Chapter 15

Intelligent data engineering

Esa Alhoniemi, Jaakko Hollmen, Johan Himberg, Sampsa Laine, Golan Lampi, Pasi Lehtimäki, Teppo Marin, Jukka Parviainen, Kimmo Raivio, Timo Similä, Olli Simula, Miki Sirola, Mika Sulkava, Jarkko Tikka, Juha Vesanto

15.1 Spatio-temporal analysis of forest nutrition data

Living plants are capable of taking up substances from the environment and using them for the synthesis of their cellular components. Plant nutrients play an integral role in the physiological and biochemical processes of forest ecosystems. Therefore the nutritional status of trees provides an important diagnostic tool for estimating tree condition [1]. In this project, the nutrient concentrations of pine and spruce needles in Finland and Austria between 1987–2000 were studied using different data analysis methods [2]. The aim was to analyze the spatial and temporal distribution of the nutrients and generally find out what kind of internal structure exists in the data. The analysis methods used in [2] were spatial statistics, clustering of the self-organizing map and time series modeling. The work was done in collaboration with the Finnish Forest Research Institute, Parkano Research Station.

The clustering method of the self-organizing map [3] provided new information about the relations of the nutrients between different years and locations. The clustering method was able to represent the structure of the relations of nutrient concentrations in a new informative way. Using the clustering, we were able to divide the measurements into groups [2]. The clustering result of Finland on the geographical map in different years is presented in Figure 15.1. In each group the growth of the needles and the amounts of the nutrients were different and thus, different groups represented different kinds of growing conditions. Using the result of the clustering method, it was possible to construct a temporal model that characterizes the development of the forests of Finland.

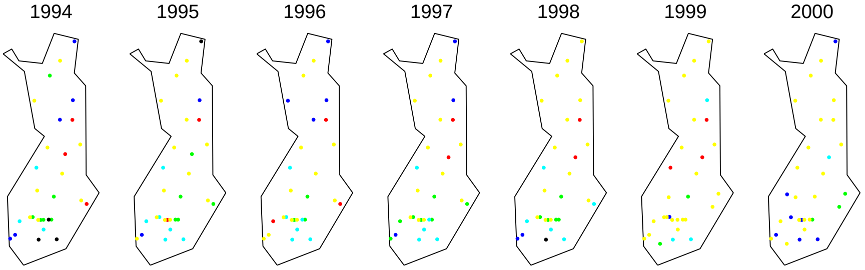


Figure 15.1: Clustering of the measurement stands of Finland for years 1994–2000. Colors indicate different clusters.

References

[1] Sebastiaan Luyssaert, Hannu Raitio, and Alfred Fürst. Forest nutrition at the Finnish and Austrian level I plots in 1987-2000. Technical report, The Finnish Forest Research Institute and Austrian Federal Office Research Centre for Forests, 2003.

[2] Mika Sulkava. Identifying spatial and temporal profiles from forest nutrition data. Master's thesis, Helsinki University of Technology, May 2003.

[3] Juha Vesanto and Mika Sulkava. Distance matrix based clustering of the self-organizing map. In José R. Dorronsoro, editor, *Artificial Neural Networks - ICANN 2002*, volume 2415 of *Lecture Notes in Computer Science*, pages 951–956, Madrid, Spain, August 2002. Springer.

[4] Mika Sulkava and Jaakko Hollmén. Finding profiles of forest nutrition by clustering of the self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 243–248, Hibikino, Kitakyushu, Japan, September 2003.

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

Although the engineers of industry have access to process data, they seldom use advanced statistical tools to solve process control problems. Why this reluctance? I believe that the reason is in the history of the development of statistical tools, which were developed in the era of rigorous mathematical modelling, manual computation and small data sets. This created sophisticated tools. The engineers do not understand the principles of these algorithms related. If algorithms are fed with unsuitable data, or are parameterized poorly, they produce biased results, which probably leads an engineer to descregard statistical tools.

My thesis work [1] proposes algorithms that probably do not impress the champions of statistics, but serve process engineers. I advocate the three properties: supervised operation, robustness and understandability. Supervised operation allows and requires the user to explicate the goal of the analysis. Robustness allows the analysis of raw process data. Understandability is essential, as the user must know how to parameterize the algorithm, and how to interpret the results.

To realise a methodology that complies with the above criteria, I studied three types of algorithms: visualization, variable selection and feature extraction. Variable selection helps the user to find relevant variables among the hundreds of variables provided by an automation system; Feature extraction helps the user to mathematically manipulate the variables to surface relevant information; Visualization provides understandable presentation of the results. Figure 15.2 illustrates these three tools together with the three criteria.

