Variable selection and grouping for large-scale data-driven modelling

Esko K. Juuso

Control Engineering, Environmental and Chemical Engineering, Faculty of Technology,
University of Oulu,Finland
esko.juuso@oulu.fi

Abstract

For large-scale systems, the number of possible variable combinations becomes very large. Variable grouping means finding feasible groups of variables for modelling. Systems can be divided into subsystems but even then the number of available variables is often impractically high to be used with the data-based methods. Interactive variable selection and grouping by comparing the performance of the model alternatives is a good solution if there are not too many variables. This paper describes possibilities of variable selection in large-scale industrial systems. It classifies the variable selection and grouping into four categories: knowledge-based grouping, grouping with data analysis, decomposition, and model-based grouping and selection. The data analysis part consists of correlation analysis and handling of high dimension data with principal components. These originally linear methodologies were extended to nonlinear systems by using the nonlinear scaling approach. Decomposition can be realised with various clustering methods or learning with case-based reasoning. The multimodel systems are handled with fuzzy set systems. Numerous studies based on linear multivariate statistical modelling have been reported in literature. The methodologies approaches have been tested in several applications: bioprocesses, continuous brewing, condition monitoring, web break sensitivity analysis and wastewater treatment. Industrial process data, a pilot system and a test rig were used in the analysis. Uncertainty handling is a part of the analysis method: uncertainty is represented with the degrees of membership.

Keywords: variable selection and grouping, data analysis, intelligent methods, data-driven modelling

1 Introduction

Data-driven modelling always requires variable grouping and selection. In small systems, expert knowledge gives a clear basis for the variable selection since possible interactions and causal effects are known fairly well. For these cases, few modelling alternatives can be compared interactively. Variable selection becomes important when the number of variables increases, especially when normal process data is used. In large-scale systems, the number of possible variable combinations becomes easily very large (Figure 1), This rapidly increasing number of combinations, known as the combinatorial explosion (Pyle, 1999), can easily defeat even powerful computers.

In big data processing, the analysis is even more challenging (Hashem et al., 2015). The amount of different types of data generated from different sources is increasing fast. Discovering hidden values from these large heterogeneous datasets requires versatile methodologies (Juuso, 2020a). Neural deep learning methods provide highly complex structures (Schmidhuber, 2015) but because of a huge set of parameters they are not easy to assess with expert knowledge.

The model assessment becomes easier through the better process insight provided by the modules based on analyzed variable groups. Already, the development and tuning require that the models should not include too many variables. In practical cases, variable selection is necessary either because it is computationally infeasible to use all available variables, or when limited data samples have numerous variables. Finding feasible groups and combinations of variables for modelling is closely connected with data clustering since the interactions can depend on the operating area. Variable selection, also known as subset selection or feature selection, is a process commonly used in machine learning, wherein a subset of the features available from the data are selected for application of a learning algorithm.

Systems can be divided into subsystems but even the number of available variables is often impractically high to be used with the data-based methods. Interactive variable selection and grouping by comparing the performance of the model alternatives is a good solution if there are not too many variables. This approach can be extended to a wider set of variables by evolutionary computation. As the number of variables becomes too big even for these methods, the number of alternatives must be reduced before model development.

There is a lot of literature on both the model and data-based techniques. Spectroscopic data, multi-sensor systems, multivariate analysis and modelling of large-scale systems seem to require efficient methods for variable selection. A large number of different methods have also been used in process monitoring (Venkatasubramanian et al., 2003). Commonly used data-based monitoring methods are reviewed in (Vermasvuori, 2006). Partitioning-based clustering algorithms are compared in

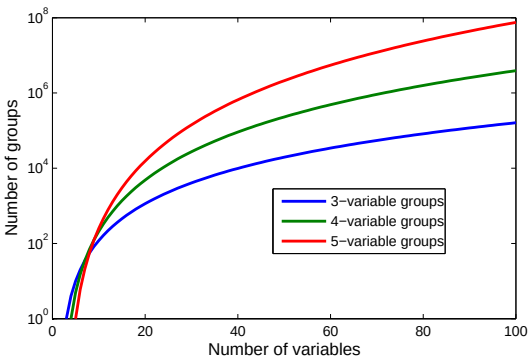

Figure 1. Variable combinations (Juuso and Ahola, 2008).

(Äyrämö and Kärkkäinen, 2006). A literature review of inference and decision making methods in fault diagnosis is presented in (Cheng, 2006).

