

## Chapter 11

# Intelligent data engineering

Olli Simula, Jaakko Hollmen, Sampsa Laine, Kimmo Raivio, Miki Sirola, Pasi Lehtimäki, Timo Similä, Mika Sulkava, Jarkko Tikka, Jukka Parviainen, Teppo Marin, Golan Lampi, Mikko Multanen, Risto Hakala, Petri Saarikko

### 11.1 A knowledge-based model for analysis of GSM network performance

The performance of the mobile network is measured based on thousands of counters, describing the numbers of the most important events over a measurement period (typically one hour). In order to allow more efficient performance monitoring, a set of high-level key performance indicators (KPIs) are derived from the counter data. Such indicators are traditionally used in resource management [1] and they are well suited for performance monitoring [4], but there are several drawbacks when they are used in fault diagnosis [2]. One of the main problems is that most of the widely used performance indicators describe the operation of the network at the BTS or BSC level. As a result, the performance degradations originating from interaction between several BTSs become very difficult to observe. In many cases, however, the operation of the close-by BTSs are highly dependent on each other. In this project [3], we aimed to avoid the above mentioned KPI-related problems by using a novel performance analysis approach based on counter data. Due to the significant increase in number of variables, a knowledge-based model was used to divide the analysis process into a set of small system identification problems in order to keep the overall analysis process tractable.

In order to determine the knowledge-based model, the available measurements were divided into variable sets describing the performance of the different subsystems of the GSM network. Then, a simple mathematical input-output model for each of the subsystems were proposed. In Figure 11.1(a), the subsystem hierarchy for the overall performance model is shown. The model parameters were estimated from the available data record using quadratic programming. Then, the parameter estimates were used to find the input-output variable pairs involved in the most severe performance degradations. Finally, the resulting variable pairs were visualized as a tree-shaped cause-effect chain in order to allow user friendly analysis of the network performance (see Figure 11.1(b)). The provided information can be used to enhance the current radio resource usage in the network.

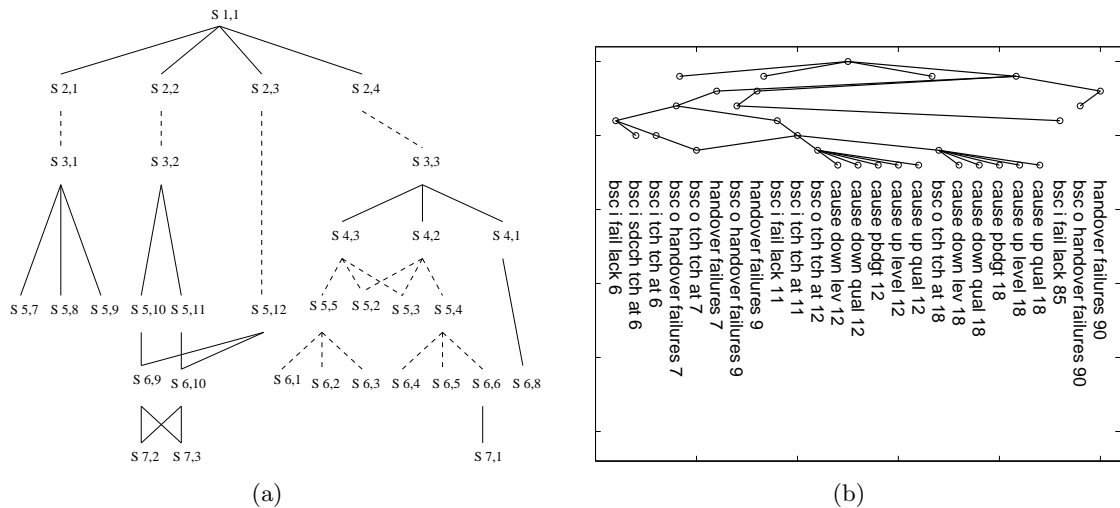


Figure 11.1: (a) The subsystem hierarchy. (b) The dependency tree describing the cause-effect chains of the handover failures.

