DATA MINING CAN HELP HUMAN SUPERVISORY CONTROL OF CRITICAL COMPLEX INDUSTRIAL SYSTEMS

K. Li*, S. Thompson†, J. Peng†, Xinglin Chen‡

*Corresponding author, Intelligent Systems and Control Group
School of Electrical and Electronic Engineering, Queen's University Belfast
Ashby Building, Stranmillis Rd., Belfast BT9 5AH, UK
E-mail: K.Li@qub.ac.uk; Fax: +44 (0)28 90 667023
†School of Mechanical and Manufacturing Engineering
Queen's University Belfast, Belfast BT9 5AH, UK
‡School of Astronautics, Harbin Institute of Technology, Harbin, China

Keywords: Data overload, Data mining, Human supervisory control, Critical systems, Power plant emissions

Abstract

The problem of data overload in supervisory control of critical complex industrial systems is addressed. As a solution data mining has been shown as an effective tool to help human supervisory control of a coal-fired power generation plant in reducing pollutant emissions.

1 Introduction

Rapid progress in technology, together with the scale enlargement, social, financial and political stimuli have changed the scope of society. Systems are becoming more complex and consequently dependent on automation increasingly. Increased complexity naturally gives rise to the fear that even the smallest error may cause a disaster [9]. The mechanism underlying such fears may be compared with the butterfly effect described in a book on chaos [7]: small numerical errors in the initial conditions of a deterministic weather model can produce massive changes in the simulated predictions.

Industrial processes including many critical systems such as power generation plants are also becoming more complex with an increasing number of interactions between different domains and plants. For example, very tight links can exist between plants owned and operated by different companies. Consequently, control centres can control several plants. This centralised control is seen as advantageous, since many apparently simple functions can be combined and executed by a small number of operators. In addition, concentrating tasks can produce many similar tasks, of which some may be independent of the normal operating modes. Examples of such tasks are plant optimisation, maintenance and administration, which may all be executed by specialists outside of the control room. Since the control room operators are only in control of the plant during start-up, shutdown and malfunction, the danger is that the *situation awareness* of those operators who remain in the control room is reduced due to the following reasons:

1) The frequency of system malfunctions is extremely low for highly automated critical systems. On the other hand, the workloads between normal and abnormal operations are also extremely distinct.

2) Operators are increasingly required to cope with situations that are outside their immediate base of expertise on a short deadline due to the organizational trend of reducing operational staffing and expertise during nominal situations.

3) As the system complexity increases, data overload poses a generic and difficult problem for system operators and operators may find it difficult to cope with a situation that generates an avalanche of electronic data and information.

Although the advances in artificial intelligence, computer graphics, or electronic connectivity etc. promise to help people better understand and manage wider range of activities more efficiently and effectively, from financial analysis to monitoring and control of critical systems such as various power generation plants and space missions, our ability to interpret this avalanche of data has expanded much more slowly [14]. As a matter of fact, examinations across different engineering fields show that engineering practitioners are always bombarded with computer-processed data, particularly when anomalies occur. Operators may from time to time find themselves lost in massive networks of computer based displays, options, modes, and their associated engineering contexts.

The data overload problem in human supervisory control of complex engineering systems can be characterised in different ways 1) as a clutter problem where there is too much data on the screen. This problem can be solved by reducing the number of data units on each screen and by grouping data units of the same nature; 2) as a workload bottleneck where there is too much to do in the time available. This problem can be solved by using automation or other techniques to disperse the activities of the operators in the case of burst of workload; 3) as a problem in extracting significant patterns or identifying the significance of data from data avalanche generated during the operation when it is unknown in advance what data is important. Data mining techniques can solve this problem.

Data mining is defined as the process of extracting patterns as well as predicting (previously unknown) trends from large quantities of data by posing (automatically) repeated queries [2]. It is an information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of various techniques, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. While various forms of data mining have existed for quite a while, it is only during the past decade that data mining has emerged as a technology area for a wide range of applications. The combination of statistic analysis, machine learning and database management has resulted in the emerging technology area - data mining [1,2,5,6].

For data mining to be effective, several technologies have to work together. First of all, statistical analysis and machine-learning techniques have to be applied successfully to databases to extract patterns and to predict trends. Visualization techniques are important to provide visual understanding of data, patterns and trends and subsequently guide the user in carrying out further data mining. Data warehousing is a critical technology for organizing and cleaning the data to prepare for mining. Parallel processing techniques provide important enabling technology to speed up the mining process for large-scale data sets. Network-computing infrastructures are an important consideration especially for distributed data mining. That is, various technologies have to be integrated to carry out successful data mining, leading to a need for standards [6]. As a result of the developments in data mining during the past decade, numerous commercial products and research prototypes have been developed. A consortium of data mining vendors and early adopters of data mining technology, through a European Commission funded effort, have developed the Cross-Industry Standard Process for Data Mining. This is a hierarchical process model that breaks the data mining process into several phases, each with a variety of tasks [1], 1) Business understanding; 2) Data understanding; 3) Data preparation; 4) Modelling; 5) Evaluation; 6) Deployment.