The final selection and grouping step is based on modelling. Multivariate statistical modelling and structural relationships are widely used. In linguistic equation (LE) models, nonlinear scaling is used together with one or more linear equations (Juuso, 2004). Equations can be generated for all the combinations, e.g. three-variable combinations, or selected combinations. Selected combinations can be constructed also for several groups of variables by generating all the combinations within each variable group. The set of variable groups may also contain groups with different number of variables. For small systems, these groups can be defined manually. Non-informative groups can be removed manually but the variable selection need to be partially automated for large-scale systems.

This paper focuses on the methodologies of variable selection for large-scale modelling. The analysis starts with knowledge-based methods (Section 2) before going to data-based grouping (Section 3). Decomposition is needed for more complex structures (Section 4). The selection and grouping are finalized with modelling (Section 5). Several applications are discussed in Section 6. The classification of methodologies is discussed in Section 7. Conclusions and future studies are presented in Section 8.

2 Knowledge-based variable grouping

Knowledge-based information is essential for all types of variable selection and grouping. For small systems, only expert knowledge is needed. In large-scale systems, expertise is used for selecting variable combinations which should be avoided, e.g. calculated variables should not be used together with the variables used in calculating them. This is important if indirect measurements are used. A group containing a controlled variable and its set point is not usually appropriate. These problems are avoided by defining the inappropriate groups as non-groups, i.e. as variable groups which should not be a part of any acceptable variable group.

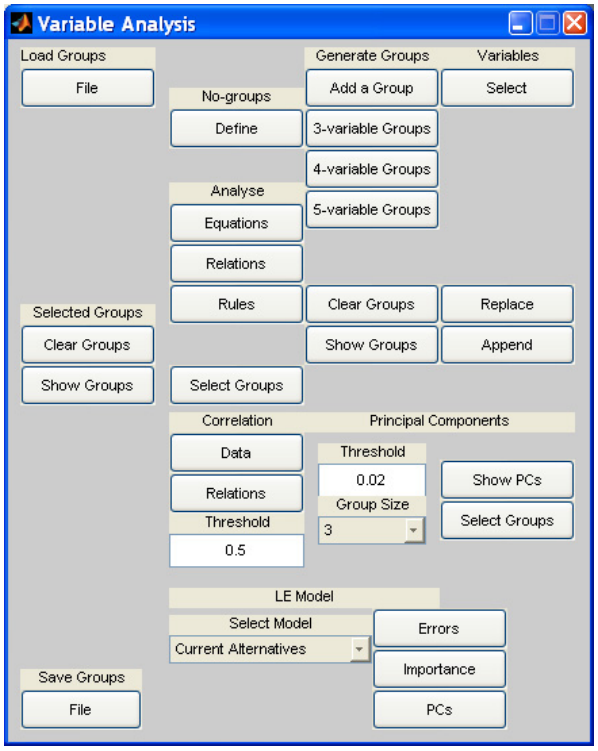

Figure 2. Variable selection (Juuso et al., 2008).

Too few variables mean that models cannot capture the phenomena. Too many variables may cause overfitting and models which are difficult to understand and evaluate. This is especially important for frequency range analysis. Redundancy, i.e. practically the same measurement is obtained by several sensors, the measurements can be combined as a weighted sum. However, sensor failures must be taken into account in real-time applications.

Knowledge-based information can effect in many ways on the variable grouping, which reduces strongly the variable combinations. For large-scale systems, some variables are suitable for the system decomposition and some for developing specialized models. The working point variables, which are used for defining different operating areas, are not necessarily in the corresponding specialized models. The specialized models can be totally different in different operating conditions.

Each equation has normally from three to five variables, and the FuzzEqu Toolbox (Juuso, 2000) is designed for analyzing these variable groups (Figure 2). The generation of the alternatives is based on three variable groups: one variable is selected to be in all four variable groups, and two variables are selected to be in all five variable groups.

The interactive variable grouping shown in Figure 2 can be used as a tool in learning the system. Three variable groups are the basic elements. The subsets of the variables and the important variables in the four and five variable groups can be based on process knowledge. This analysis already reduces the number of alternatives with 70 percent (Juuso et al., 2008).

DOI: 10.3384/ecp2118538

3 Grouping with data analysis

Correlation analysis provides methods for pruning variables. Principal component analysis (PCA) is a well-known method for variable selection in large data sets. PCA explains variations within one data set.

3.1 Correlation analysis

Correlation coefficients are indicators of the strength of the linear relationship between two variables. The most common coefficient is called Pearson's product-moment correlation coefficient (Ranta et al., 1999; Karttunen, 1994). Statistical correlations are not indications of real causal interactions. Statistical reasoning based on correlation coefficients presumes bivariate normal distribution between variables (Ranta et al., 1999). This assumption is fairly seldom true in process data.

