ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Hierarchical analysis of GSM network performance data

Mikko Multanen, Kimmo Raivio and Pasi Lehtimäki

Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400, FI-02015 HUT - Finland

Abstract. In this study, a method for hierarchical examination and visualization of GSM data using the Self-Organizing Map (SOM) is described. The data is examined in few phases. At first temporally averaged data is used and then, in each phase some of the data is discarded and the rest is examined in more detail. The SOM is used both in clustering and in visualization. The actual clustering is performed to the nodes of the SOM to lower the computational cost and to help to understand better the properties of the clusters.

1 Introduction

The purpose of this project was to develop a method to explore the data of an entire GSM network. This method is an extension to the methods developed by the group for the analysis of mobile radio access networks [1] [2] in which only small geographical areas of the network were used. The main problem in the current project was on the one hand to reduce the amount of data so that the results would not be overwhelming to go through and on the other hand to give an extensive picture of the performance data.

To achieve this goal a hierarchical method was developed where the sensitivity of the examination is increased gradually. In every phase, a proportion of the data samples are discarded so that the rest can be examined in more detail. The data samples are clustered at each phase and one cluster is chosen to the next phase. The whole process builds a tree-like structure where similar data samples are in the same branch of the tree.

The main tool used in the project was the Self-Organizing Map (SOM) [3]. The SOM was used because of its data reduction abilities and mainly because it is highly visual. The SOM is a good method when high dimensional data has to be represented. It maps multi-dimensional data to a two-dimensional grid which can easily be visualized. The visualization of the data was one of the main goals of this project. Also a method for ordering the clusters was developed so that the interestingness between the clusters can be compared more easily.

2 Network and data

A simplified description of the architecture of a GSM (Global System for Mobile communications) network is shown in Fig. 1. The subscriber carries the Mobile Station (MS). The Base Transceiver Station (BTS) contains radio transceivers

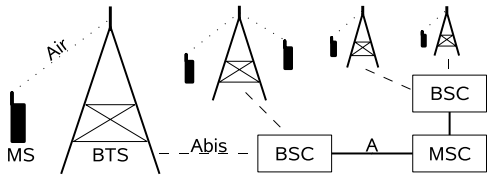

Fig. 1: The architecture of a GSM network.

and handles the radio-link protocols. The Base Station Controller (BSC) manages the radio resources for one or more BTSs. BTSs and BSCs together control the Air interface to the MS and they communicate across the Abis interface. The network is connected to other telephone networks through the Mobile service Switching Center (MSC). The BSC and the MSC communicate across the A interface [4].

The data had been collected from a real operative network. It was stored as counters of different events in the network elements. The counters were stored every hour and then reseted for the next cycle. There were thousands of counters in the database and only about hundred of them were used in this project.

A Key Performance Indicator (KPI) is calculated from one or more counters. The KPIs were defined by the network manufacturer. The KPIs were used because they are much more intuitive than the plain counters. It is much easier to understand what one KPI represents than what the plain counters from which the KPI is calculated represent. On the other hand some information is also lost with KPIs because one KPI value contains the information of many counters.

The most critical KPIs were TCH (Traffic CHannel) Drop Rate, which indicates abnormal service interruptions, and TCH Blocking, which indicates service denials. These KPIs have a direct effect to the quality experienced by the subscribers [5].

3 Methodology

3.1 Two-level clustering

The Self-Organizing Map (SOM) [3] consists of a regular low dimensional grid, in this case two dimensional. The units of the grid are represented by prototype vectors whose dimension is the same as the dimension of the data vectors. The prototype vectors form an elastic net. When the SOM is trained the net settles so that the prototype vectors represent the training data as well as possible without losing the net structure. Because of the net structure it is straightforward to visualize the SOM in two dimensions.

The clustering of the data vectors is done in two stages [6]. At first a SOM is trained with normalized data. The prototype vectors also known as the nodes of the SOM can be seen as the clusters of the first level. Then the nodes are clustered. This is the second level of the clustering. The number of the prototype

vectors is much smaller than the number of the data points. So by using the SOM the data for the actual clustering is reduced considerably.

