Process Development Based on Model Mining and Experiment Design Techniques

J. Abonyi
University of Pannonia, Department of Process Engineering, Veszprém, Hungary,
abonyij@fmt.uni.pannon.hu

Abstract— **Modern product and process development tools must meet wide-range of requirements. Minimizing the number of experiments while maximizing of the amount of information generated is only one aspect. Time restrictions and constraints imposed by the technology as well as specific customer demands are imperative boundary conditions which must be considered when planning an experiment. Furthermore, the experiment must yield reliable information regarding the feasibility of a project as early as possible. Commercial tools for statistical design of experiments alone cannot meet these requirements at an acceptable cost-benefit ratio. That is why the key of the proposed approach is to integrate the existing methods, models and information sources to explore useful knowledge. To explore and transfer all the useful knowledge needed to operate and optimize products, technologies and the business processes, the research of the applicant aimed the development of a novel methodology to integrate heterogeneous information sources and heterogeneous models. The proposed methodology can be referred as model mining, since it is based on the extraction and transformation of information not only from historical process data but also from different type of process models. The introduction of this novel concept requires the development of new algorithms and tools for model analysis, reduction and information integration. For this purpose fuzzy systems based modeling, clustering and visualization algorithms have been developed. To handle complex and contradictory goals a novel approach has been worked out based on visualization an interactive evolutionary algorithms. The aim of this paper is to provide an overview of these approaches.**

I. INTRODUCTION

As it is emphasized by the 7th Framework Programme, it is essential to develop new methods in order to speed up the transformation of the European industry and the economy and to increase industrial competitiveness and high quality products. During the last decade, a major shift has begun in the chemical and process industry, since there is an urgent need for new tools which are able to support the optimization of already operating and new production technologies. Approaches of this shift differ from company to company but one common feature is that it requires the intensive communication between design, manufacturing, marketing and management. Such communication should be centred on modelling and simulation, which integrates not only the whole product and process development chain, but all the process units, plants, and subdivisions of the company.

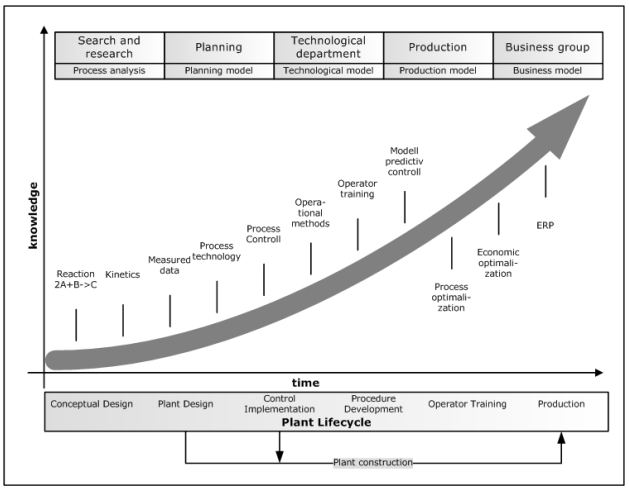

Figure 1. Knowledge management in "life-cycle modeling" and integrated modeling. Models are applied at every level of a technology and used to transfer information from conceptual design to the optimization of the production

Hence, engineers and directors of leading companies, e.g. DuPont and Dow Chemical, think that "model integrates the whole organization" [15]. They believe that the extensive use of models is the way that data, information, and knowledge should be conveyed from research to engineering, to manufacturing, and on to the business team. According to that, modelling and simulation will have a much greater role in bio-, chemical, and process engineering; it is prognosticated as a key feature of modern process maintenance in the future.

Officials of AspenTech and other companies dealing with simulation technologies talk about "life-cycle modelling" and integrated modelling technology, i.e. a model that is applied at every level of a technology (see Fig. 1 for the details of this concept).

However, the concept of life-cycle modelling is only a vision of how companies should operate in the future, in the 21st century. It is only a future concept, an ideal case. But instead of this there are only "model islands" for the time being. Isolated models are used for different and limited purposes on different levels of the technology (if they exist at all). These models are heterogeneous not only because they have different purposes, but also because they process data or information taken from different sources and apply different methods [13] which are often not compatible with other methods.

