Chapter 16

Intelligent data engineering

Miki Sirola, Kimmo Raivio, Pasi Lehtimäki, Jukka Parviainen, Jaakko Talonen, Golan Lampi, Teemu Poikela, Eimontas Augilius, Olli Simula

16.1 Data analysis in industrial operator support

Miki Sirola, Jukka Parviainen, Jaakko Talonen, Golan Lampi, Teemu Poikela, Eimontas Augilius

Early fault detection with data-analysis tools in nuclear power plants was one of the main goals in NoTeS2-project (test case 4) in TEKES technology program MASI. The industrial partner in this project was Teollisuuden Voima Oy, Olkiluoto nuclear power plant. Data analysis was carried out with real failure data, training simulator data and design based data, such as data from isolation valve experiments. A control room tool, visualization tools and various visualizations were developed.

Fault dynamics and dependencies of power plant elements and variables was inspected to open the way for modelling and creating useful statistics to detect process faults. In our research we succeeded to use data mining to learn from industrial processes and find out dependencies between variables by Principal Component Analysis (PCA) and Self-Organizing Map (SOM). Also a segmentation method was developed to detect automatically different process states of stored datasets.

An adaptive model was developed to primary circulation system to detect leakage in steam lines. A fault was defined as an unpermitted deviation of the variable. Also K-means clustering in time was used for monitoring and detecting pre-stage of process fault [1].

When fault or its pre-stage is detected, current process state should be diagnosed and operators should be informed efficiently. Process monitoring was improved by concepts of generated control limits and alarm balance. All these fault detection and diagnosis methods were programmed with Matlab. Data Management Tool (DMT) is an interface for off-line analysis of stored Olkiluoto datasets including preprocessing, variable selection and other developed methods, see Figure 16.1.

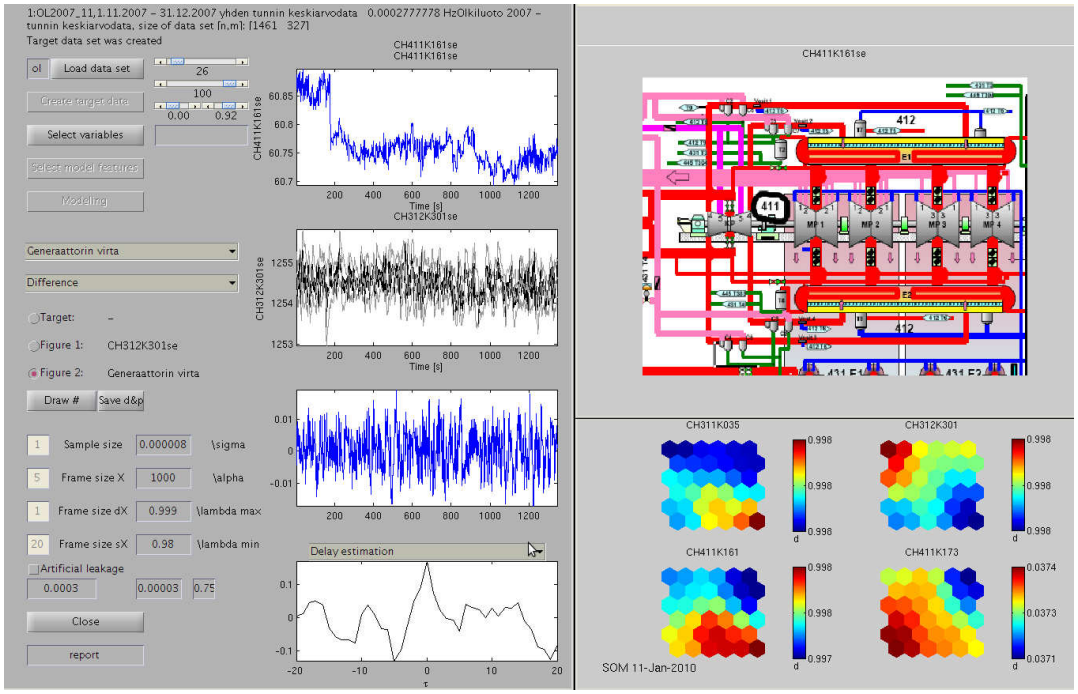

Figure 16.1: DMT User Interface.