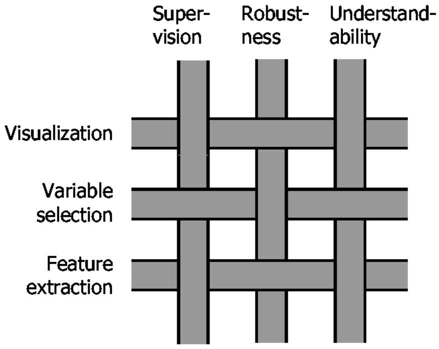


Figure 15.2: The three criteria and tool types discussed in my thesis

I illustrated my approach by analysing an industrial case: the concentrator of the Hitura mine. A significant benefit of algorithmic study of data is efficiency: the manual approach reported in my early publications took approximately six man months to produce; the automated approach of this thesis created comparable results in few weeks.

References

[1] S. Laine, Using visualization, variable selection and feature extraction to learn from industrial data, PhD thesis, Helsinki University of Technology, 2003. Available in `http://lib.hut.fi/Diss/2003/isbn9512266709/`.

15.3 Decision models for computerized decision support

Computerized decision support system prototypes have been developed and summarized in [1], where decision making problem formulation of failure management in safety critical processes was studied. The main application area was the nuclear power plants. Decision support in both the control room and maintenance were covered. One of the models, the use of decision analysis methodology in maintenance problems, is presented in [2].

After these works the objective has been to build decision models for certain applications, mostly for practical purposes, and also try to find out more general decision making principles. This approach leads easily to single case studies that are difficult to generalize. The large amount of possible methodologies and the narrowness of application areas are also known difficulties.

To find out general principles from separate case studies, to formulate more comprehensive decision concepts, and to build more general decision models are difficult tasks. While such studies produce tested models and concepts, evaluation of these results is difficult, because there are no competent measures for such purposes. The only really clear result is the decision support achieved in each particular case.

How to utilize data analysis in computerized decision support systems has been outlined. A prototype is being built to demonstrate how to utilize Self-Organizing Map in a computerized decision support system.

An old decision case that has been analyzed with rule-based methodologies in [1] has been solved with multi-criteria decision analysis method in [3]. A Comparison with the elder case has been made in the analysis. The problem is to choose the right control action in a situation where a leak has appeared in the primary circuit of a BWR nuclear power plant.

Decision concepts have been reviewed and a conceptual decision model has been built by case-based means [4]. This model utilizes rule-based methodologies, numerical algorithms and procedures, statistical methodologies including distributions, and visual support. Probability models are used in handling uncertainties.

References

[1] Sirola M. Computerized decision support systems in failure and maintenance management of safety critical processes. VTT Publications 397. Espoo, Finland, 1999.

[2] Laakso K., Sirola M., Holmberg J. Decision modelling for maintenance and safety. International Journal of Condition Monitoring and Diagnostic Engineering Management. Birmingham, England, July 1999.

[3] Sirola, M. Applying decision analysis method in process control problem in accident management situation. International Conference on Systems Science. Wroclaw, Poland, September 2001.

[4] Sirola M. Using conceptual decision model in a case study. International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'2003). Oxford, United Kingdom, September 2003.

15.4 Context awareness

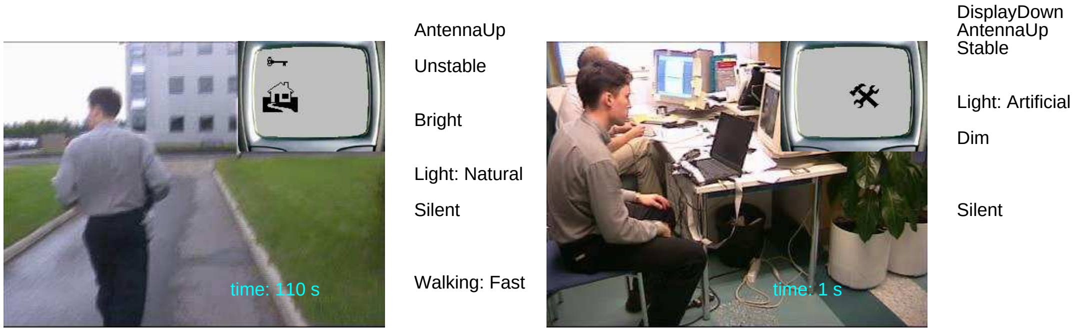
Context awareness [4] has become a major topic in human-computer interaction. The need for context awareness is especially large in mobile communications, where the communication situations can vary a lot. A mobile terminal is often expected to enable connections all the time. At the same time, it should not irritate the user by signalling in a wrong way at the wrong moment, or by requiring constant attention to keep it working in the right way for the situation. In addition, the hand-held terminals are becoming more and more sophisticated in their function yet smaller in their size. The interaction could be made easier and less intruding if the mobile device recognized the user's current context and adapted its functions accordingly.

Information of user's preferences is obtained from the logs of different applications, e.g., calling, messaging, using calendar, or profiling. Piece of ambient information can be obtained by directly monitoring the user's physical environment using on-board sensors and information of user's location. The operating network itself can offer information, e.g., on location. Setting explicit information sources, context tags, located in a short range network is another approach.