In this paper, data mining will be used to help operator to ease the data overload and to control critical systems in normal and abnormal circumstances.

2 Using data mining to help human supervisory control of power plants

2.1 The complexity in human supervisory control of coal-fired power generation plants

Operation and control of pulverised fuel thermal power plants (TPPs) is getting increasingly complex due to several reasons.

1) More stringent legislation on pollutant emissions. The contribution of solid fuels in TPP has increased since 1995. It is shown that 32% of the world's power generation is based on inputs of hard coal and combustion of pulverized coal in large power station boilers accounts for over 50% of total world coal consumption. During the combustion process, various pollutants are produced such as oxides of carbon (COx), oxides of sulphur (SOx), oxides of nitrogen (NO_x) and particulates. SO_2 and NO_x can cause acid rain and CO_2 is the most important greenhouse gas responsible for climate changes. As more stringent emission limitations have been legislated for power generation stations worldwide, the limitation of current instrumentation and control technology for coal-fired TPPs are being exposed. In order to meet tighter legislations, more complex instrumentation and control systems have been installed in new and modified TPPs worldwide.

2) Changing of coal supply resources. Because of the international trade in coals the quality and other properties of the coals used in TPPs will vary frequently. This strategy is followed for economic reason, but the wide variety in characteristics of the different coals can lead to unpredictable milling and combustion behaviour. Therefore the whole combustion process needs to be monitored more intensively to determine and control the effects of blends on emissions and combustion efficiency.

3) The changing of electricity demands. Another force for introducing more complex control systems is the changing electricity demands. In many European countries as well as other countries, the increase of alternative power generation means that the majority of coal-fired TPPs are no longer operated on base load (at a high steady power output), instead most coal-fired TPPs operate at the specific load following a pattern of two-or-more shift operation, i.e. a plant will operate at high output (~80%) for daylight hours and low output at night. When the electricity demands for a TPP vary there must be corresponding changes in fuel, air and water inputs as well as burner operation patterns. To maintain close control over boiler conditions, currently separated loop operation and control systems are used, such as boiler air control, boiler steam temperature control and boiler feed water control, etc.

It is clear that the above listed reasons for increase of system complexity will affect the plant operations, and eventually more data will be generated and has to be coped with in human supervisory control. Human operators in conventional coal-fired power generation boilers play an essential role in plant operations. In addition to plant start-up, shut-down, monitoring and intervention; plant operators also adjust various parameters related to steam pressure and generation control, combustion process control, steam temperature control, and water-steam circuit control. Particularly to the combustion process control, operators are required to adjust control parameters for pulverised coal-distribution system, boiler/burner system and combustion emission as coal and electricity demand change over time. These operational demands comprise many peculiarities which the operators should know about and obviously they need assistance tools to handle the data overload problem and an advisory system to include them into operations.

In this paper our focus is to use data mining techniques to help human supervisory control of pollutant emissions. In particular, software is designed to help operators to modelling the relations of pollutant emissions with manipulated variables, which then may be used to extract useful operation patterns for human operators.

2.2 Pollutant emission reduction and requirements for data mining software in human supervisory control

The principal pollutants emitted from a TPP are nitrogen oxides (NO_x), sulphur oxides (SO_x, carbon dioxide ($CO_{2)}$), and particulates (soot, flyash, and aerosols) [10]. Among these, emission of dust can be eliminated or reduced by for instance electrostatic precipitators or filters. The best way to reduce CO_2 emissions is to improve the efficiency of coal combustion process. Sulphur oxides can be removed by a flue gas desulphurisation (FGD) system which is extensively used throughout the world. Nitrogen oxides can form the photochemical smog and damage the ozone layer. It can also cause acid rain. NO_x can be reduced by operational modifications or by installing "low NO_x burners". Although low NO_x burners are usually sufficient to achieve the required target, it is often at the expense of other important operational parameters such as incomplete combustion, steam temperature and boiler performance. The current legislations require that power plants make significant reductions in pollutant emissions especially in NO_x emissions (Holmes and Mayes, 1994), and these restrictions are going to be more stringent. Therefore reducing the environmental impact of TPPs by NO_x is one of the important targets that will have to be achieved by introducing more sophisticated operation and control systems [8,11].