Information for the modelling and identification of the processes can be obtained from different sources: mechanistic knowledge obtained from first-principles (physics and chemistry) [15], empirical or expert knowledge, expressed as linguistic rules [13], measurement data, obtained during normal operation or from an experimental process [1]. Different modelling paradigms should be used for an efficient utilization of these different sources of information [1].

The aim of our research is to develop an approach, a methodology to integrate heterogeneous information sources and heterogeneous models to cover the whole process, technology and company. This is an extremely complex task considering the time scales for a production system (see also Fig. 2) [10, 14]. The key idea is the utilization of data mining and computational intelligence techniques since the synergistic integration of qualitative and quantitative models require tools to handle uncertain information (by fuzzy logic), systems complexity (by neural networks and evolutionary algorithms), and expert knowledge (by rule-based systems) (see Fig. 3).

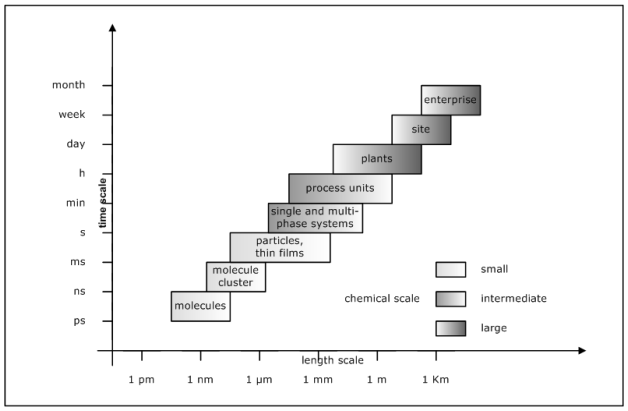

Figure 2. Time and scale levels in process engineering. Different scales define different engineering problems. The information gained in one scale could be useful in other scales, hence the research is focused on mining and transferring information from one scale to another

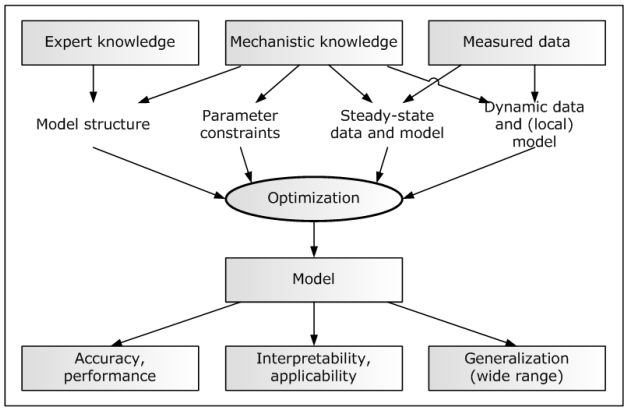

Figure 3. The proposed scheme combines expert and mechanistic knowledge and measured data in the form of rule-based systems, model structure, parameter constrains and local models. The model is optimized to ensure booth good prediction performance and interpretability.

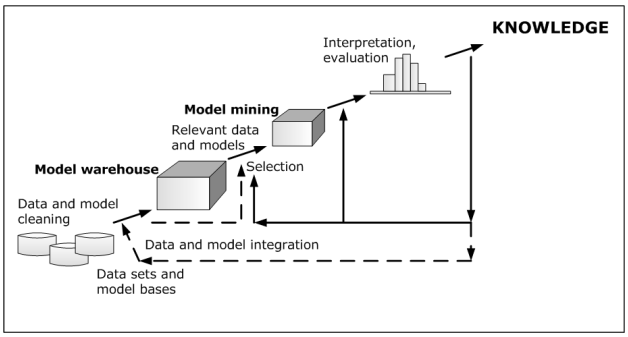

Figure 4. Scheme of model mining. The concept is similar to data mining, but in this schemes not only historical process data but different type of models are also analyzed.

II. Model Mining – A Novel Methodology

It is clear that a created global model cannot be developed by the improvement and refinement of existing models. The key of the proposed approach is to *integrate* the existing models and information sources to explore useful knowledge. The whole process can be called model mining. This type of knowledge discovery is relatively new [23], its importance has only been recognized in the automated detection and mining of atmospheric phenomena relationships by the Global Modelling and Assimilation Office at NASA. Hence, our research can be considered as the first concentrated attempt to develop a new methodology for model mining. According to our current ideas, the proposed approach, depicted in the Fig 4., consists of the following steps.