The clustering of the SOM is made with a hierarchical agglomerative clustering algorithm. At first each node is assigned to its own cluster. Then distances between all the clusters are calculated and two nearest ones are merged. The distance d between two clusters C_k and C_l is calculated by using the average linkage:

$$d(C_k, C_l) = \frac{\sum_{i,j} \|x_i - x_j\|}{N_k N_l}, \tag{1}$$

where $x_i \in C_k$, $x_j \in C_l$, $k \neq l$ and N_k is the number of data points in the cluster C_k. The merging of the clusters is continued until there is only one cluster left which contains all the nodes. These operations build a clustering tree (dendrogram). Several different clusterings of the data can be made by cutting the tree from different levels. In this project, the number of the clusters was fixed to the square-root of the number of the data vectors used in training of the SOM.

This two-level approach decreases the computational cost of the clustering compared to the clustering of the data outright. If there were N data vectors which should be clustered using this two-level clustering only M nodes of the SOM have to be clustered where M is much smaller than N. However the data can be clustered only to M clusters at the most and also the SOM must be trained. The overall computational cost is smaller with a two-level clustering than with a straightforward clustering of the data samples.

3.2 Interest values

The clusters of the SOM can be compared by calculating interest values for them. At first interest values are calculated for the individual data vectors. The interest value of one cluster is got by taking the average of the interest values of the data vectors which belong to the cluster. KPI values are mapped nonlinearly to interest values. Expert knowledge is used in the construction of the mapping. A simple sigmoid function can be used. The interest value is near one when the KPI gets interesting values and otherwise near zero.

The following sigmoid function is used in the calculation of the interest value of a cluster for a KPI:

$$\mathtt{interest}_{i,k} = \sum_{x \in C_i} \frac{\mathtt{scale}_k}{1 + e^{-\mathtt{steepness}_k (x_k - \mathtt{threshold}_k)}} \, w(x_k), \tag{2}$$

where x is a data vector, i the index of the cluster and k the index of the KPI. C_i is the set of the data vectors of the cluster i. $w(x_k)$ is a traffic weight value which is calculated separately for SDCCH, TCH and HO related KPIs. The more traffic there is the more interesting it is. The weight value is calculated using the following formula:

$$w(x_k) = a \frac{t_k}{m_{t_k}} + (1 - a), \tag{3}$$

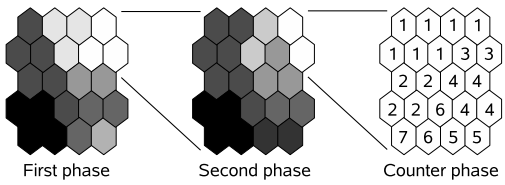

Fig. 2: A simple view of the system.

where t_k is the traffic related to the KPI k, m_{t_k} is the mean of the traffic values in the data set and a is a parameter which defines how much traffic has influence. The relative influence of different KPIs to the interest value is defined by `scale` parameter. The accepted operational area of a KPI is defined with `steepness` and `threshold` parameters. For example, `threshold` can be set to the border of the unacceptable behavior. If `steepness` is positive values bigger than `threshold` are unacceptable and if `steepness` is negative values smaller than `threshold` are unacceptable. Expert knowledge was used in setting these parameters. The overall interest value of a cluster is calculated by summing the interest values of the cluster over all the KPIs.

4 Data analysis

4.1 First and second phase

The following procedure is used in finding the most interesting parts from the data of the entire network. In the first phase the feature vector contained the averages of the selected KPIs over the whole two month measurement time for every BTS. A SOM was trained with this data and the nodes of the SOM were clustered as described above in Section 3.1. After the clustering interest values were calculated for the clusters using the method described in Section 3.2. One cluster was selected and the BTSs which hit the selected cluster were chosen to the next phase.

In the second phase, the feature vector contained the averages of the KPIs over one day for every BTSs which were selected in the first phase. Different days were now separate feature vectors. A new SOM was trained and clustered. Interest values were calculated for the clusters and one cluster was selected and the days and BTSs which hit the cluster were chosen to the next phase. Also one KPI of the used KPIs was chosen so that in the counter phase there was only one KPI and the most interesting days and BTSs to examine.