## References

- [1] Sofoklis A. Kyriazakos and George T. Karetsos. Practical Radio Resource Management in Wireless Systems. Artech House, Inc., 2004.
- [2] P. Lehtimäki and K. Raivio A SOM Based Approach for Visualization of GSM Network Performance Data. *Proceedings of the 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, pages 588 – 598, Bari, Italy, June 22–24, 2005.
- [3] P. Lehtimäki and K. Raivio A Knowledge-Based Model for Analyzing GSM Network Performance. *Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA)*, pages 205 – 215, Madrid, Spain, September 8–10, 2005.
- [4] 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.

## 11.2 Analysis of mobile radio access network

Radio access networks produce a huge amount of data. In this project, the Self-Organizing Map has been used to analyse mobile data [1][2] from GSM and 3G networks. Earlier analysed 3G network data was generated using a radio network simulator. Now, both 3G and GSM data have been collected from real network. The goal is to develop efficient adaptive methods for monitoring the network behavior and performance. Special interest is on fault detection and on finding clusters of mobile cells or mobile cell pairs. Cells of one cluster can be configured using similar parameters.

The method utilizes the SOM algorithm several times when clustering mobile cells or cell pairs. At first, the Self-Organizing Map is used together with some clustering algorithm to cluster feature vectors of single cells or cell pairs. The clustering can be performed repeatedly to cluster mobile cells using more complex features or to zoom stepwise into some part of the data. In the latter method, at each round the most interesting subset of the data is selected for further analysis. This two phase clustering algorithm [3] begins with training a SOM with the data vectors. The codebook vectors of the SOM are clustered using K-means or some hierarchical clustering method with a validity index.

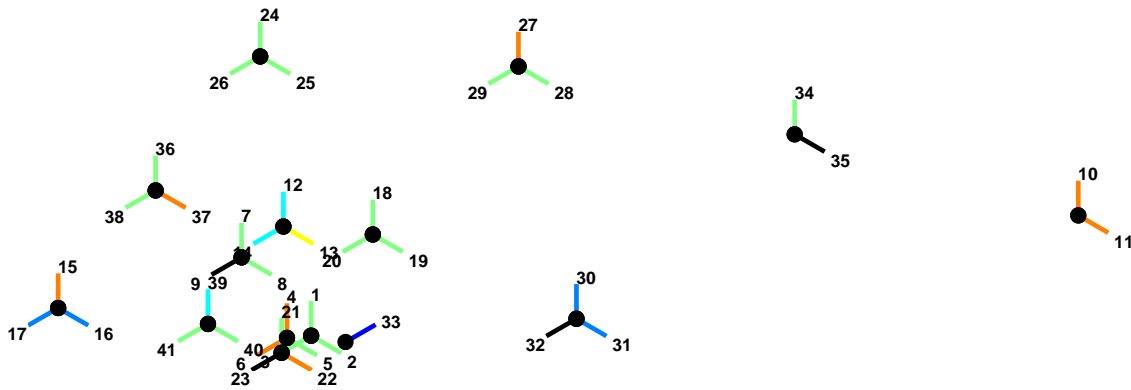


Figure 11.2: Classified GSM cells

When clustering mobile cells a histogram can be computed for each mobile cell. The histogram describes how the data from one cell fall into the data clusters. These histograms are used as profiles in cell classification. The profiles are fed into second SOM, which is clustered to find the classes of cell profiles. The classified mobile cells and their locations are presented in Figure 11.2. In this method, two level clustering procedure has been used because long term cell profiles are desired. The method gives us more reliable classification results.

## 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 SOM based approach for visualization of GSM network performance data. In M. Ali and F. Esposito, editors, *IEA/AIE*, volume 3533 of *Lecture Notes in Computer Science*, pages 588–598. Springer, 2005.
- [3] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

### 11.3 Using visualization, variable selection and feature extraction to learn from industrial data

Co-operation between the Academia and the general society has been defined by the Finnish Parliament as a major task for Finnish Universities. One of the tactics of the Laboratory of Computer and Information Sciences to meet this challenge is the development of the HAHMO-tool depicted in Fig. 11.3. This tool is intended to facilitate the definition of the goals and realization of an applied research project. This brief description illustrates the basis of the method and the current implementation.