Decision support prototype DERSI for failure management in nuclear power plants was developed [2]. It is a control room tool for operator or analysis tool for expert user. It combines neural methods and knowledge-based methods. DERSI utilizes Self-Organizing Map (SOM) method and gives advice by rule-based reasoning. The operator is provided by various informative decision support visualizations, such as SOM maps for normal data and failure data, state U-matrix, quantization error for both component level and state U-matrix, time-series curves and progress visualizations. DERSI tool has been tested in fault detection and separation of simulated data.

All visualizations developed in the project are collected for making a first proposal for wide monitoring screens in [3]. Industrial applications by using SOM are presented in [4]. Also a seminar course on this topic was held in the autumn term 2009.

References

[1] J. Talonen and M. Sirola. Abnormal Process State Detection by Cluster Center Point Monitoring in BWR Nuclear Power Plant. In *Proceedings of the International Conference on Data Mining (DMIN)*, volume I, II, pages 247–252, July 2009.

[2] G. Lampi. Self-organizing maps in decision support: a decision support system prototype. Master's thesis, Helsinki University of Technology, 2009.

[3] M. Sirola, J. Parviainen, J. Talonen, G. Lampi, T. Alhonnoro, and R. Hakala. Early fault detection with SOM based methods and visualizations - new contents for wide monitoring screens. *EHPG-Meeting of OECD Halden Reactor Project*, May 2008. Loen, Norway. 11p.

[4] M. Sirola, J. Talonen, and G. Lampi. SOM based methods in early fault detection of nuclear industry. In *Proceedings of the 17th European Symposium On Artificial Neural Networks ESANN'09*, April 2009.

16.2 Cellular network optimization

Kimmo Raivio, Pasi Lehtimäki

Structure of mobile networks gets more and more complicated when new network technologies are added to the current ones. Thus, advanced analysis and tuning methods are needed to optimize the performance of the network. Adaptive methods can be utilized, for example, to detect anomalous behavior of network elements [3] and to adjust configuration parameters of the network [1].

In order to automate the configuration parameter optimization, a computational method to evaluate the performance of alternative configurations must be available. In data-rich environments like cellular networks, such predictive models are most efficiently obtained with the use of past data records.

In blocking prediction, the interest is to compute the number of blocked requests at different conditions. This can be based on the use of well known Erlang-B formula. The expected value for the number of blocked requests is obtained by multiplying the number of arriving requests with the blocking probability, leading to $B = \lambda p(N_c|\lambda, \mu, N_c)$. The expected value for the congestion time is $C = p(N_c|\lambda, \mu, N_c)$ and the expected value for the number of channels in use is $M = \sum_{n=0}^{N_c} np(n|\lambda, \mu, N_c)$.

In [1], it was shown that the Erlang-B formula does not provide accurate predictions for blocking in GSM networks if low sampling rate measurements of arrival process are used in the model. More traditional regression methods can be used for the same purpose with the assist of knowledge engineering approach in which Erlang-B formula and regression methods are combined. With the use of Erlang-B formula, the dependencies between B, C and M that remain the same in each base station system need not be estimated from data alone. The data can be used to estimate other relevant and additional parameters that are required in prediction. In this research, a method to use Erlang-B formula and measurement data to predict blocking has been developed. The regression techniques are used to estimate the arrival rate distribution describing the arrival process during short time periods. The Erlang-B formula is used to compute the amount of blocking during the short time periods.