Binary correlations and their combinations are used for pruning the set of acceptable groups defined by domain expertise (Section 2). For forecasting models, input variables should have high correlation to the output variable but low correlation between each other. For case detection, this requirement is not necessary (Section 3.3).

Low correlation may be caused by noise and observation errors. The result is improved by using appropriate filtering and correct time delays between the variables. Moving averages, medians or value ranges have time delays which depend on the calculation window and the applied methodology.

3.2 Correlations in nonlinear systems

For the nonlinear systems, basic correlation analysis and rank correlation methods like Spearman rank correlation coefficients and Kendall rank correlation coefficient do not give reliable results although Spearman correlation coefficient can sometimes identify also nonlinear interaction between variables (Ranta et al., 1999).

There are extensions of the analysis to nonlinear systems (Juuso et al., 2008). A nonlinear correlation of a binary relationship can be implemented for example by using time sequential joint transfer correlation (JTC), morphological correlation (MC) and sliced orthogonal nonlinear generalized decomposition (SONG) (Oton et al., 2005). A weighted SONG (WSONG) correlation has been presented in (Garcia-Martinez et al., 2002). The WSONG correlation is based on the sum of many linear correlations between binary images.

The nonlinear scaling brings measurements and features to the same scale by using monotonously increasing scaling functions $x_j = f(X_j)$ where x_j is the variable and X_j the corresponding scaled variable. The function $f()$ consist of two second order polynomials, one for the negative values of X_j and one for the positive values, respectively. The corresponding inverse functions $X_j = f^{-1}(x_j)$ based on square root functions are used for scaling to the range [-2, 2], denoted as linguistification. The monotonous functions allow scaling back to the real values by using the function $f()$. (Juuso, 2004)

The nonlinear scaling functions are developed by using central tendencies, statistical dispersion and shape of the data distribution. The data-based method for developing the scaling functions is presented in (Juuso and Lahdelma, 2010). Nonlinear models can be developed by using these scaled values and linear relationships. This approach extends the correlation analysis for curvilinear relationships.

3.3 Correlations in variable groups

For the multivariable correlation, Kendall's coefficient of concordance is a measure of agreement among raters is often used (Ranta et al., 1999). It is based on the rank values of observations. In variable group correlation analysis, the scaled values of variables are used and the evaluation of interaction is based on multivariate regression.

Combinations of the binary correlation coefficients are used in the FuzzEqu Toolbox. The methodology depends on the model type. For forecasting models, the correlations between the input variables should be low, and each input variable and the output variable should have high correlation. For detecting operating conditions, there are not necessarily any output variable, i.e. also groups where several variables have high correlation between each other are acceptable. Both alternative approaches are used for variable grouping for the detection of operating conditions.

3.4 High-dimensional data

Principal component analysis (PCA) is a data reduction method using mathematical techniques to identify patterns in a data matrix. The principal components are a small set of new orthogonal, i.e. non-correlated, variables derived from a linear combination of the original variables. They do not necessarily have any meaning as they are combinations of initial variables. However, these new axes provide the angles to see the data by representing the directions of the data which explain a maximal amount of variance.

The main element of this approach consists of the construction of PCAs which compres the data by reducing the number of dimensions with minor losses of information. Each principal component is a linear combination of the original variables.

The full set is as large as the original set of variables but it is common that the sum of the variances of the first few principal components to exceed 80 percent of the total variance of the original data. All the principal components are orthogonal to each other so there is no redundant information. The plots of these new variables help to understand the driving forces of the system.

Principal components can be used In data-based process monitoring by tracking how well the data points are explained with the PCA model (Vermasvuori, 2006).

Testing of loadings and their estimated standard uncertainties are used to calculate significance on each variable for each component (Westad et al., 2003). Variable selection can also mean identifying a k-subset of a set of

original variables that is optimal for a given criterion that adequately approximates the whole data set (Cadima et al., 2004).

The static PCA can be extended to dynamic systems by using several past measurements. The number of the lagged variables is selected in tuning (Ku et al., 1995; Li and Qin, 2001; Vanhatalo et al., 2017). Another variant is moving PCA developed by (Kano et al., 2001) for detecting changes in operating conditions. A special multiway approach has been developed for analyzing variations from the normal trajectories in batch processes (Nomikos and MacGregor, 1994). Multiscale PCA combines PCA with wavelet decompositions (Bakshi, 1998; Aminghafari and Cheze, 2006).

PCA is a linear method, which does not produce accurate results for nonlinear processes. Linguistic principal component analysis (LPC) extends the linear combinations for nonlinear systems by combining the nonlinear scaling with PCA. In the FuzzEqu Toolbox, the scaling functions can be used also for the normal principal components.