The HAHMO is directly based onto the CRoss Industrial Standard Platform for Data Mining (CRISP-DM), developed by Chapman et al. in [1]. The CRISP-DM is a six-phase method for the definition and realization of a statistically motivated project. The first two phases require the user to define the Business Objectives and assess the quality and quantity of data. The next two phases lead to the pre-treatment of the data and to the actual modelling. The last two phases require assessment of the business potential of the statistical results, and require a deployment plan to be drawn. The CRISP-DM has been created to support industrial data mining projects. The HAHMO is a computer interface to this analysis approach.

The HAHMO tool leads the user through the CRISP-DM -process by asking questions and allowing the user to fill in his or her answers. These answers are stored into a data base to allow later study and refinement. The HAHMO also creates a report of the data fed into the data base.

The HAHMO has been, so far, piloted once. In Autumn 2005 we invited Isto Halinen from Efeko to pilot the HAHMO-tool. We held a one hour meeting in which we used the HAHMO to define their needs and our capability of serving those needs. The benefits of the HAHMO were as follows:

1. the meeting proceeded in a well organized manner, and
2. all relevant aspects of the task were traversed.

Mr. Halinen was pleased with the results and intended to forward the created plan in Kunnallisliitto, an organization to which they are related.

## References

- [1] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth CRISP-DM 1.0 step-by-step data mining guide. *Technical report, CRISM-DM consortium, <http://www.crisp-dm.org>, 2000.*

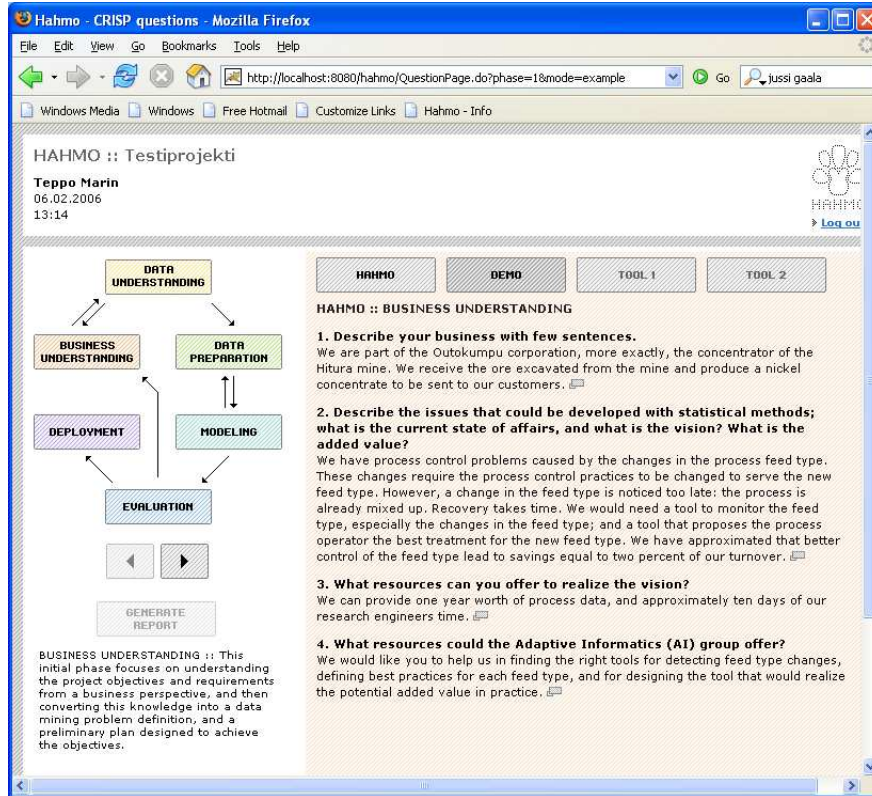


Figure 11.3: A screen shot from the HAHMO-tool.