Suppose that the time period is divided into N_s segments of equal length. Also, assume that we have a vector $\boldsymbol{\lambda} = [0 \;\; 1\Delta_\lambda \;\; 2\Delta_\lambda \;\; \ldots \;\; (N_\lambda - 1)\Delta_\lambda]$ of N_λ possible arrival rates per segment with discretization step Δ_λ. Let us denote the number of blocked requests during a segment with arrival rate λ_i with $B_i = \lambda_i p(N_c|\lambda_i, \mu, N_c)$, where $p(N_c|\lambda_i, \mu, N_c)$ is the blocking probability given by the Erlang distribution. Also, the congestion time and the average number of busy channels during a segment with arrival rate λ_i are denoted with $C_i = p(N_c|\lambda_i, \mu, N_c)$ and $M_i = \sum_{n=0}^{N_c} np(n|\lambda_i, \mu, N_c)$. In other words, the segment-wise values for blocked requests, congestion time and average number of busy channels are based on the Erlang-B formula.

Now, assume that the number of segments with arrival rate λ_i is θ_i and $\sum_i \theta_i = N_s$. Then, the cumulative values over one hour for the number of requests T, blocked requests B, congestion time C and average number of busy channels M can be computed with

$$\begin{bmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_{N_\lambda} \\ B_1 & B_2 & \ldots & B_{N_\lambda} \\ \frac{C_1}{N_s} & \frac{C_2}{N_s} & \ldots & \frac{C_{N_\lambda}}{N_s} \\ \frac{M_1}{N_s} & \frac{M_2}{N_s} & \ldots & \frac{M_{N_\lambda}}{N_s} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{N_\lambda} \end{bmatrix} = \begin{bmatrix} T \\ B \\ C \\ M \end{bmatrix} \tag{16.1}$$

or in matrix notation $\mathbf{X}\boldsymbol{\theta} = \mathbf{Y}$.

Now, the problem is that the vector $\boldsymbol{\theta}$ is unknown and it must be estimated from the data using the observations of $\mathbf{Y}$ and matrix $\mathbf{X}$ which are known a priori. Since the output vector $\mathbf{Y}$ includes variables that are measured in different scales, it is necessary to include weighting of variables into the cost function. By selecting variable weights according to their variances estimated from the data, the quadratic programming problem

$$\min_{\boldsymbol{\theta}} \quad \left\{ \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} + \mathbf{f}^T \boldsymbol{\theta} \right\} \tag{16.2}$$

$$w.r.t \quad 0 \leq \theta_i \leq N_s, \ i = 1, 2, ..., N_\lambda, \tag{16.3}$$

$$\sum_{i=1}^{N_\lambda} \theta_i = N_s \tag{16.4}$$

is obtained where $\mathbf{f} = -\mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{Y}$ and $\mathbf{H} = \mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{X}$ include the weighting matrix $\mathbf{W}$. In other words, the goal is to find the vector $\boldsymbol{\theta}$ that provides the smallest prediction errors for variables T, B, C and M.

The optimization problem could be solved for each of the N_d observation vectors separately, leading to N_d solution vectors $\boldsymbol{\theta}$ for hour h. Since we are interested in long-term prediction of blocking, we should somehow combine the solution vectors so that behavior common to all solution vectors are retained and non-regular properties of the demand are given less attention.

Using probabilistic models solutions of different arrival rates can be combined. At first the total number of arrived requests is estimated from probabilities of observing a segment with certain arrival rate. The same model can be used to map segment-wise blocking candidates to the total number of occurrences of blocked requests during one period. Similarly, the cumulative values for the average number of busy channels and the congestion time can be computed [2].

References

[1] P. Lehtimäki and K. Raivio. Combining measurement data and Erlang-B formula for blocking prediction in GSM networks. In *Proceedings of The 10th Scandinavian Conference on Artificial Intelligence (SCAI)*, Stockholm, Sweden, May 26 - 28 2008.

[2] Pasi Lehtimäki. *Data Analysis Methods for Cellular Network Performance Optimization.* Doctoral dissertation, TKK Dissertations in Information and Computer Science TKK-ICS-D1, Helsinki University of Technology, Faculty of Information and Natural Sciences, Department of Information and Computer Science, Espoo, Finland, April 2008.

[3] M. Multanen, K. Raivio, and P. Lehtimäki. Outlier detection in cellular network data exploration. In *Proceedings of the 3rd International Workshop on Performance Analysis and Enhancement of Wireless Networks (PAEWN)*, Okinawa, Japan, March 25 - 28 2008.