems

Olli Simula, Kimmo Raivio, Jaakko Hollmén, Kimmo Hätönen, Sampsa Laine, Pasi Lehtimäki, Timo Similä

17.1 Introduction

Large amounts of data are generated by the telecommunications systems in order to log events for later analysis in problem cases or to charge for the services used. We have applied our earlier experience in studying the industrial processes in the telecommunications context: the data-driven approach may be applied equally well in this context as in the analysis of industrial processes.

17.2 Analysis of mobile radio access network

The 3G cellular systems will offer services beyond the capabilities of today's networks. The variety of services requires some modifications on the radio network planning and optimization process. One of the modifications is related to the quality of service (QoS) requirements and control. For each service the QoS targets have to be set and naturally also met. With configuration parameters optimum operation point for each cell should be found.

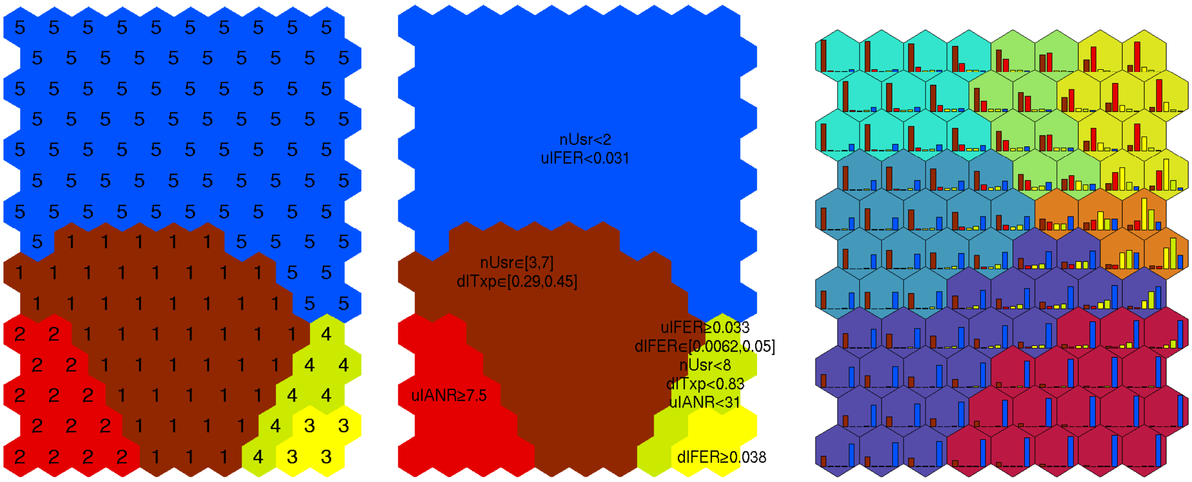
In this project, wideband code division multiple access (WCDMA) mobile network has been analyzed using the Self-Organizing Map [3] [2]. The goal is to develop efficient adaptive methods for monitoring the network behavior and performance. Special interest is on finding clusters of mobile cells, which can be configured using similar parameters. The data has been generated using a WCDMA radio network simulator [1].

Figure 17.1: Model used in classification of mobile cells.

The method utilizes the SOM algorithm both in feature extraction and clustering the cell features. A block diagram of the method is shown in Figure 17.1.

