A SOM Based Approach for Visualization of GSM Network Performance Data

Pasi Lehtimäki, Kimmo Raivio

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

Abstract. In this paper, a neural network based approach to visualize performance data of a GSM network is presented. The proposed approach consists of several steps. First, a suitable proportion of measurement data is selected. Then, the selected set of multi-dimensional data is projected into two-dimensional space for visualization purposes with a neural network algorithm called Self-Organizing Map (SOM). Then, the data is clustered and additional visualizations for each data cluster are provided in order to infer the presence of various failure types, their sources and times of occurrence. We apply the proposed approach in the analysis of degradations in signaling and traffic channel capacity of a GSM network.

Keywords. data mining, neural networks, visualization, self-organizing map, telecommunications.

1 Introduction

The radio resource management in current wireless communication networks concentrates on maximizing the number of users for which the quality of service (QoS) requirements are satisfied, while gaining the maximal profitability for the operator [12]. In practice, the goal is to obtain an efficient usage of the radio resources (i.e. maximal coverage and capacity with the given frequency spectrum) while keeping the infrastructure costs at the minimum. Currently, the variety of services is developing from voice-oriented services towards data-oriented services, causing new difficulties for the network resource management due to the increased diversity of QoS requirements.

The most severe performance degradations of wireless networks from the user point of view involve the reduced availability (blocking) of the services as well as the abnormal interruption of the already initiated services (dropping). In principle, such performance degradations may result from unpredictable hardware breakdowns or temporary changes in the operating environment (i.e in traffic flow), but on the other hand, they may originate from incorrect (or unsuitable) network configuration, causing bad performance more regularly.

The system knowledge required to optimize GSM system performance is very difficult to formalize as a mathematical model and therefore, automatic control of many configuration parameters is unfeasible. Instead, the network optimization is carried out by application domain experts having a long experience in the problem field. In such a case, it is more efficient to exploit the existing expert knowledge and to try to represent the most informative portion of the measurement data in an efficient form in order to support performance optimization.

In this paper, an analysis process based on Self-Organizing Map (SOM) to visualize GSM network performance data is presented. The SOM has been applied in the analysis of 3G network performance, including advanced network monitoring and cell grouping [7,6].

Next, the basic SOM algorithm is presented. Then, the overall SOM based analysis process for GSM performance data is outlined. Then, we demonstrate the use of the analysis process in two problem scenarios in which the capacity problems in the signaling and traffic channels are analyzed.

2 Methods

2.1 Self-Organizing Map

One of the most widely used neural network algorithms is the Kohonen's Self-Organizing Map [5]. It consists of neurons or map units, each having a location in a continuous multi-dimensional measurement space as well as in a discrete two-dimensional output grid. During the so-called training phase, a multi-dimensional data collection is repeatedly presented to the SOM until a topology preserving mapping from the multi-dimensional measurement space into the two-dimensional output space is obtained. This dimensionality reduction property of the SOM makes it especially suitable for data visualization.

The training phase of SOM consist of two steps: the winner map unit search, followed by application of an update rule for the map unit locations in the measurement space. In winner search, an input sample $\mathbf{x}$ is picked up randomly from the measurement space and the map unit c closest to the input sample $\mathbf{x}$ is declared as the winner map unit or the best-matching map unit (BMU):

$$c = \arg\min_i ||\mathbf{x} - \mathbf{m}_i||, \tag{1}$$

in which $\mathbf{m}_i$ is the location of the ith map unit in the measurement space and c is the index of the winner map unit in the output grid of SOM.

After the winner search, the locations of the map units in the measurement space are updated according to the rule:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t)h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)], \tag{2}$$

in which $0 < \alpha(t) < 1$ is a learning rate factor and $h_{ci}(t)$ is usually the Gaussian neighborhood function

$$h_{ci}(t) = \exp\left(-\frac{||\mathbf{r}_c - \mathbf{r}_i||}{2\sigma^2(t)}\right), \tag{3}$$

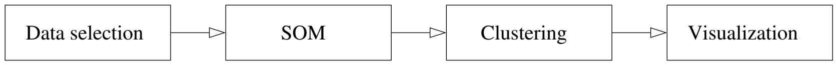


Fig. 1. A block diagram illustrating the phases of the proposed analysis process.

where $\mathbf{r}_c$ is the location of the winner unit and $\mathbf{r}_i$ is the location of the ith map unit in the discrete output grid of SOM. The learning rate factor $\alpha(t)$ and the neighborhood radius $\sigma(t)$ are monotonically decreasing functions of time t.