The methods to reduce NO_x emission in coal-fired TPPs can be classified into two categories, namely primary or combustion modification based technologies, and secondary or flue gas treatment based technologies [4]. Those based on combustion modification achieve reduction of NO_x by limiting the flame temperature or the availability of oxygen in the flame. For coal-fired power plants that are not installed with low NO_x burners, it is believed that the balancing of fuel and air flow to individual burners is usually a first step in controlling NO_x emission, and for some plants can achieve up to a 20% reduction in NO_x formation. The further NO_x reduction for such plants can be achieved by adopting a sort of "burner out of service" method, in which fuel and air to different burners is increased, reduced or turned off completely to create local fuel rich and fuel lean zones. Similarly burners can also be operated with different stoichiometries at different furnace levels to create a form of global furnace staging that can achieve up to 30% NO_x reduction. These operations require careful control in order to maintain the thermal efficiency. Further reduction of NO_x emissions should focus on the overall performance of combustion boiler through implementation of advisory system for human operators (which requires good models of the underlying processes), or even advanced control systems.

In general, the parameters that will affect the combustion process in pulverised fuel boilers include:

- Primary air to coal ratio
- Secondary air distribution
- Burner tilting position for tangentially boilers
- Mill firing pattern

The methods of changing the above parameters through operation are:

- Primary air/ fuel ratio can be changed through controlling the coal feed rate and primary air fan speed
- Secondary air distribution can be changed through damper positions
- Burner tilting positions are normally changed by altering each individual burner
- Mill firing patterns are changed through altering mill feed rate to bunks of burners

The change of boiler parameters cannot be made freely due the to safety consideration. For example, loss of individual flames can lead to unburned fuel into the boiler, causing accumulation of potential fuel/air mixture that can be very explosive. Irregular mill firing patterns can cause instability in the flame ignition plane, therefore affect the whole efficiency of the combustion process. In coal-fired power generation plants the NO_x emission reduction can be achieved by human operators through the help of data mining software, which may consist of two stages. The first stage is based on a priori knowledge and/or a posterior knowledge to catch the relationship between the plants operational inputs and the NO_x output. In the second stage some form of constrained optimisation is used to compute the optimal operation patterns in order to minimise the NO_x output while maintaining or increasing the combustion efficiency. These values are then presented to the operator (open-loop mode) as the operational references.

As one of the key steps in data mining is modelling. The requirement for pollutants particularly NO_x emission models is that it should be simple enough to implement in real-time but also convey rich operational information to derive operation patterns [10]. In order to achieve the above purpose, the data mining software will have the following characteristics,

1) Flexibility. The data mining software should let the operator to identify significant data visually and choose the data segment of his interest flexibly for data mining.

2) Right modelling tool. Given the complexity of NOx formation mechanism, it is essential to choose the right modelling tool to capture the relations between NOx emissions with the operation inputs.

3) Information hiding. To avoid data overload, unnecessary processing information need to hide from the operator and only meaningful information is represented.

3 The data mining software for human supervisory control of power plant emissions

3.1 The NOx emission modelling

As a special example of grey-box modelling method [3], fundamental grey-box modelling is used to model a complex engineering system where the underlying mechanisms are either too complex to build a simple model or such knowledge is only partially known a priori [10]. The model structure used by fundamental grey-box method is a non-linear regression model given that the terms are well defined a priori:

$$\text{Model:} \quad y(t) = \sum_{i=0}^{p} \theta_i \varphi_i(t) + \varepsilon(t) \qquad (1)$$

where,

$$\begin{cases} \varphi_0(t) = 1 \\ \varphi_i(t) = \varphi_i(y(t-k_y), u_j(t-d-k_{u_j})), i = 1, \dots, p \end{cases}$$

and y and u_j $j=1,2,..,m$ are NOx output and operational inputs, θ_i is the model coefficient, $\varepsilon(t)$ is a white noise series, $\varphi_i(t)$ denotes model terms or fundamental elements, and k_y and k_{uj} are time delays for inputs and output.

In the fundamental grey-box modelling approach, fundamental elements are appropriate non-linear functions that are uniquely identifiable from 'a priori' engineering knowledge and they could happen to be such functions as exponential, power, trigonometric, rational, etc. It is interesting to notice that for some function types listed above, the differentiation or integration would result in a function belonging to the same class, and superposition, subtraction or multiplication of these functions will also result in the same function type, though the parameters would be different. This property allows us to separate these functions from ODEs and PDEs, and through appropriate composition and recombination of these functions, a simplified model could then be produced to represent the original system more realistically in terms of predicting unseen data/phenomena. As these fundamental elements together with associated parameters reflect physical reality in one way or another therefore are also useful in helping system operators and control engineers to interpret and gain some physical insight into the system under control. For example, in modelling NOx emissions in coal-fired power generation plants, $\varphi_1 = (x + b_1)^{c_1}$ and $\varphi_2 = e^{(c_2/(x+b_2))}$ are two type of fundamental elements [10], where x is manipulated variables such as mass flows of fuel and air, b_1, b_2, c_1, c_2 are parameters. These two types of fundamental elements are extracted from a priori knowledge about the NOx formation equations. Depending on the type of variable x, φ_1 is related to either the temperature in the furnace, or the available oxygen or nitrogen concentrations. φ_2 is related to the Arrhenius equation, where changes of x in φ_2 reflect the change of temperature distribution in the furnace. Once these fundamental elements are extracted, then the system $f(x)$ can be approximated using the model as described in Eqn. (1).