1. Developing and understanding of the application domain and the relevant prior knowledge, and identifying the problem.

2. Creating and pre-processing the target set.

This phase starts with activities in order to get familiar with the models and data, to identify quality problems, to get first insights into the data and models to detect interesting subsets, to form hypotheses for hidden information. To support this step, the utilized models should be as transparent as possible, since model transparency allows the user to effectively combine different types of information, namely linguistic knowledge, first-principle knowledge and information extracted from data. For that purpose, a model warehouse should be developed in which the different information sources and the developed models are stored in an easily retrievable way. To achieve this goal, there is a need for a language and clear definitions of metadata that can be used to describe the models and different information sources. Such metadata can be represented by an XML based system, similar to the Predictive Model Markup Language (PMML, www.dmg.org) used to store and transfer data mining models.

3. Model Mining

The ultimate goal of the whole process is to extract potentially useful knowledge which could not be explored from one single model or database, but only from the integrated information sources and models in the model

warehouse. The goals of model mining are achieved via the following tasks:

3.a *Model Transformation and Reduction*: the different kinds of information presented by models can be used to transform or reduce other models in order to simplify them, to make them more precise and/or robust, or to expand their operational range (e.g. to improve the extrapolation capability of a black-box model using a priori knowledge). Some of the computational intelligence models lend themselves to be transformed into other model structures that allow information transfer between different models (e.g. decision trees can be mapped into feedforward neural networks, or radial basis functions are functionally equivalent to fuzzy inference systems).

3.b *Model Fusion:* the integration of the information content of different models is the key issue of the research. During this process it should be kept in mind in what range the particular models are valid and what process should be applied or developed to get the global result (e.g. voting systems for models on the same level of the technology). Information can be stored within nominal, ordinal, ratio or fuzzy data; images and pictures, multivariate time series or documents can also be information sources [6]. Besides of that, several type of models should be treated, e.g. graphs, decision trees, neural networks, rule based systems, etc., which need different types of reduction methods (e.g. spanning trees for graphs, pruning for decision trees, rule base reduction for fuzzy models etc.). According to our recent research results and know-how such models and knowledge can be integrated by fuzzy systems (extended Takagi-Sugeno fuzzy models).

3.c *Visualization* is a very promising field because it makes possible merging the best of computer (calculation) capabilities with the best of human perception (cognitive) abilities [23]. There are several types of methods to visualize multidimensional data [22]. However, the visualization of the results of the knowledge discovery process is much more complex than the visualization of multivariate data. The world-leading companies like AT&T and Microsoft are dealing with these problems. The visualization of clustering results have been also dealt with by our research group; these methods could be improved, modified and extended, and new ones would be developed according to model mining purposes. Visualization methods are based on similarity measures (for multivariate data it depends on a particular distance measure; for fuzzy sets it is based on fuzzy set similarity measures etc.). For model mining purposes, there is a clear need to define several types of similarity measures according to the heterogeneous information sources (similarity of models, images, graphs etc.). The existing methods are commonly used to project data or other types of information into two dimensions.

3.d *Choosing and application of the model mining algorithm(s):* Selecting algorithms for searching for patterns. It is essential to involve the user or the experts into this key step because it is often very difficult to formalize the complex and many times contradictory goals. One possible solution is to increase the degree of interactivity. For that purpose the visualization of partial results is needed [28].

4. Interpreting and application of the mined patterns

Whereas the "knowledge worker" judges the success of the application of modelling and discovery techniques more technically, she or he contacts domain experts later in order to discuss the results and their applicability. Based on the results it can be necessary to return to any of steps 1-3 for further iteration.