2.2 The overall analysis process

The proposed SOM based analysis process is illustrated in Figure 1. Next, the steps of the analysis process are discussed in detail.

Data selection The GSM system consists of large amount of base stations (BSs), each serving the users on distinct geographical areas (cells). The performance of the BSs is described by a large amount of variables called Key Performance Indicators (KPIs) with typical sampling frequency of one hour. For each KPI, an objective value can be defined by the network operator in order to define the acceptable performance of the network.

When projection methods such as the SOM are used in data visualization, all samples usually have equal priority when determining the projection (dimensionality reduction) function. In many trouble shooting tasks, however, more accurate visualizations of failures would be more appropriate at the expense of samples representing normal operation. When analyzing the performance degradations of a GSM network, the data subset to be used in projection function determination can be selected by choosing the KPIs of interest and removing the samples that represent normal operation (the objective values for the selected KPIs are met). For example, if an accurate visualization of traffic channel problems are desired, it would be justified to use only the samples in which traffic channel blocking or traffic channel drop rate exceed some pre-selected threshold.

SOM training After the subset of data of interest is selected, the data is normalized in order to make all variables equally important independently on the measurement unit. Then, the normalized data is used as the input data in the SOM training. The training procedure for the SOM was described in Section 2.1. The trained SOM is used to visualize the multi-dimensional input data using the component plane representation of SOM [10].

Clustering The clustering of the data aims in partitioning the data into "natural" groups, each (hopefully) describing different types of failures present in the GSM network. Therefore, the clustering of the data allows the analysis process to be divided into subproblems in which different types of failures are analyzed separately. We have adopted a clustering approach in which the clustering process is carried out for the map units of SOM (in the measurement space) instead of the original data subset [11]. We have used the k-means clustering algorithm [2] for different values of k (the number of clusters in which the data is divided). The best clustering among different values of k is selected according to the Davies-Bouldin index [1].

Visualization After the SOM training and clustering, a visualization of the selected multi-dimensional input data is obtained. This information helps the application domain expert to make inferences about the possible problem scenarios present in the data. The cluster analysis based on SOM component planes reveals the variety of failures faced by the network. It is relatively easy task for an expert to select the most important variables (KPIs) for each failure type. By analyzing the amount of samples in different fault clusters originating from each cell of the GSM network, the locations of the different failure types are efficiently obtained. Finally, the visualization of the times of occurrence of different fault types reveals additional temporal information about the faults. These three types of simple visualizations allows the selection of variables, cells and time periods that are taken into further analysis using conventional methods.

3 Experiments

3.1 Analysis of SDCCH capacity problems

In this section, we demonstrate the use of the presented analysis process by analyzing the capacity problems in the signaling channel. The available data set consists of several KPIs with sampling frequency of one hour. The measurements were made in 41 cells during 10-week time period, resulting in about 40 000 multi-dimensional data vectors. First, we selected a suitable data selection scheme in order to focus on the signaling channel capacity problems. The selected variable set consisted of SDCCH blocking and availability rates, uplink and downlink signal strengths and signal quality measurements, as well as the amount of circuit switched traffic in the cell.

We applied an inequality constraint with SDCCH Blocking > 0 % in order to filter the uninformative (normal operation) samples from the analysis. Then, we applied histogram equalization based normalization method for the selected data set in order to obtain invariance w.r.t the scales of the variables. Then, we trained a SOM in which the map units were organized in a 15×10 hexagonal grid by applying 500 epochs of batch training and 500 epochs of sequential training.