The device can infer parts of the context of the user from features extracted from on-board measurements of acceleration, noise level, luminosity, humidity, etc. In [1],[2], we have consider context recognition by fusing and clustering these context features using a recently introduced method, the Symbol Clustering Map (SCM) [3]. As such, it can be used for finding static patterns but a suitable transformation of the data allows identifying also temporal patterns. The recognized clusters/segments can then serve as "higher-level contexts" that show which combinations of the basic features form common patterns in the data.

Fig. 15.3 presents an user, the context features and the recognized context in two different situations. The context is presented here as a user interface profile. In this case, common contexts are recognized unsupervised by the SCM from training data. However, the labeling (deciding the profile) is done afterwards by the user, and the selection of the profiles/actions is based on a lookup table. A future aim is that the terminal would also learn to suggest applications according to user's spontaneous actions in different situations.

Publications [1, 2] are joint work with Dr. John Adrian Flanagan in Nokia Research Center, Helsinki, Finland and Dr. Jani Mäntyjärvi in VTT Technical Research Centre of Finland, Oulu, Finland.



(a) (b)

Figure 15.3: In panel (a) SCM has recognized a "walking outdoors" context based on the active features listed to the right of the image. "Keypad lock on" and "Outdoors profile" have been activated according the the action lookup table. In Panel (b) a "working profile" is launched due to the office context.

References

[1] J.A. Flanagan, J. Himberg, and J. Mäntyjärvi. A Hierarchical Approach to Learning Context and Facilitating User Interaction in Mobile Devices. In *Proceedings of Artificial Intelligence in Mobile System 2003 (AIMS 2003).* (in conjunction with Ubicomp 2003, October 12, Seattle, USA.)

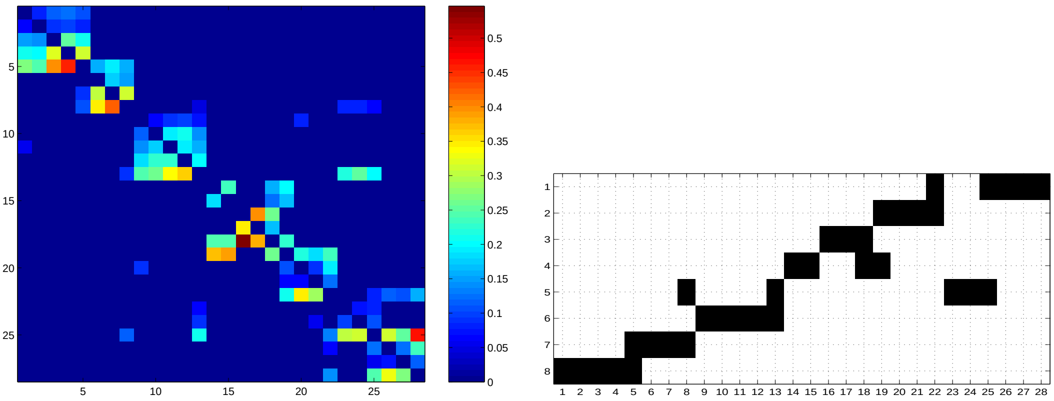
[2] J. Himberg, J. A. Flanagan, and J. Mäntyjärvi. Towards Context Awareness Using Symbol Clustering Map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pp. 249–254, Hibikino, Kitakyushu, Japan, September 2003.

[3] A. Flanagan. Unsupervised Cluster Discovery using the Self-Organizing Map. In *Proc. of International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP 2003)*, pp. 9-12, 2003, Istanbul, Turkey.

[4] A.K. Dey and G.D. Abowd and D. Salber. A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. *Human-Computer Interaction*, 16(2–4):97–166, 2001.

15.5 Dependency trees from industrial time-series data

Industrial processes generate large masses of multivariate, noisy time-series data. In exploring the database, the analyst is interested in looking at the structure of the data set, or more specifically dependencies among the variables. Our problem is to seek for dependencies between the N variables in the data set. The dependencies are defined through multiple linear regression models, which are estimated from the data. Before model fitting, time-series are denoised using the Wavelet transform. In model fitting, one variable at the time is the dependent variable and rest of the variables are possible regressors. Sparse regression algorithms are used to select the best regressors among all candidates and estimate the corresponding regression coefficients. Bootstrap is also applied on the selection and the estimation. The relative weight of the each regressor is computed from the bootstrap replications of the regression coefficients. The relative weight of the regressor is a measure of belief that the regressor belongs to the estimated linear model. These relative weights are thresholded to yield graphs, some graph operations are performed to define dependent variables. Taken together, the method defines a dependency tree, or possibly a dependency forest. More detailed results will be reported in a Master's thesis.



(a) (b)

Figure 15.4: Examples from an artificial data set. In (a), the relative energies of the sparse linear model coefficients estimated from the bootstrapped data sets are shown. In (b), dependencies between variables are defined using thresholding and operations on the resulting graph.