III. Relation to Experiment and Quality By Design Techniques

The previous section presented an iterative model development scheme. This scheme can be used to support product and process development. Of course this requires the application of sophisticated development methodologies. Beside the concept of Model Mining, the key idea of this paper is that Quality by Design and experiment design methodologies are the most suitable that can be incorporated to the model and data mining techniques. Quality by Design (QbD) is a concept first outlined by well-known quality expert Joseph M. Juran in various publications, most notably Juran on Quality by Design. Juran believed that quality could be planned, and that most quality crises and problems relate to the way in which quality was planned in the first place. While Quality by Design principles have been used to advance product and process quality in every industry, and particularly the automotive industry, they have most recently been adopted by the U.S. Food and Drug Administration (FDA) as a vehicle for the transformation of how drugs are discovered, developed, and commercially manufactured. According to the FDA draft guidance, the desired state of pharmaceutical manufacturing is that:

- product quality and performance are ensured through the design of effective and efficient manufacturing processes
- product and process specifications are based on a mechanistic understanding of how process factors affect product performance
- quality assurance is continuous and real time
- relevant regulatory policies and procedures are tailored to accommodate the most current level of scientific knowledge
- risk-based regulatory approaches recognize both the level of scientific understanding and the capability of process control related to product quality and performance

The key idea of this approach is that once the properties of the product components are understood, the processing variables that control the relevant properties must be identified. Identification of these variables necessarily requires a multivariate approach. The required methodology should involve the design of manufacturing processes based on a thorough scientific understanding of the properties of the product at critical points throughout manufacturing. Process models play important role in this computer aided process engineering. The accuracy of the resulted models (estimated parameters) and the success of a development action largely depends on the information content of the experiment designed to obtain useful information for the identification of the unknown model parameters. Using the tools of Optimal Experiment Design (OED) we can maximize the confidence on the model parameters. The essence of the method is to determine those input trajectories which has the potential to generate system output trajectories with high information content. OED could be easily inserted to the classical scheme of experiment design and might be used integrated with the process (see Fig. 5).

Statistical design of experiments is based on, among other things, the application of regression models. It is particularly well suited for the design of experiments for the identification of key variables at the beginning of a study (screening) when there is little or no prior knowledge. If the parameter space is high dimensional, neural networks rather than regression models should be used in the modeling phase to reduce the number of additional experiments. Regression models are much too expensive for optimization in high dimensional parameter spaces with discrete input variables. In such cases, evolutionary experiment design based on genetic algorithms can be used. The efficiency of these processes can be further improved if data mining methods are applied to data analysis. The significance of results, which are often difficult to interpret, can be improved and data mining can identify new correlations across several process levels with no prior knowledge or reveal hidden information. So far, this form of integrated design of experiments has been applied successfully to catalyst screening (high throughput screening) and product quality optimization.

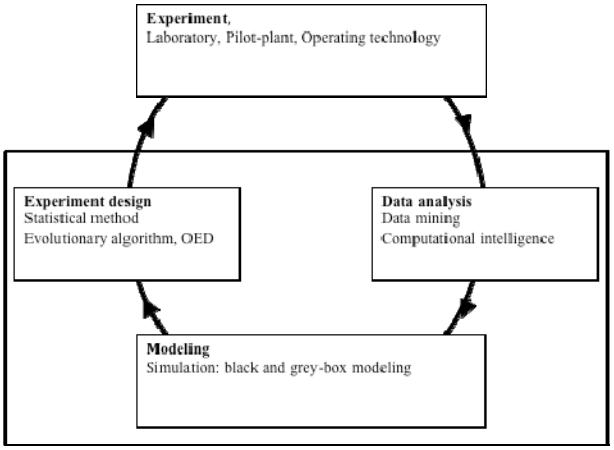

Figure 5. Scheme of data and model mining based experiment design

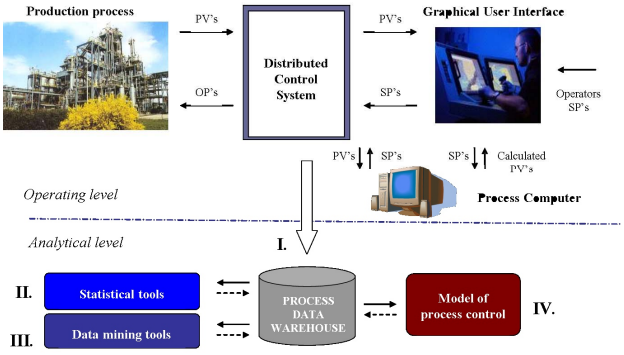

Figure 6. Integrated information system for process analysis. The proposed analytical level consists the models of the process and control systems. Stored data can be used to re-simulate the passed critical situations of the operation and the measured and the calculated data can be analyzed by statistical and data mining tools.