For comparison purposes, we used Principal Component Analysis (PCA) [3] and Independent Component Analysis (ICA) [4] methods to obtain alternative coordinate axes in the original data space along which the data samples were to be projected. We compared the quality of the projections using the measures of trustworthiness and preservation of neighborhoods [9]. We found out that the SOM and PCA based projections were equally good, outperforming the ICA based projection in both measures. We evaluated the same accuracy measures

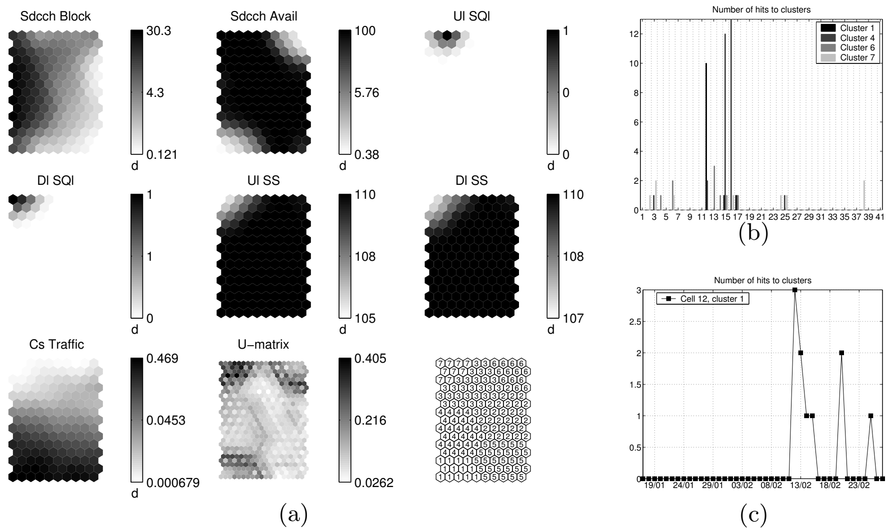


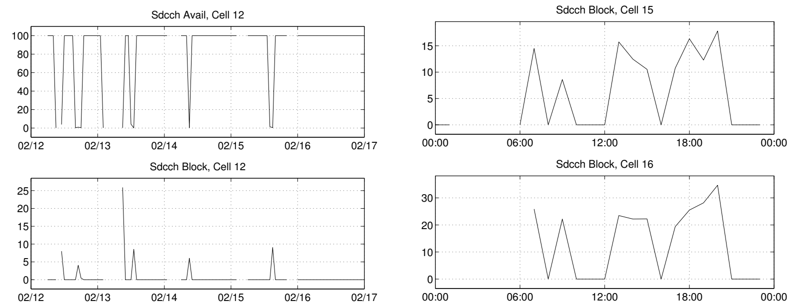
Fig. 2. (a) SOM of data samples representing possible signaling channel capacity problems. Clusters 1, 4, 6 and 7 represent possible signaling channel capacity problems. (b) Cells 12, 15 and 16 contribute the most to the fault clusters. (c) In cell 12, the failures appear mostly during a 4-day period.

for the same data subset in cases where the data selection had no impact on actual projection function (i.e all the data was used in forming the projection). We found out, that the SOM and ICA based projections lost in representation accuracy when data selection was not used. The PCA based projection performed equally well in both cases (with and without data selection).

In Figure 2(a), the so-called component planes of the SOM are shown. In order to visualize the cluster structure of the data, we clustered the map units of SOM using k-means clustering algorithm and plotted the resulted clustering with the U-matrix representation [8] of SOM. The numbers in the map units of SOM indicate the cluster memberships of the map units.

By analyzing the properties of each cluster using the component planes, four clusters that represent possible signaling channel capacity problems can be identified: cluster 4 contains high values for signaling channel blocking, with moderate amount of traffic. Clusters 1 and 6 represent behavior in which a drop in channel availability is likely to cause the high blocking values. Cluster 7 represents channel blockings that are likely to be a result of bad signal quality, i.e the connection is refused because the required channel quality could not be provided. The U-matrix reveals, that the clusters 1, 6 and 7 are located further apart from the other clusters.

By analyzing the number of hits into different fault clusters (see Figure 2(b)), it was evident that nearly all of the samples in the fault clusters were generated by only three cells of the network. Hits from other cells can be viewed instan-



(a) (b)

Fig. 3. (a) In cell 12, peaks in signaling channel blocking appear during a drop in channel availability. (b) The blocking rates of the cells 15 and 16 are very correlating.

taneous situations that do not give reasons to configuration adjustments and therefore can be ignored.