An overview of the system is shown in Fig. 2. In the first and the second phase, the SOM is colored with the help of the interest values of the clusters so that it is easier to see which parts of the map are more interesting. The number of data vectors is kept roughly the same in the phases because the number of temporally averaged data vectors is decreased by selecting only one cluster and before the SOM of the next phase the temporal precision of the residual data is

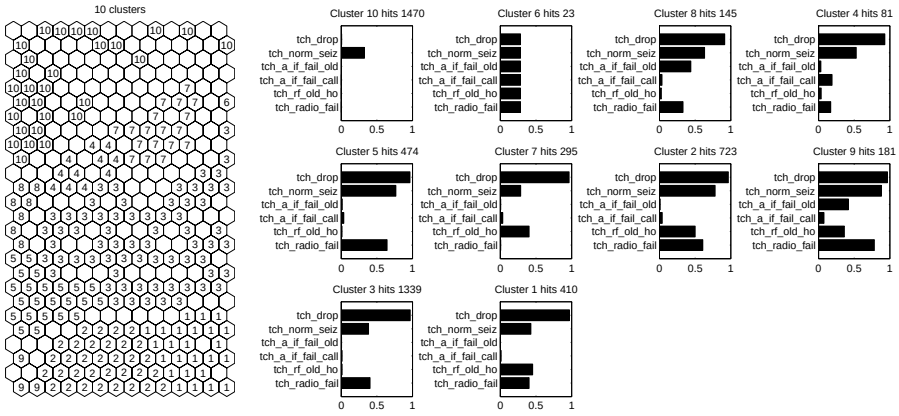

Fig. 3: On the left side is the clustering of the TCH Drop Rate SOM of the counter phase and on the right side are the normalized mean counter values of the clusters. These can be used in the comparison between the clusters. Diagrams are sorted according to the value of the TCH Drop Rate KPI.

increased. There could have been more phases, if for example the measurement time had been longer, but to this data these two were sufficient.

4.2 Counter phase

Now the feature vector was created from the counters of the chosen KPI of the second phase cluster. The hours of the chosen days for the selected BTSs were investigated as separate vectors. A SOM was trained and clustered. The value of the KPI depends on different counters so the clusters are distinguished by the fact that different counters might be dominant in different clusters. After this phase the most interesting data can be chosen to a more case-specific examination.

The clustering of the counter phase SOM and the bar graphs of the normalized means of the counters in the clusters are shown in Fig. 3. TCH Drop Rate KPI was chosen to the counter phase because it got the highest interest value in the second phase. In the first and the second phase, the most interesting clusters were chosen. This explains why the TCH Drop Rate KPI gets high values all over the SOM, except in the clusters 10 and 6, in the counter phase. Although the TCH Drop Rate is near constant the values of the counters vary. In this case the counters `tch norm seiz`, `tch a if fail old`, `tch a if fail call`, `tch rf old ho` and `tch radio fail` had the biggest increasing effect to the average value of the TCH Drop Rate KPI on the map. Counters whose influence was much smaller than these were dropped out from the visualization.

5 Conclusions

In this project, a hierarchical method to examine the data of an entire mobile network has been presented. The data is examined hierarchically in steps. KPI data is used in the first phases because it is much more intuitive than raw counter data. On every phase, data is clustered and one cluster is selected. Interest values are calculated for the clusters to help the selection. The amount of the data is reduced in every phase so that it can be examined in more detail in the next phase. This method creates a tree like structure where the data vectors of one branch are similar with each other. It gives a quick representation of the data of the network.

The developed method takes advantage of the expert knowledge of the network. The KPI formulas are defined by experts and the parameters of the interest function must be defined before the method can be used. Also the KPIs which will be used in the method have to be chosen. However, the method gives some feedback of the chosen KPIs, how interesting they are according to the interest function.

Hierarchical examination proved to be a workable method. It reduces effectively the amount of data which have to be examined manually. In this project, the rough estimate of the data was created by taking average which might hide some of the interesting parts of the data. So, a more intelligent method should be developed. Perhaps some kind of weighted average which uses the interest function would be a better one. Also the KPI interest calculation function could be replaced with a more advanced one. Although GSM data was used in the project the method described in this paper can be used as well for the analysis of 3G radio access networks.

References

[1] J. Laiho, K. Raivio, P. Lehtimäki, K. Hätönen, and O. Simula. Advanced analysis methods for 3G cellular networks. *IEEE Transactions on Wireless Communications*, 4(3):930–942, May 2005.

[2] P. Lehtimäki and K. Raivio. A knowledge-based model for analyzing GSM network performance. In A. F. Famili, J. N. Kok, J. M. Peña, A. Siebes, and A. J. Feelders, editors, *IDA*, volume 3646 of *Lecture Notes in Computer Science*, pages 204–215. Springer, 2005.

[3] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[4] J. Scourias. Overview of the global system for mobile communications. Technical report, Department of Computer Science, University of Waterloo, 1996.

[5] S. A. Kyriazakos and G. T. Karetsos. *Practical radio resource management in wireless systems*. Artech House, Norwood, 2004.

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