IV. RESULTS

The research group of the author developed a novel know-how for the design and implementation of process data and model warehouses that integrates plant-wide information, where integration means information, location, application and time integrity.

The process data warehouse contains non-violate, consistent and preprocessed historical process data and works independently from other databases operating at the level of the control system. To extract information from historical process data stored in the process data warehouse tools for data mining and exploratory data analysis have been developed (see Figure 6).

In case of complex production processes it is often not sufficient to analyze only input-output data for process monitoring purposes. The reasons may be that historical process data alone do not have enough information content, it can be incomplete, not measured frequently or not at regular intervals. In these cases it is important to obtain information about state variables; therefore (nonlinear) state estimation algorithm is needed. This phenomenon has been proved experimentally at the time-series segmentation based process monitoring, where the result of the segmentation was much more reliable when the estimated state variables or the error covariance matrices computed by the state estimation algorithm have been also utilized by the segmentation algorithms.

When models of process and control systems are integrated to a process Data Warehouse the resulted structure support engineering tasks related to analysis of system performance, process optimization, operator training (OTS), reverse engineering, and form decision support (DSS) systems [9].

1. Constrained (grey-box) identification of operating regime based models enable the effective use of prior knowledge and results in a robust modeling approach

Fuzzy model identification is an effective tool for the approximation of uncertain nonlinear systems on the basis of measured data. Among the different fuzzy modeling techniques, the model introduced by Takagi and Sugeno has attracted most attention. The Takagi-Sugeno (TS) model formed by logical rules; that consists of a fuzzy antecedent and a mathematical function as consequent part. The construction of a TS model is usually done in two steps. In the first step, the fuzzy sets in the rule antecedents are determined that partition the input space into a number of fuzzy regions. In the second step, the rule consequents are determined which means identification of (usually linear) models. TS fuzzy model identification is a complex task; there are non-trivial problems as follows: (1) how to automatically partition the input space, (2) how many fuzzy rules are really needed for properly approximating an unknown nonlinear system, and (3) how to construct a fuzzy system from data examples automatically. These problems can be partially solved by the recent developments of fuzzy systems.

A novel approach to data-driven identification of Takagi-Sugeno fuzzy models have been worked out. It allows to translate prior knowledge about the process (including stability, minimal or maximal static gain and settling time) into constraints on the model parameters. This approach has been successfully applied in (adaptive) model predictive control. The advantage of the algorithm is that by constraining the parameters of the local linear models, it is possible to speed up the adaptation and avoid unrealistic model parameters that could result bad control performance. This grey box fuzzy model approach allows the development of TS models also in cases where little experimental data are available. It has been shown that fuzzy models built on the basis of data combined with prior knowledge perform better in control than models obtained from data only. [1, 2, 3, 4]

2. Model based cluster analysis is a useful tool to solve process engineering problems

Clustering, as a special area of data mining is, one of the most commonly used methods for discovering the hidden structure of the considered data set. The main goal of clustering is to divide objects into well separated groups in a way that objects lying in the same group are more similar to each another than to objects in other groups. In the literature several clustering and visualization methods can be found. However, due to the huge variety of problems and data sets, it is a difficult challenge to find a powerful method that is adequate for all problems. [11, 12, 17].

Segmentation of multivariate time-series, and application of state estimation algorithm to detect changes in the state of the system and for product quality estimation [5].

3. Fuzzy clustering for nonlinear regression

Takagi-Sugeno (TS) models formed by logical rules consist of a fuzzy antecedent and a mathematical function as consequent part. The construction of a TS model is usually done in two steps. In the first step, the fuzzy sets in the rule antecedents are determined that partition the input space into a number of fuzzy regions. In the second step, the rule consequents are determined which means identification of (usually linear) models. TS fuzzy model identification is a complex task; there are non-trivial problems as follows: (1) how to automatically partition the input space, (2) how many fuzzy rules are really needed for properly approximating an unknown nonlinear system, and (3) how to construct a fuzzy system from data examples automatically. These problems can be partially solved by the recent developments of fuzzy systems.

It is recognized that fuzzy clustering algorithms are able to automatically divide the input space, and developed clustering algorithm that fits a local models beside the clustering simultaneously [6, 7]. Transparency of the model is enhanced by the tree representation. The number of rules can be given by the user or it can be identified by validation. Comparing with well-known methods it can be determined that the developed method gives the most transparent results at similar accuracy to these methods.