4 Decomposition

A modelling problem can be divided into smaller parts by developing separate models for independent subprocesses. In addition to spatial or logical blocks decomposed modelling can be based on different frequency ranges. Different operating conditions can be detected with cluster analysis, model-based reasoning, rule-based reasoning, or learned with case-based reasoning. Fuzzy set systems provide feasible techniques for handling the resulting partially overlapping models.

4.1 Clustering

Clustering consists of partitioning a data set into subsets (clusters), so that the data in each subset share common similarities or proximities for some defined distance measures. Cluster analysis, also called segmentation analysis or taxonomy analysis, is a way to create groups of objects, or clusters, in such a way that the profiles of objects in the same cluster are very similar and the profiles of objects in different clusters are quite distinct. Similarity criteria (distance based, associative, correlative, probabilistic) among the several clusters facilitate the recognition of patterns and reveal otherwise hidden structures.

- Hierarchical clustering groups data, simultaneously over a variety of scales, by creating a cluster tree. The tree is not a single set of clusters, but rather a multi-level hierarchy, where clusters at one level are joined as clusters at the next higher level. Hierarchical clustering produce a set of solutions with different numbers of clusters. The level or scale of clustering is chosen according to the application.

- Partitioning-based clustering algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained. In a basic iterative algorithm, such as K-means- or K-medoids, convergence is local and the globally optimal solution can not be guaranteed. (Äyrämö and Kärkkäinen, 2006) The fuzzy c-means algorithm imposes a spherical shape on the clusters, regardless of the actual data distribution (Babuška, 1998).

- Fuzzy clustering -based clustering algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained (Bezdek, 1981). In a basic iterative algorithm, such as K-means- or K-medoids, convergence is local and the globally optimal solution can not be guaranteed. (Äyrämö and Kärkkäinen, 2006)

- Neural computing can be used for clustering. Self-organising maps (SOM) (Kohonen, 1995) can be used for finding operating conditions or simply for clustering. Also radial basis networks (Chen et al., 1991) combine clustering with models.

- Nonlinear clustering is aimed to detect clusters of different geometrical shapes. (Gustafson and Kessel, 1979) extended the standard fuzzy c-means algorithm for this. The nonlinear scaling extends the clustering methods to different shapes.

- Robust clustering uses spatial medians to reduce effects of erroneous and missing values (Äyrämö and Kärkkäinen, 2006).

4.2 Reasoning

Decision making between different operating conditions is based on different types of reasoning. These methods can also be considered as inference fault diagnosis methods (Cheng, 2006).

Model-based reasoning. Causal directed graphs represent physical cause-effect relations between variables. Fault tree analysis provides graphical models of the pathways within a system that interconnect the basic cause events and conditions using standard symbols, and lead to a foreseeable, critical event. Bayesian networks represents probability relations among random variables as a graphical model to be used in probability inference.

Rule-based reasoning. Rule-based systems use IF-THEN rules representing domain expertise. Inference methods are data-driven forward chaining and goal-driven backward chaining. Conflict resolution is applied to select one rule out of the active ones. Preferences and priorities may be utilized in the conflict resolution. Fuzzy logic provides feasible solutions for resolving the conflicts and handling the uncertainty.

Case-based reasoning. Case-based reasoning (CBR) is a problem solving paradigm for finding out the solution to a new problem by remembering previous similar situations and by reusing information and knowledge of that situation (Aamodt and Plaza, 1994). The solutions are maintained in carefully indexed memory. The case base containing the previous cases with possible general knowledge of the problem area and the problem solving starts with the identification of the current problem situation (Figure 3). With information on new cases, the most similar case is retrieved from the case base. The retrieved case is reused to solve the current problem. During the revise step the suggested solution is evaluated to get the confirmed solution. Finally, the useful solutions with related case information are retained as new learned cases to the case base.

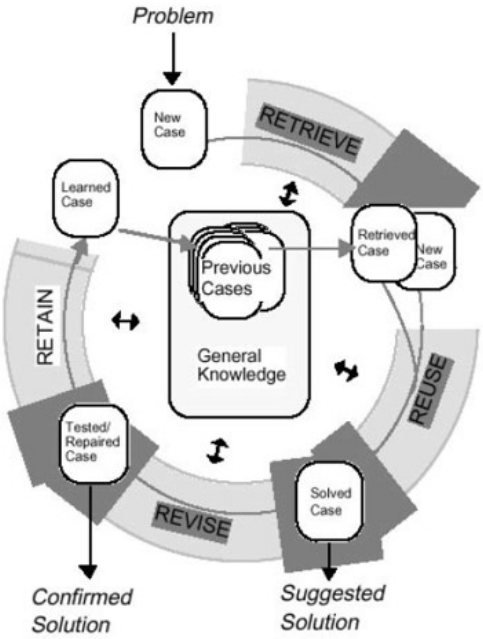

Figure 3. Case-based reasoning (CBR) (Aamodt and Plaza, 1994).