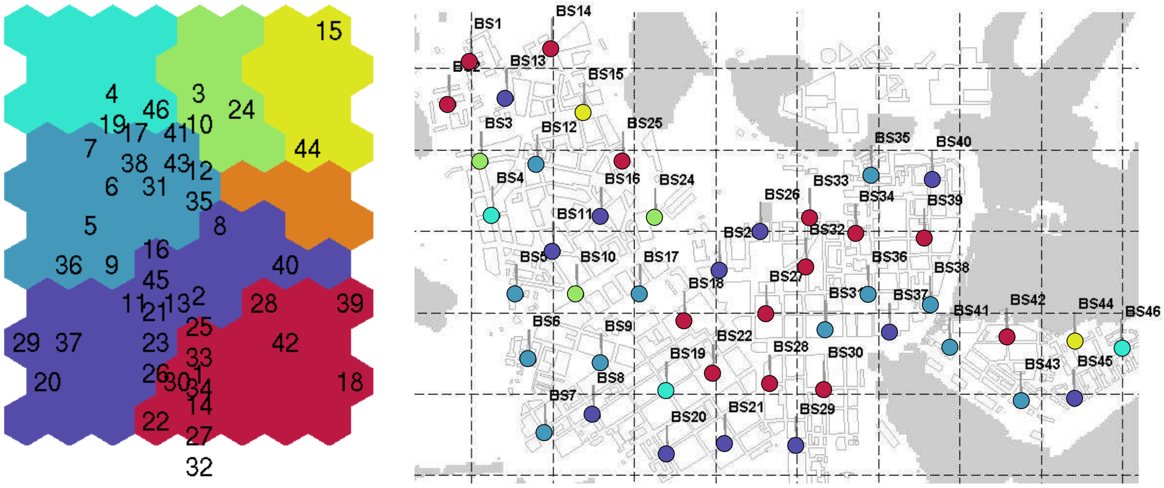
The data vectors of all the cells are clustered using a two phase clustering algorithm [4]. In that algorithm, a Self-Organizing Map is trained using the data vectors. The codebook vectors of the SOM are clustered using K-means or some hierarchical clustering method with a validity index so that exact clusters can be defined. The input data vectors are classified using labeled SOM codebook vectors.

Next, some results of the network analysis are presented. Five variables are used in the analysis and two of them describe the QoS level in the network. Clusters of SOM codebook and some automatically generated rules of the clusters are shown in Figure 17.2a and 17.2b.



(a) (b) (c)

Figure 17.2: Clusters of the map of data vectors (a), rules of clusters (b) and clusters of histogram map (c).



(a) (b)

Figure 17.3: BMUs of mobile cells on histogram map (a) and locations of classified cells (b).

For each mobile cell a histogram is computed. The histogram describes how the data from one cell fall into the data clusters. These histograms are used as profiles in cell classification. The clusters of histogram SOM are shown in Figure 17.2c and the best-matching units (BMU) of the mobile cells in Figure 17.3a. The classified cells and their locations are presented in Figure 17.3b.

In this method, two level clustering procedure has been used because of high data rate. In 3G systems, the sampling rate can be from ten to hundred samples per second. Similar methods can be used also in lower sampling rate systems like GSM, but then it is enough to process the data by using only one clustering level.

References

[1] S. Hämäläinen, H. Holma, and K. Sipilä. Advanced WCDMA radio network simulator. In *Personal, Indoor and Mobile Radio Communications*, volume 2, pages 951–955, Osaka, Japan, September 12-15 1999.

[2] P. Lehtimäki, K. Raivio, and O. Simula. Mobile radio access network monitoring using the self-organizing map. In *Proceedings of European Symposium on Artificial Neural Networks*, Bruges, Belgium, April 24 - 26 2002. (accepted).

[3] K. Raivio, O. Simula, and J. Laiho. Neural analysis of mobile radio access network. In *IEEE International Conference on Data Mining*, pages 457–464, San Jose, California, USA, November 29 - December 2 2001.

[4] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

17.3 The use of preprocessing and visualisation techniques to analyze telecommunications data

The GSM and third generation cellular phone networks produce masses of data. As the networks develop fast, the operating personnel can not keep up-to-date with the new conditions of the system. They have problems to evaluate the data: to select the most important variables and cells for study, and to assess the state of the network. This study searches for the preprocessing and visualization techniques that support inexperienced process operators. The display designed to given an overview to the GSM-network is presented in Figure 17.4. The nodes of the graph are the GSM-transmitters; the arcs indicate cells. The color of the node indicates the status of each GSM-cell.