When plotting the times of occurrences of the hits to the fault clusters from these three cells, it was found that the cell 12 had a 4-day period when most of the samples into fault cluster 1 were generated (see Figure 2(c)). This suggests that the signaling channel availability were temporarily reduced (i.e the amount of available channels dropped) and therefore, some of the channel requests were blocked. In order to verify this assumption, we plotted the signaling channel availability and blocking from that time period (see Figure 3(a)). According to this figure, it is even more clear that it is the drops in availability that causes the requests to be blocked.

Most of the samples of cluster 4 were generated by cells 15 and 16 (in the same site), suggesting that they suffer from signaling channel blocking at high amounts of users. In addition, these samples were generated mostly during one day. The signaling channel blockings of these cells from that day are shown in Figure 3(b). Clearly, the blocking rates of the two cells are strongly correlating. Such behavior can be due to a failure in a close-by cell, causing requests to originate from larger geographical area than normally. Therefore, the amount of channel requests is abnormally high. On the other hand, such increase in signaling channel traffic may also be caused by a configuration error leading to increased amount of location updates or paging traffic. It should be noted that the amounts of signaling traffic capacity problems in this network are relatively small (only less or equal to 10 hits per cell into any of the problem clusters).

3.2 Analysis of TCH capacity problems

In this experiment, we repeated the same analysis procedure for traffic channel data. The selected variables consisted of TCH blocking, dropping and availability rates, uplink and downling signal strengths as well as the amount of circuit

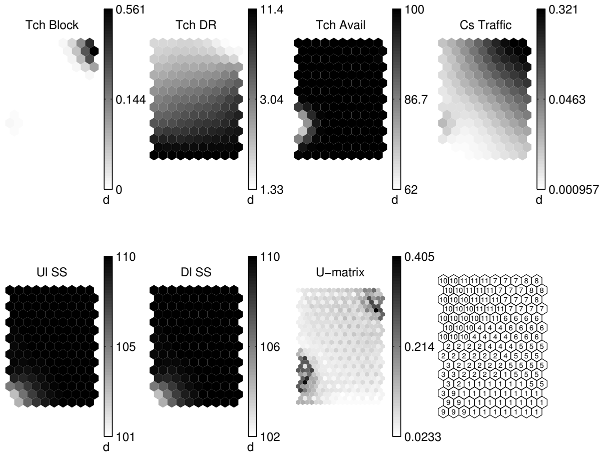


Fig. 4. The SOM of traffic channel capacity problems and the corresponding clusters on top of SOM.

switched data traffic. In addition, we applied the inequality constraint requiring that TCH Blocking > 0 % or TCH Drop Rate > 2 %.

Then, a SOM was trained using the normalized data subset as the input data. The training procedure of SOM was similar to the one with the signaling channel data. Also, we trained SOMs with bigger map grids, but they did not provide any clear benefits over the map of size 15×10. When comparing the SOM based projection with the PCA and ICA projections, we found out that all the projections were equally good. The SOM based projection provided the worst values of trustworthiness measure with small neighborhoods, but the best values with large neighborhoods. Also, the ICA based projection gave worse values of preservation of neighborhoods as the size of the neighborhood increased. The importance of data selection in forming the projection function was not as clear as in the signaling channel capacity analysis. This is due to the fact that in the signaling channel capacity analysis, the data selection retained only 0.4 % of the samples, and in the traffic channel capacity analysis, the used data selection scheme retained up to 27 % of the samples.

In Figure 4, the SOM of the traffic channel capacity problems is shown with the corresponding clusters. From the figure, several fault clusters can be identified: cluster 1 represents samples with relatively high drop rate and low amount of traffic. Cluster 3 represents moderate amount of traffic channel drops and degraded traffic channel availability. In cluster 8, blocking appears with relatively high amount of traffic. Cluster 9 contains samples with high drop rate, low uplink and downlink signal strengths, and low amount of traffic.