5 Model-based selection and grouping

Multivariate statistical tools are used for analyzing data matrices with regression and/or pattern recognition techniques. These methodologies are primarily linear. Nonlinear systems can be handled with fuzzy set systems, artificial neural networks or their combinations. In LE models, linear methodologies are combined with the nonlinear scaling discussed in Section 3.2.

Multivariate statistical modelling. The application of principal component regression (PCR) to the trajectories of the process variables (block-wise PCR) has given straightforward results without requiring a deep knowledge of the process (Zarzo and Ferrer, 2004). Partial least squares (PLS), also known as projection to latent structures, is a robust multivariate generalized regression method using projections to summarize multitudes of potentially collinear variables (Gerlach et al., 1979).

Fuzzy set systems. Different types of fuzzy set systems have been compared in (Juuso, 2004). Prior knowledge can be used in constructing rule-based fuzzy models: qualitative knowledge can be incorporated in linguistic fuzzy models (Driankov et al., 1993), or in fuzzy relational models if there are several alternative rules (Pedrycz, 1984); locally valid linear models can be collected by Takagi-Sugeno (TS) fuzzy models (Takagi and Sugeno, 1985).

Artificial neural networks. Artificial neural networks (ANNs) are commonly used in modelling of large scale systems. ANNs are nonlinear and aimed for strongly nonlinear systems. Multilayer perceptrons (MLPs) are supervised feedforward networks, which are mainly used for approximating nonlinear behaviour. Another popular network, the self-organising map (SOM) (Kohonen, 1995) based on unsupervised competitive learning, can be considered as a clustering method (Section 4.1). Radial basis networks (Chen et al., 1991) provide an interesting alternative as they can be used both as a clustering tool and a modelling tool.

Neurofuzzy methods. Neurofuzzy methods provide various techniques for generating fuzzy set systems, e.g. ANFIS method (Adaptive-Network-based Fuzzy Inference Systems) is a well-known neurofuzzy method which is suitable for tuning of membership functions (Jang, 1993). Partitioning clustering is used in this tuning (Section 4.1).

Linguistic equations. The nonlinear scaling transforms the nonlinear problem to a linear one. A LE model with several equations is represented as a matrix equation

$$AX + B = 0, \tag{1}$$

where the interaction matrix A contains all coefficients A_{ij} and the bias vector B all bias terms B_i. Each equation has from two to five variables.

6 Application cases

The variable selection and grouping methods described above have been tested in several applications.

Bioprocesses. Batch bioprocesses are difficult to model due to strong nonlinearities, dynamic behaviour, lack of complete understanding and unpredictable disturbances. The fed-batch fermentation model has three growth phases, each including three interactive models.A decision system based on fuzzy logic to provide smooth gradual changes between phases. (Juuso, 2019)

Continuous brewing. Brewing is based on ethanol fermentation but the most important aim is a balanced flavour

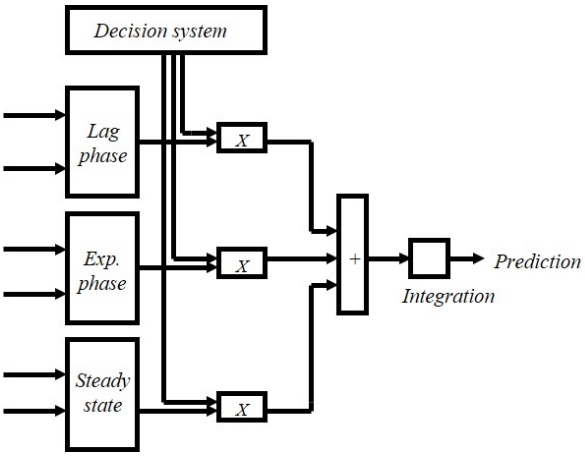

Figure 4. Subprocesses with interactions (Juuso, 2008).

not the highest possible ethanol yield. The desired traditional flavour is based on a balance of numerous compounds. Experiments with immobilized yeast were carried out in a pilot scale. The LE models were generated for five variable groups, each group contains two of the three control variables (air, CO_2 and flow) and three flavour compounds or attenuation. Fluctuations from the normal operation are detected to warn process operators. (Juuso and Kronlöf, 2005).