4. Supervised clustering for classifier induction

The classical fuzzy classifier consists of rules each one describing one of the classes. In this paper a new fuzzy model structure is proposed where each rule can represent more than one classes with different probabilities. The obtained classifier can be considered as an extension of the quadratic Bayes classifier that utilizes mixture of models for estimating the class conditional densities. A supervised clustering algorithm has been worked out for the identification of this fuzzy model [8]. The relevant input variables of the fuzzy classifier have been selected based on the analysis of the clusters by Fisher's interclass separability criteria. This new approach is applied to the well-known wine and Wisconsin Breast Cancer classification problems.

The proposed method can be used for the discretization of continuous features to form efficient fuzzy decision tree based classifiers. The resulted fuzzy classifiers are and well interpretable while the accuracy is still comparable to the best results reported in the literature.

5. Rule-based systems for the analysis of process data

Fuzzy association rule mining is applicable for process data analysis.

(a) Fuzzy association rule mining is applicable for feature and model structure selection. [20]

(b) Compact and accurate fuzzy classifiers can be constructed by fuzzy association rule mining. [21]

(c) Visual post- analysis of mined association rules can be done by Sammon Mapping and a novel similarity measure [19].

6. Multi- and conflicting- objective process optimization problems can be effectively solved by interactive optimization

Process optimization problems often lead to multi-objective problems where optimization goals are non-commensurable and they are in conflict with each other. In such cases, the common approach, namely the application of a quantitative cost-function, may be very difficult or pointless. For these problems, we developed a method that handles these problems by introducing a human user into the evaluation procedure. Namely, the proposed method uses the expert knowledge directly in the optimization procedure [18].

V. CONCLUSIONS

In this paper a novel model mining framework has been presented based on machine learning, data mining and computational intelligence techniques. The framework can be used to support experiment design and quality by design methodologies, so improved quality and efficiency are expected:

- reduction of cycle times
- prevention of reject product and waste
- increased use of automation
- facilitation of continuous processing using small-scale equipment, resulting in improved energy and material use and increased capacity

ACKNOWLEDGEMENT

The support from the TAMOP-4.2.2-08/1/2008-0018 (Livable environment and healthier people – Bioinnovation and Green Technology research at the University of Pannonia, MK/2) project is gratefully acknowledged

REFERENCES

[1] J. Abonyi: Fuzzy Model Identification for Control, Birkhauser Boston, 2003

[2] J. Abonyi, R. Babuska, H. Verbruggen, F. Szeifert, "Incorporating Prior Knowledge in Fuzzy Model Identification", Int. Journal of Systems Science, 31(5), 657-667, 2000, IF 0.268

[3] J. Abonyi, R. Babuska, F. Szeifert, "Fuzzy modeling with multivariate membership functions: Gray-box identification and control design", IEEE Systems, Man and Cybernetics, Part B: Cybernetics, 31 (5), pp. 755-767 IF 0.789, 2001

[4] J. Abonyi and R. Babuska and M. Ayala Botto and F. Szeifert and L. Nagy, "Identification and Control of Nonlinear Systems Using Fuzzy Hammerstein Models", Industrial and Engineering Chemistry Research, 39, 4302-4314, 2000., IF 1.294

[5] J. Abonyi, B. Feil, S. Nemeth, and P. Arva, "Modified Gath-Geva Clustering for Fuzzy Segmentation of Multivariate Time-series, Fuzzy Sets and Systems", Data Mining Special Issue, 2005, 149 39-56, IF 0.734

[6] J. Abonyi, B. Feil, Cluster Analysis for Data Mining and System Identification, Birkhauser, 2007, 300 pages

[7] J. Abonyi, F. Szeifert and R. Babuska, "Modified Gath-Geva Fuzzy Clustering for Identification of Takagi-Sugeno Fuzzy Models", IEEE Trans. on Systems, Man and Cybernetics, Part B, 612-621, Oct, 2002, IF 0.63

[8] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers", Pattern Recognition Letters, 24(14) 2195-2207, October 2003, IF: 0.809

[9] B. Balasko, S. Nemeth, J. Abonyi, "Application of integrated process and control system model for simulation and improvement of an operating technology", 6th European Congress of Chemical Engineers, Copenhagen, September 16-20, 2007

[10] J.C. Charpentier, "Four main objectives for the future of chemical and process engineering mainly concerned by the science and technologies of new materials production", Chemical Engineering Journal 107:3–17, 2005

[11] B. Feil, J. Abonyi, F. Szeifert, "Model Order Selection of Nonlinear Input-Output Models – A Clustering Based Approach", Journal of Process Control, Volume 14, Issue 6, 593-602, IF: 1.241 2004.