## 11.4 SOM in decision support

The usability of Self-Organizing Map (SOM) method in computerized decision support systems is studied [1]. A prototype (DERSI) that combines neural methods and knowledge-based methodologies is being built (Fig. 11.4). DERSI is intended to be used first in fault diagnosis, but also other application areas are possible. Decision making problems in fault detection and identification are met for instance in the control rooms of power plants.

One goal of the study is to develop the control room tool with co-operation of end-users and other specialists. In addition the intention is to make data analysis with real data and simulated data of nuclear power plants, and analyze various failures. Some co-operation with the TVO Olkiluoto and Finnish Radiation Centre (STUK) have been initialized. One set of TVO simulator data is already being analyzed. One important research issue is to find out in what kind of failures the analysis can really help to find out the reason for the fault early enough, and concentrate on such scenarios in more detailed analysis. In initial studies e.g. leaks were found out promising in this respect.

DERSI is a general framework that can be used also in other purposes and application areas. For instance, paper industry could be a potential process where the same framework could be utilized.

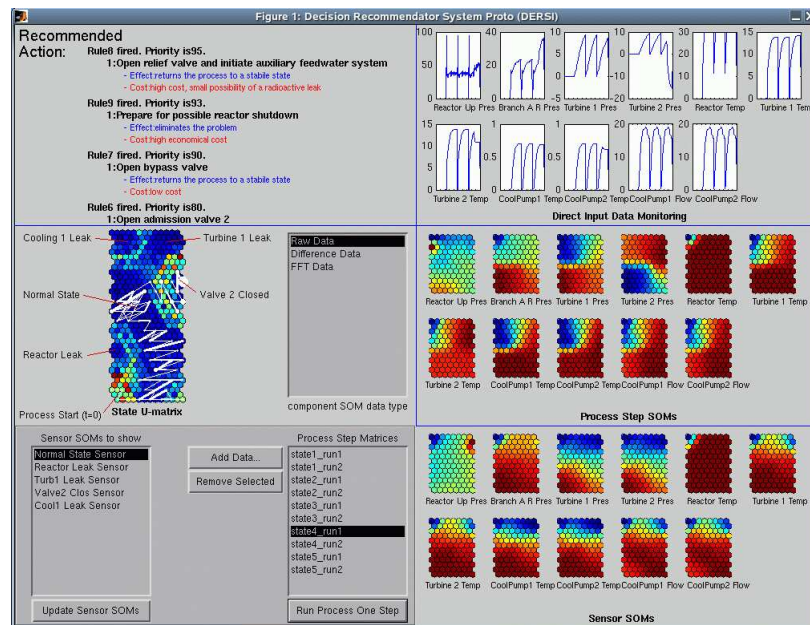


Figure 11.4: DERSI Man-Machine Interface (MMI) including different decision support visualizations.

## References

- [1] M. Sirola, G. Lampi, J. Parviainen. SOM based decision support in failure management. *International Journal of Computing*, 4(3): 124–130, 2005.

## 11.5 Interpreting dependencies in data using the Self-Organizing Map

The Self-Organizing Map (SOM) is a widely used method for the visualization of multivariate data. This work presents various applications of the SOM.

**Model interpretation.** Traditional regression means predicting the mean value of the output for a given value of the input. Instead of the mean value, quantile regression tries to predict the median or any other quantile of the output given the input. In [1] nonlinear quantile regression models and their interpretation is considered. Model interpretation is important if the purpose is not only the prediction itself, but the aim is to understand dependencies between the inputs and outputs. A novel model visualization technique is proposed based on the SOM.

**Variable selection.** Relations between two distinct sets of multivariate data are studied in [2, 3]. The SOM is used to visualize the dependent set of data. Then the proposed method finds variables of the independent set that are related to the visualization. To illustrate the method, we applied it to the task of finding the common properties of various car models that explain their safety and economic aspects.