Condition monitoring. Reliability of operation, high quality, safety and environmental issues are increasingly important and machine condition monitoring enables reliable and economical way of action. Overhaul before a breakdown is in many cases more effective than run to failure. The earlier model-based approach discussed in (Juuso et al., 2008) was an interesting case for variable selection and grouping. New features and indicators have completely changed in this application. The methodologies have been tested in test rigs and industrial processes. (Juuso, 2017).

Paper machine. The ambition to increase the production of paper has made the paper machine runnability important. The paper web breaks when the strain on it is greater than the strength of paper. The machine can be run at the desired speed with the least possible number of breaks if the runnability is good. The web break sensitivity indicator was developed as a CBR application which combines LE models and fuzzy logic (Juuso et al., 1998; Ahola et al., 2003). The analyses are based on the online process data. There are several runnability categories, each including several case models defined by several equations based on up to five scaled variables. The final selection of the active cases, corresponding categories and the value of the break sensitivity are obtained by fuzzy logic (Juuso and Ahola, 2008).

Wastewater treatment. In the biological wastewater treatment, the model consists of three interactive models (Figure 5): the biomass quality obtained in Model B has a

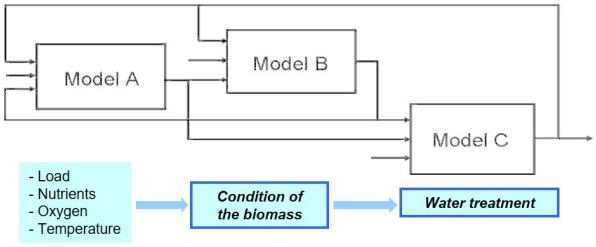

Figure 5. Subprocesses with interactions (Juuso, 2009).

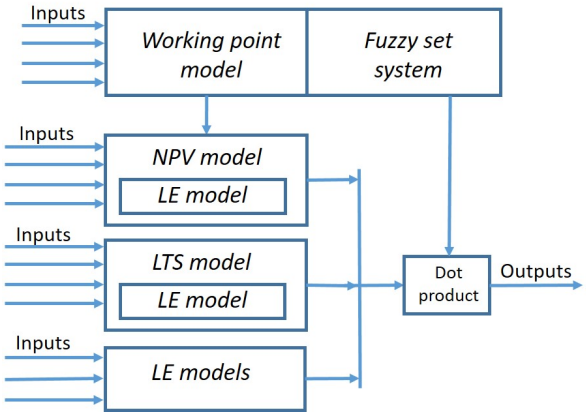

Figure 6. Multimodel LE system with a fuzzy decision module (Juuso, 2020b).

strong effect on the treatment (Model C). In Model A, the effects of the incoming wastewater and return sludge are combined. The chemical oxygen demand (COD) and suspended solids are used in the calculations. Nutrients, oxygen and temperature effect on both the biomass (Model B) and operating conditions Model C. The condition of the biomass has a key effect on the treatment performance (Juuso, 2009). Long periods of high load reduce the biomass quality which deteriorates the treatment performance and it takes time to get the good performance back.

7 Discussion

Variable selection was in all cases started with manual methods, continued with data analysis, and the final variable selection is based on generated alternative interactions assessed with domain expertise. The automatic analysis has several phases and alternative approaches. Correlation analysis is used for selecting interesting groups from the acceptable groups and principal components extend this analysis to the high dimensional systems. Several clustering methods are used for dividing the data sets into different operating areas. The nonlinear scaling provides new possibilities for these analysis methods as an essential part of the model-based analysis of interactions.

In these cases, the submodels are based on LE models Smooth operation and high quality products are the main goals of all these applications, and this can be achieved

by combining the LE models with fuzzy logic. Parametric LE models can make the model structure very compact (Figure 6). For diagnostics, the degree of membership calculated for the normal operation is a good indicator. This was used for the brewing case. The memberships of different faults are used in the condition monitoring. In the paper machine case, the case structure is highly complicated.

The data-based grouping provide high performance solutions but the phases, cases and interactions between models reduce the automatic analysis to the submodels. Also, the indirect measurements based on a set of measurements needs to be analyzed separately. New indicators can combine several measurements and several new features can be developed from individual signals. These case types which require decomposition are challenging for machine learning.

8 Conclusions and future studies

Variable selection and grouping are an essential part in the developing model-based applications. Domain expertise is used for removing useless combinations of variables. Data-based methods are divided into three classes: data analysis, decomposition and modelling. The model-based analysis is the final step. The originally linear methodologies were extended to nonlinear systems by using the nonlinear scaling approach. Applications are based on integrated approaches which combine all the techniques. The presented classification of methodologies was successfully used in the case studies.