[12] B. Feil, B. Balasko, J. Abonyi, "Visualization of Fuzzy Clusters by Fuzzy Sammon Mapping Projection – Application to the Analysis of Phase Space Trajectories", special issue of "Soft Computing" on "Soft Computing for Information Mining", 11, (5), 479-488,. 2006 (IF: 0.33)

[13] N. Gershenfeld: "The nature of mathematical modeling", Cambridge University Press, 1999

[14] I.E. Grossmann, AE. Westerberg, "Research challenges in process systems engineering", AIChE J. 9:1700–1703, 2000

[15] JH. Krieger, "Process Simulation Seen As Pivotal In Corporate Information Flow", Chemical & Engineering News, March 27, 1995

[16] J. Madár, J. Abonyi, H. Roubos, F. Szeifert, "Incorporating Prior Knowledge in Cubic Spline Approximation - Application to the Identification of Reaction Kinetic Models", Industrial and Engineering Chemistry Research, 42, 4043-4049, 2003, IF: 1.252

[17] J. Madar, J. Abonyi, F. Szeifert, "Genetic Programming for the Identification of Nonlinear Input-Output Models", Industrial and Engineering Chemistry Research, 44, 3178-3186, 2005, IF: 1.29

[18] J. Madár, J. Abonyi, F. Szeifert, "Interactive Evolutionary Computation in Process Engineering", Computers & Chemical Engineering, Volume 29, Issue 7, 15 June 2005, Pages 1591-1597, IF: 1.678

[19] Peter Matyus, Pach F. Peter, Janos Abonyi, Attila Gyenesei, "Visualization of Fuzzy Association Rules Representing High-Dimensional Problems", 11th IPMU International Conference. July 2-7, 2006 Paris

[20] Pach F.P., Gyenesei A., Abonyi J., MOSSFARM: Model Structure Selection by Fuzzy Association Rule Mining, Journal of Intelligent and Fuzzy Systems, accepted

[21] Pach F.P., Gyenesei A., Abonyi J.,"Compact fuzzy association rule-based classifier", Expert systems with applications, 2008,34,4,2406-2416

[22] JB. Tenenbaum, V. Silva, JC. Langford, " A global geometric framework for nonlinear dimensionality reduction", Science, 290:2319–2323, 2000

[23] J. Valdes, A. Barton: "Virtual Reality Spaces: Visual Data Mining with a Hybrid Computational Intelligence Tool", NRC/ERB-1137 (NRC 48501), 2006

[24] A. Vathy-Fogarassy, A. Kiss, J. Abonyi,"Improvement of Jarvis-Patrick clustering based on fuzzy similarity",2007, Vol:4578 Lecture Notes in Computer Science, Applications of Fuzzy Sets Theory, 195-202

[25] A. Vathy-Fogarassy, A. Kiss, J. Abonyi, "Hybrid Minimal Spanning Tree and Mixture of Gaussians based Clustering Algorithm", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Foundations of Information and Knowledge Systems, 3861 LNCS, pp. 313-330 (IF: 0.513)

[26] A. Vathy-Fogarassy, A. Werner-Stark, B. Gal, J. Abonyi, "Visualization of topology representing networks", Intelligent Data Engineering and Automated Learning - IDEAL 2007, Lecture Notes in Computer Science,4881, 557-566, 2007

[27] A. Vathy-Fogarassy, A. Kiss, J. Abonyi, "Topology Representing Network Map-A new Tool for Visualization of High-Dimensional Data", Transactions on Computational Science, Lecture Notes in Computer Science,4750/2008, 61-84, 2008

[28] L. Yan, IDL. Bogle, "A visualization method for operating optimization", Computers and Chemical Engineering 31:808–814, 2007