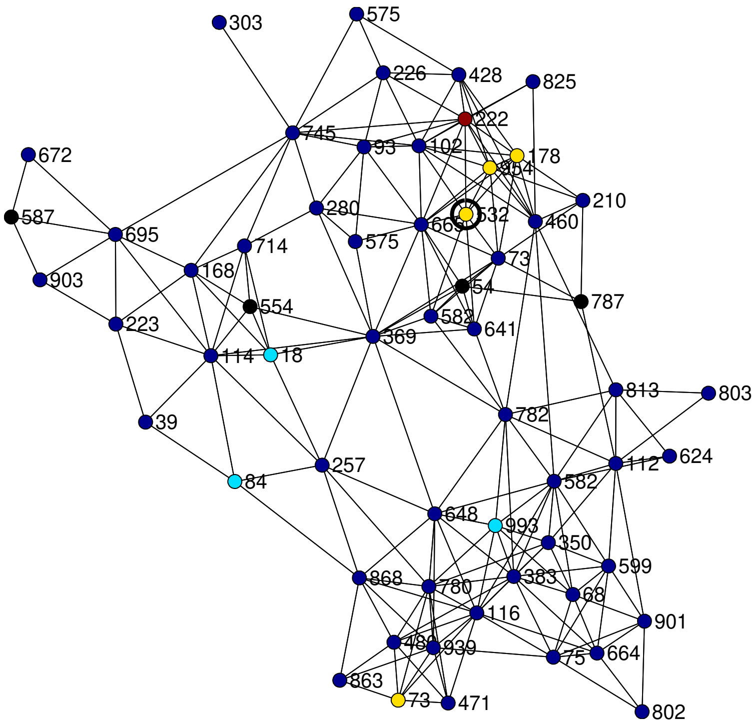


Figure 17.4: A graph depicting the status of a GSM-network.

The colors are a result of a data analysis process. The process is based on the proper value range of each variable defined by an expert into fuzzy membership functions. These membership functions are used to pretreat the data: to scale the variables and to remove outliers. The resulting data is used to train a SOM, which is clustered [1] to find the typical states of the GSM-cells. The clusters are allocated a color known to the user. For example, red color can be used to indicate problems.

The approach has been created together with a senior expert of the GSM-networks in the Nokia company, Mikko Toivonen. The research is based on the funding of the Suomen Akatemia and the Nokia Foundation.

References

[1] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

17.4 Fraud detection in mobile communications networks

Telecommunications fraud, the illegitimate use of services, presents a financial burden to the network operators, since the service charges from the services used by fraudsters can not be collected. In fraud detection, the goal is to discover the presence of illegitimate calling activity of telecommunications customers. What makes calling activity illegitimate in the above sense is that the subscriber's intention at the time of calling is not to pay for the services used. The intentions can not be observed directly, but they are reflected in the calling behavior. The calling behavior is collectively described by the call data, which in turn can be observed. Therefore, it is reasonable to use call data as a basis for subsequent analysis, both in formulation of models through learning and in detection of fraudulent behavior.

Whereas the call data exemplifies individual patterns of normal or fraudulent behavior, the goal is to formulate a *model* so that it would essentially articulate the same thing as data but on a more abstract and general level. The approach taken in this work is to *learn* a representation of normal and fraudulent behavior that can be used in decision making. For this end, probabilistic models and neural networks have been applied [1]. These models handle uncertainty in the domain in a favorable fashion and they can process the abundance of data present in the telecommunication domain. In this work, we have applied dynamical modeling of behavior in fraud detection. This has proved to be a promising line in fraud detection, and in user profiling and classification in general, since behavior at the most natural level is defined in a temporal context.

User profiling and classification for fraud detection is reported in [1] (see also the section on doctorate theses).

References

[1] J. Hollmén. *User Profiling and Classification for Fraud Detection in Mobile Communications Networks*. PhD thesis, Helsinki University of Technology, 2000.