Future studies are needed for applying these methodologies iteratively for the expanding heterogeneous data available in big data.

References

A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations and system approaches. *AICom - Artifical Intelligence Communications*, 7(1):39–59, 1994.

T. Ahola, H. Kumpula, and E. Juuso. Case based prediction of paper web break sensitivity. In *Proceedings of Eunite 2003 - European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, July 10-11, 2003, Oulu, Finland*, pages 161–167. Wissenschaftsverlag Mainz, Aachen, 2003.

Mina Aminghafari and Nathalie Cheze. Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis*, 50:2381–2398, 2006.

S. Äyrämö and T. Kärkkäinen. Introduction to partitioning-based clustering methods with a robust example. Reports of the Department of Mathematical Information Technology Series C. University of Jyväskylä, Software and Computational Engineering No. C. 1/2006. Jyväskylä, 2006.

R. Babuška. *Fuzzy Modeling and Identification*. Kluwer Academic Publisher, Boston, 1998.

B. Bakshi. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE Journal*, 44:1596–1610, 1998.

J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function*. Plenum Press, New York, 1981.

J. Cadima, J. Orestes Cerdeira, and M. Minhoto. Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis*, 47(2):225–236, 2004.

S. Chen, C.F.N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.

H. Cheng. Inference and decision making methods in fault diagnosis. A literature review. Helsinki University of Technology, Laboratory of Process Control and Automation. no. 9. Espoo, 2006.

D. Driankov, H. Hellendoorn, and M. Reinfrank. *An Introduction to Fuzzy Control*. Springer, Berlin, Germany, 1993.

P. Garcia-Martinez, M. Tejera, C. Ferreira, D. Lefebve, and H. H. Arsenault. Optical implementation of the weighted sliced orthogonal nonlinear generalized correlation for nonuniform illumination conditions. *Applied Optics*, 41 (32):6867–6873, 2002.

R. W. Gerlach, B. R. Kowalski, and H. O. A. Wold. Partial least squares modelling with latent variables. *Anal. Chim. Acta*, 112(4):417–421, 1979.

D. E. Gustafson and W. C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In *Proceedings of IEEE CDC, San Diego, CA, USA*, pages 761–766. IEEE Press, 1979.

I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47: 98–115, 2015. doi:10.1016/j.is.2014.07.006.

J.-S. R. Jang. ANFIS: Adaptive-Network-based Fuzzy Inference Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–685, 1993.

E. Juuso and S. Lahdelma. Intelligent scaling of features in fault diagnosis. In *7th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, CM 2010 - MFPT 2010, 22-24 June 2010, Stratford-upon-Avon, UK*, volume 2, pages 1358–1372, 2010. URL `www.scopus.com`.

E. Juuso, T. Ahola, K. Oinonen, and K. Leiviskä. Web break sensitivity indicator for a paper machine. In H.-J. Zimmermann, editor, *Proceedings of the 6th European Congress on Intelligent Techniques & Soft Computing -EUFIT'98, Aachen, September 7 - 10, 1998*, volume 3, pages 1549–1553, Aachen, 1998. Mainz.

E. Juuso, T. Ahola, and K. Leiviskä. *Variable selction and grouping. Report A 36, August 2008.* Control Engineering Laboratory, University of Oulu, Oulu, 2008.

E. K. Juuso. Linguistic equations for data analysis: FuzzEqu toolbox. In L. Yliniemi and E. Juuso, editors, *Proceedings of TOOLMET 2000 Symposium - Tool Environments and Development Methods for Intelligent Systems, Oulu, April 13-14, 2000*, pages 212–226, Oulu, 2000. Oulun yliopistopaino.

E. K. Juuso. Integration of intelligent systems in development of smart adaptive systems. *International Journal of Approximate Reasoning*, 35(3):307–337, 2004. doi:10.1016/j.ijar.2003.08.008.

E. K. Juuso. Intelligent dynamic simulation of a fed-batch enzyme fermentation process. In *Tenth International Conference on Computer Modelling and Simulation, EUROSIM/UKSim, Cambridge, UK, April 13, 2008.*, pages 301–306. The Institute of Electrical and Electronics Engineers IEEE, 2008. doi:10.1109/UKSIM.2008.133.

E. K. Juuso. Hybrid models in dynamic simulation of a biological water treatment process. In J. Kunovský, P. Hanácek, F. Zboril, Al-Dabass, and A. Abraham, editors, *Proceedings First International Conference on Computational Intelligence, Modelling and Simulation, 7- 9 September 2009, Brno, Czech Republik*, pages 30–35. IEEE Computer Society, 2009. doi:10.1109/CSSim.2009.52.