Similarly to the signaling channel capacity analysis, the contributions of the cells into these clusters were analyzed. In the analysis, it was found that cells 39 and 9 generated many samples into cluster 3, stating that they suffer from

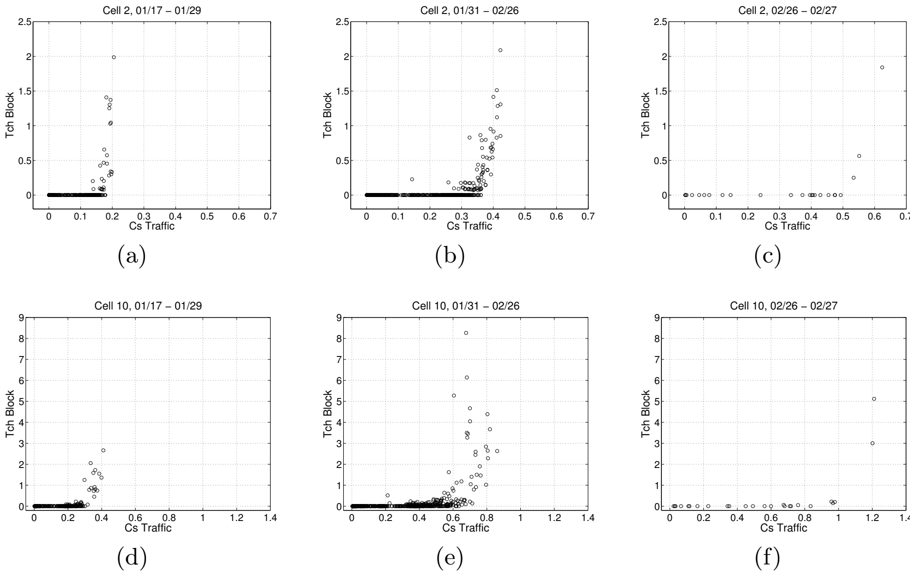


Fig. 5. The amount of traffic channel blocking at different amounts of traffic in cells 2 and 10. In these cells, the capacity is increased twice during the analyzed period.

call dropping due to low availability. By plotting the drop rate and channel availability as time series for both cells (not shown), it became clear that the drop rates in these cells were in the same level also when the availability was full. Therefore, the call dropping is not likely to be caused by the reduced channel availability. However, it is interesting that these problems appear simultaneously in these cells (see Figure 6(a)) and that they were located on nearly overlapping coverage areas.

Cells 2 and 10 generated most of the samples in cluster 8, i.e they seem to suffer from blocking at high amounts of traffic. The amount of resources (frequencies) may not be appropriate for these cells. They did not suffer from signal quality or channel availability problems, and therefore, the blockings are exclusively due to insufficient resources. Figure 5 shows the amount of blocking vs. the amount of traffic of these cells in three different time periods. In Figures 5(a) and (d), the first time period is shown. It is clear that the blockings start to increase when the amount of traffic in cell 2 is more than 0.15 Erlangs and in cell 10, more than 0.25 Erlangs. However, during the next time period shown in Figures 5(b) and (e), the blockings start to appear when the amount of traffic exceeds 0.3 Erlangs in cell 2 and 0.5 Erlangs in cell 10. It seems likely that the amounts of resources were increased between these two periods. However, blocking still appears. In Figures 5(c) and (f), the third time period is shown. This time period lasts only two days, but the same behavior seems to continue: blocking starts to appear when the amount of traffic exceeds 0.5 Erlangs in cell 2 and 1.0 Erlangs in cell 10.

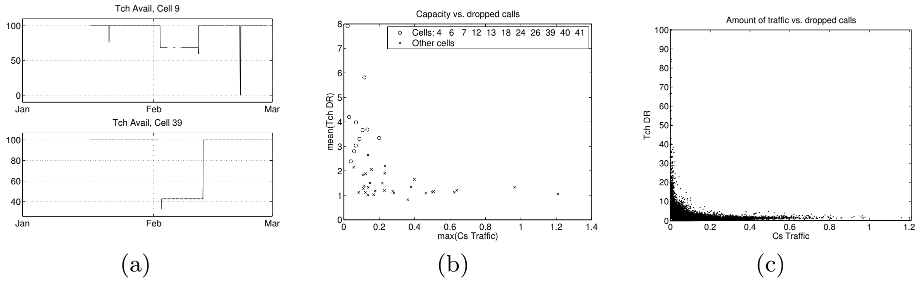


Fig. 6. (a) The drop in traffic channel availability in cells 9 and 39 was not explaining the inadequate drop rate values. Instead, the drop rate levels were high constantly. Interestingly, the channel availabilities dropped simultaneously in both cells. (b) The cells with lowest capacity tend to suffer from higher drop rates. (c) The highest drop rates occur at very low amount of traffic. This behavior is common to all cells.