**Manifold learning.** A modification of the SOM algorithm for visualizing a special type of data is presented [4]. It is known that the SOM has problems with data, which form a nonlinear manifold (for instance a heavily twisted sheet) in a high-dimensional space. On the other hand, some of the more recently presented projection methods can handle these cases. The proposed technique is a hybrid of the traditional SOM and one of these recent methods, see Fig. 11.5. Although the technique may seem to have a narrow area of applicability, there are problems, for instance in the area of image analysis, where intrinsically low dimensional and nonlinear data structures lie in a very high-dimensional pixel space.

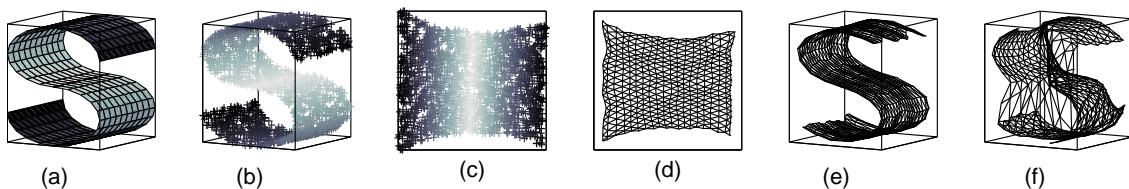


Figure 11.5: The problem of manifold learning for three-dimensional data (b) sampled from an intrinsically two-dimensional S-curve manifold (a). The Locally Linear Embedding algorithm [5] discovers the projection onto the internal coordinates on the manifold (c). The map trained by the algorithm proposed in [4] learns similarities in the internal coordinates (d) and forms a successful representation of the manifold in the observation coordinates (e). The basic SOM algorithm fails in this task (f).

## References

- [1] T. Similä. Self-organizing map visualizing conditional quantile functions with multi-dimensional covariates. *Computational Statistics & Data Analysis*, 50(8):2097–2110, 2006.



- [2] S. Laine and T. Similä. Using SOM-based data binning to support supervised variable selection. *Proceedings of the 11th International Conference on Neural Information Processing (ICONIP)*, pages 172–180, Calcutta, India. November 22–25, 2004.
- [3] T. Similä and S. Laine. Visual approach to supervised variable selection by self-organizing map. *International Journal of Neural Systems*, 15(1–2):101–110, 2005.
- [4] T. Similä. Self-organizing map learning nonlinearly embedded manifolds. *Information Visualization*, 4(1):22–31, 2005.
- [5] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

## 11.6 Analysis of forest nutrition data

Forests are complex ecosystems. Gaining an insight into the condition of forests and the assessment of the future development of forests under the present and predicted environmental scenarios requires large data sets from long-term monitoring programmes. In this project the development of forests in Finland has been studied using data from the International Cooperative Programme on the Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forests).

Plant nutrients play an integral role in the physiological and biochemical processes of forest ecosystems. We have analyzed the development of foliar nutrient concentrations in coniferous trees in Finland using clustering of the Self-Organizing Map [1, 2]. Based on these results further analysis of the effect of nitrogen and sulfur deposition on the mineral composition of foliage was conducted [3]. It was concluded that evidence for deposition-induced changes in needles has clearly decreased during the nineties.

Various environmental factors and past development affect the growth and nutritional composition of tree needles as they are aging. Different regression models were compared to find out how these effects [4] could be modelled effectively and accurately during the second year of the needles. We found that sparse regression models are well suited for this kind of analysis. They are better for the task than ordinary least squares single and multiple regression models, because they are both easy to interpret and accurate at the same time. Average  $R^2$  values of the different regression models for different element concentrations and mass of the needles (NM) are presented in Figure 11.6.

Good quality of analytical measurements techniques is important to ensure the reliability of analyses in environmental sciences. We combined foliar nutrition data from Finland and results of multiple measurement quality tests from different sources in order to study the effect of measurement quality on conclusions based on foliar nutrient analysis [5]. In particular, we studied the use of weighted linear regression models in detecting trends in foliar time series data and showed that good precision of the measurement techniques may decrease the time needed to detect statistically significant trends in environmental time series by several years.