E. K. Juuso. Intelligent performance analysis with a natural language interface. *Management Systems in Production Engineering*, 25(3):168–175, 2017. doi:10.1515/mspe-2017-0025.

E. K. Juuso. Intelligent dynamic simulation of fed-batch fermentation processes. In E. Dahlquist, E. Juuso, B. Lie, and L. Eriksson, editors, *Proceedings of The 60th Conference on Simulation and Modelling (SIMS 60), 13-16, 2019, Västerås, Sweden*, number 170 in Linköping Electronic Conference Proceedings, pages 132–138. Linköping University Electronic Press, Linköpings universitet, 2019. doi:10.3384/ecp20170132.

E. K. Juuso. Expertise and uncertainty processing with nonlinear scaling and fuzzy systems for automation. *Open Engineering*, 10(1):712–720, 2020a. doi:10.1515/eng-2020-0080.

E. K. Juuso. Intelligent methodologies in recursive data-based modelling. In E. Juuso, B. Lie, E. Dahlquist, and J. Ruuska, editors, *Proceedings of The 61st Conference on Simulation and Modelling (SIMS 61), 22-24, 2020, Virtual Conference, Finland*, number 176 in Linköping Electronic Conference Proceedings, pages 466–474. Linköping University Electronic Press, Linköpings universitet, 2020b. doi:10.3384/ecp20176466.

E. K. Juuso and T. Ahola. Case-based detection of operating conditions in complex nonlinear systems. *IFAC Proceedings Volumes*, 41(2):11142–11147, 2008. doi:10.3182/20080706-5-KR-1001.01888.

E. K. Juuso and J. Kronlöf. Model-based monitoring of immobilized yeast fermentation using fuzzy logic and linguistic equations. *IFAC Proceedings Volumes*, 38(1):97–102, 2005. doi:10.3182/20050703-6-CZ-1902.02220.

M. Kano, S. Hasebe, I. Hashimoto, and H. Ohno. A new multivariate statistical process monitoring method using principal component analysis. *Computers and Chemical Engineering*, 20:1103–1113, 2001.

H. Karttunen. *Datan käsittely*. CSC-Tieteellinen laskenta, Yliopistopaino, Helsinki, 1994.

T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1995.

W. Ku, R. Storer, and C. Georgakis. Disturbance detection and isolation by dynamic principal components. *Chemometrics and Intelligent Laboratory Systems*, 30:179–196, 1995.

W. Li and S. Qin. Consistent dynamic PCA based on errors-in-variables subspace identification. *Journal of Process Control*, 11:661–678, 2001.

P. Nomikos and J. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40 (8):1361–1375, 1994.

J. Oton, P. Garcia-Martinez, I. Moreno, and J. Garcia. Phase joint transform sequential correlator for nonlinear binary correlations. *Optical Communications*, 245:113–124, 2005.

W. Pedrycz. An identification algorithm in fuzzy relational systems. *Fuzzy Sets and Systems*, 13(2):153–167, 1984.

D. Pyle. *Data preparation for data mining*. Morgan Kaufmann Publishers, San Francisco, 1999.

E. Ranta, H. Rita, and J. Kouki. *Biometria – Tilastotiedettä ekologeille*. Yliopistopaino, Helsinki, 1999.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61 (Supplement C):85 – 117, 2015. ISSN 0893-6080. doi:https://doi.org/10.1016/j.neunet.2014.09.003. URL `http://www.sciencedirect.com/science/article/pii/S0893608014002135`.

T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, 15(1):116–132, 1985.

Erik Vanhatalo, Murat Kulahci, and Bjarne Bergquist. On the structure of dynamic principal component analysis used in statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 167:1–11, 2017. doi:10.1016/j.chemolab.2017.05.016.

V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri. A review of process fault detection and diagnosis part i: Quantitative model-based methods. *Computers and Chemical Engineering*, 27:293–311, 2003.

M. Vermasvuori. Data-based methods and prior knowledge in process monitoring. A literature review. Helsinki University of Technology, Laboratory of Process Control and Automation. Report series no. 10. Espoo, 2006.

F. Westad, M. Hersleth, P. Lea, and H. Martens. Variable selection in PCA in sensory descriptive and consumer data. *Food Quality and Preference*, 14(5-6):463–472, 2003.

M. Zarzo and A Ferrer. Batch process diagnosis: PLS with variable selection versus block-wise PCR. *Chemometrics and Intelligent Laboratory Systems*, 73(1):15–27, 2004.