Cells 6 and 26 generated most of the samples in cluster 9, indicating that they suffer from call dropping when the received signal strengths were low and the amount of traffic was low. Further analysis revealed that the low signal strengths did not explain the amount of dropping, since the variation in signal strengths did not cause variation in drop rates. Also, if low signal strength would have caused traffic channel dropping, power control should have increased the power levels in order to place it at appropriate level. Instead, these cells behave similarly to cells 4, 7, 12, 13, 18, 24, 26, 39, 40 and 41 that generated a lot of samples into cluster 1. Cluster 1 is essentially similar to the cluster 9, except the higher signal strength levels.

The key observation here is that all these cells suffer from high drop rates at low amount of traffic. There are at least two reasons that might explain this behavior. First, the so-called microcells characterized by low power levels, low capacity and small coverage areas are frequently positioned to cover the busiest locations such as city areas. Often, such areas also represent the most difficult propagation environments and the highest user mobility, causing serious variations in radio channel conditions. In Figure 6(b), the capacity of the cells (measured as maximum amount of traffic observed in the data) is plotted against the average drop rate. From the figure it is clear, that highest average drop rates are observed in the cells that also have the smallest capacity. As mentioned before, such small capacity cells are frequently used in hot-spot areas where user mobility is high and propagation conditions are difficult. Therefore, the call dropping in these cells are probably caused by physical constraints of the environment and it might be difficult to achieve better performance in such cells.

Secondly, it is evident that the highest drop rate values are observed when the amount of traffic is close to zero (see Figure 6(c)). When the amount of calls is very low, a drop of only a few connections may cause very high drop rates. This is due to the fact that the formula used to generate the traffic channel drop rate from several low-level counters exaggerates the seriousness of the fault at low number of calls during the measurement period. The inaccuracy of the traffic channel drop rate causes similar behavior in all cells, but is obviously more frequently present in the cells with higher drop rates at all amounts of traffic.

It can be concluded, that the traffic channel problems are much more regular than the signaling channel problems. The traffic channel problem clusters are hit about 50 - 500 times and the signaling channel problem clusters were hit at most 13 times. The most typical problem type was traffic channel dropping in low capacity cells.

4 Conclusion

In this paper, a SOM based approach to visualize GSM network performance data was presented. This visualization process allowed us to efficiently locate the problem cells and find the times of occurrences of problems, and to select the appropriate variables in order to continue the analysis by conventional analysis methods. By visualizing all the possible variable pairs over the whole time period from all the cells would have produced a very high number of graphs, making the manual analysis of such results unfeasible. Therefore, the use of the proposed analysis process helped to achieve a higher degree of efficiency in the analysis of multi-dimensional GSM network performance data.

References

1. D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, April 1979.
2. Brian Everitt. *Cluster Analysis.* Arnold, 1993.
3. Simon Haykin. *Neural Networks: a comprehensive foundation, 2nd edition.* Prentice-Hall, Inc., 1999.
4. A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley and Sons., 2001.
5. Teuvo Kohonen. *Self-Organizing Maps, 3rd edition.* Springer, 2001.
6. Jaana Laiho, Kimmo Raivio, Pasi Lehtimäki, Kimmo Hätönen, and Olli Simula. Advanced analysis methods for 3G cellular networks. *IEEE Transactions on Wireless Communications (accepted)*, 2004.
7. Jaana Laiho, Achim Wacker, and Tomáš Novosad, editors. *Radio Network Planning and Optimisation for UMTS.* John Wiley & Sons, Ltd, 2002.
8. A. Ultsch and H. P. Siemon. Kohonen's self-organizing feature maps for exploratory data analysis. In *Proceedings of the International Neural Network Conference (INNC 90)*, 1990.
9. Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: an experimental study. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pages 485–491, 2001.
10. Juha Vesanto. Som-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126, 1999.
11. Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.
12. Jens Zander. *Radio Resource Management for Wireless Networks.* Artech House, Inc., 2001.