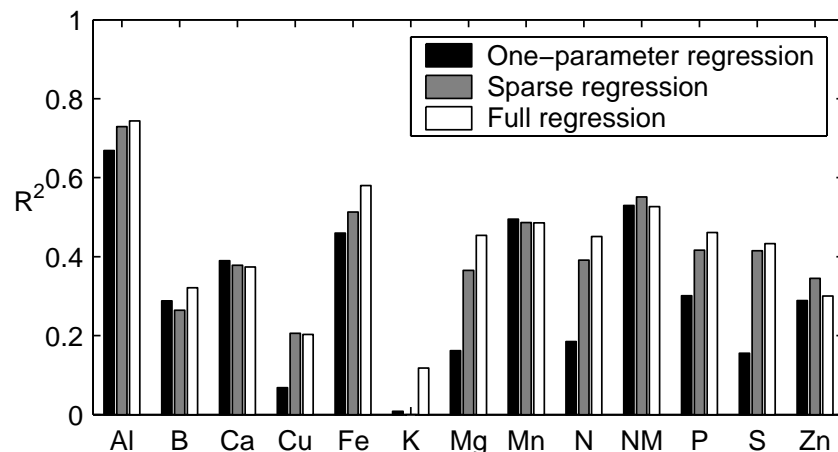


Figure 11.6: Average  $R^2$ -values of one-parameter regression, sparse regression and full regression models for foliar measurements from pine needles obtained using cross-validation. Results are for validation sets.

## References

- [1] Juha Vesanto and Mika Sulkava. Distance matrix based clustering of the self-organizing map. In José R. Dorronsoro, editor, *Artificial Neural Networks - ICANN 2002: International Conference, Proceedings*, volume 2415 of *Lecture Notes in Computer Science*, pages 951–956, Madrid, Spain, August 2002. Springer-Verlag.
- [2] Sebastiaan Luyssaert, Mika Sulkava, Hannu Raitio, and Jaakko Hollmén. Evaluation of forest nutrition based on large-scale foliar surveys: are nutrition profiles the way of the future? *Journal of Environmental Monitoring*, 6(2):160–167, February 2004.
- [3] Sebastiaan Luyssaert, Mika Sulkava, Hannu Raitio, and Jaakko Hollmén. Are N and S deposition altering the chemical composition of Norway spruce and Scots pine needles in Finland? *Environmental Pollution*, 138(1):5–17, November 2005.
- [4] Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. In Sašo Džeroski, Bernard Ženko, and Marko Debeljak, editors, *Proceedings of the Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)*, pages 57–58, Bled, Slovenia, September/October 2004.
- [5] Mika Sulkava, Pasi Rautio, and Jaakko Hollmén. Combining measurement quality into monitoring trends in foliar nutrient concentrations. In Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrozny, editors, *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005: 15th International Conference, Proceedings, Part II*, volume 3697 of *Lecture Notes in Computer Science*, pages 761–767, Warsaw, Poland, September 2005. Springer-Verlag.

## 11.7 Parsimonious signal representations in data analysis

Data mining is the analysis of often large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner [1]. While utility is a natural starting point for any analysis, understandability often remains a secondary goal. In this research, improved understandability of data-analytic models is sought by investigating sparse signal representations that are learned from data.

In order to learn about the relationships between variables, a method for learning parsimonious dependency structures has been presented in [4]. The class of dependency structures is the set of linear regression models. Each variable is taken as the output variable in turn and a sparse regression algorithm is used to select a parsimonious set of inputs from the rest of variables. A bootstrap-based procedure has been developed in order estimate robust dependency structures. A linear dependency forest of System data is shown in Figure 11.7. The System data consist of measurements from nine variables of a single computer which is connected to a network [7]. In Fig. 11.7, `idle`, `ipkts`, and `blks/s` are the output variables and the rest are the input variables. The markers + and - indicate positive and negative effect of the input variable on the output variable, respectively.

In a time series context, parsimonious modeling techniques can be used in estimating a sparse set of autoregressive variables for time series prediction [5]. We present an algorithm in the spirit of backward selection, which removes variables sequentially from the prediction models based on the significance of the individual regressors using bootstrap-based confidence intervals for the prediction error.

In ecology, a needle aging prediction problem has been casted into the framework of parsimonious modeling. In [2, 3], we show how linear sparse regression models can be used to represent the relations between different foliar nutrient concentration measurements of coniferous trees in consecutive years. In the experiments, the models proved to be capable of providing relatively good and reliable predictions of the development of foliage with a considerably small number of regressors. Two methods for estimating sparse models were compared to more conventional linear regression models. Differences in the prediction accuracies between the sparse and full models were minor, but the sparse models were found to highlight important dependencies between the nutrient measurements better than the other regression models. The use of sparse models is, therefore, advantageous in the analysis and interpretation of the development of foliar nutrient concentrations.

The problem of estimating sparse regression models in a case of multi-dimensional input and output variables has been investigated in [6]. A forward-selection algorithm that extends the Least Angle Regression algorithm (LARS) is presented. The proposed method is also applied to the task of selecting relevant pixels from images in multidimensional scaling of handwritten digits.

The research work on parsimonious modeling will be continued in various applications in bioinformatics, ecology and general data mining problems. The projects involving domain expertise will be especially useful to see the how well the results can be interpreted.

## References

- [1] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. Adaptive Computation and Machine Learning Series. MIT Press, 2001.

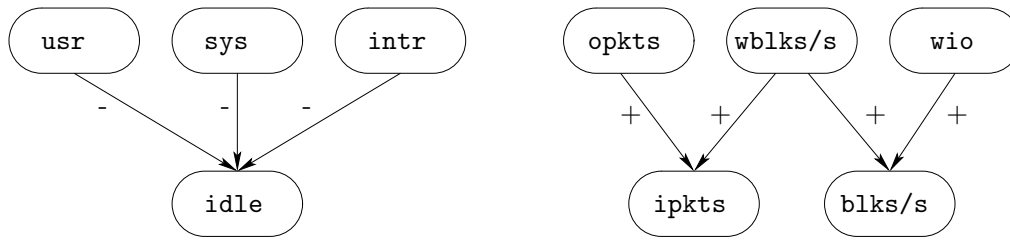


Figure 11.7: The linear dependency forest obtained from the System data. The arrows point from the input variables to the output variables.

- [2] Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. In Sašo Džeroski, Bernard Ženko, and Marko Debeljak, editors, *Proceedings of the Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)*, pages 57–58, 2004.
- [3] Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. *Ecological Modeling*, 191(1):118–130, 2006.
- [4] Jarkko Tikka and Jaakko Hollmén. Learning linear dependency trees from multivariate time-series data. In *Proceedings of the Workshop on Temporal Data Mining: Algorithms, Theory and Applications (in conjunction with The Fourth IEEE International Conference on Data Mining)*, Brighton, U.K., 2004.
- [5] Jarkko Tikka, Jaakko Hollmén, and Amaury Lendasse. Input Selection for Long-Term Prediction of Time Series. In Joan Cabestany, Alberto Prieto, and Francisco Sandoval, editors, *Proceedings of the 8th International Work-Conference on Artificial Neural Networks (IWANN 2005)*, volume 3512 of *Lecture Notes in Computer Science*, pages 1002–1009. Springer-Verlag, June 2005. Vilanova i la Geltrú, Barcelona, Spain.
- [6] Timo Similä and Jarkko Tikka. Multiresponse Sparse Regression with Application to Multidimensional Scaling. In Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrozny, editors, *Proceedings of the 15th International Conference on Artificial Neural Networks (ICANN)*, volume 3967 part II of *LNCS*, Springer-Verlag, pages 97–102, Warsaw, Poland, September 2005.
- [7] Juha Vesanto and Jaakko Hollmén. An Automated Report Generation Tool for the Data Understanding Phase. In Ajith Abraham, Lakhmi Jain, and Berend J. van der Zwaag, editors, *Innovations in Intelligent Systems: Design, Management and Applications*, Springer (Physica) Verlag, Studies in Fuzziness and Soft Computing, 2003.