In fundamental grey-box modelling, two technical problems stand out in grey-box modelling:

1. How to identify the unknown parameters in the Fundamental Elements φ_i in (1)?
2. How to select the model structure? The number of terms that might be included into the system model can be very large and there is a combination problem.

One of the solutions to the above problem is to use genetic optimisation algorithms [12].

3.2 The data mining software

The data mining software illustrated in Fig. 1 comprises 5 major modules:

- Data mining modelling module
- Genetic algorithms module
- MATLAB source code generation module
- Module for Communication with MATLAB.
- Interface module

The screenshots of the data mining software are illustrated in Fig. 2 and Fig. 3.

The main characteristics of this data mining software are:

1) Flexibility. The user may select any data segment for modelling and mining using sliding window (see Fig. 4). The user can define system variables such as the name and the order of variables. The user can assign the ranges for all parameters to be optimised, and assign the values for key genetic operation parameters for data mining. Multiple types of data sources and different formats can be accepted for data mining, e.g. MATLAB files (*.m), ASCII coded text data files (*.txt) and binary data files (*.dat). The software is also interfaced with MATLAB, and can accept and interpret Matlab files and execute Matlab commands. The application also has the ability to generate MATLAB source code with any chromosome as the resultant model. The resulted code is in the form of a MATLAB function and can be saved to a file for further use.

2) Visualisation. Different “views” are designed for the data-mining process. The main “views” are:

- Project view – allows the user to access all relevant information and parameters for data mining.
- Plant data view – plots and displays plant data used in data mining and operators can access and select the data segment of his interest.
- Population view for data mining using genetic algorithm – displays, in graphical or text form, population/sub-populations and their properties during the evolution process.
- Chromosome view for data mining using genetic algorithm – displays a selected chromosome. Each chromosome corresponds to a potential practical model, the chromosome is showed in the decoded forms.
- Validation view – displays the prediction performance of a model decoded from any chosen chromosome for any data set.

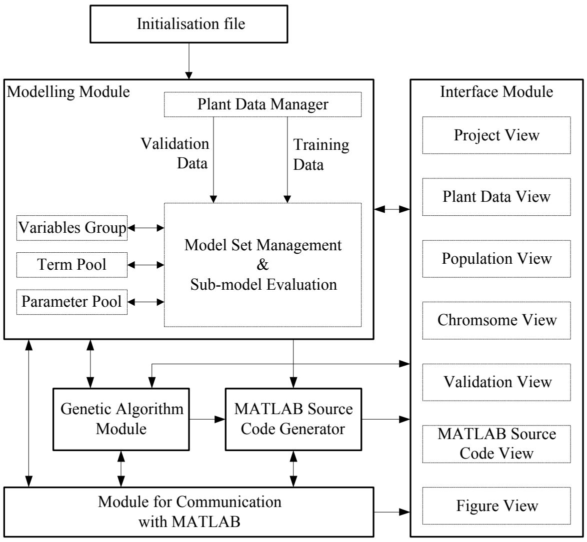

Fig.1 An overview of the application software architecture

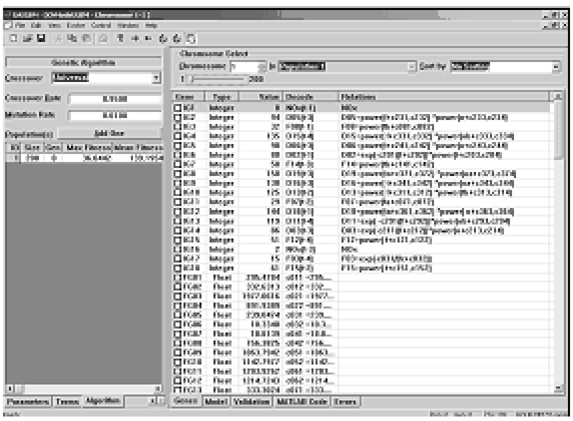

Fig 2. Screenshot of the data mining modelling software (The operators can define fundamental elements and associated the parameters in modelling)

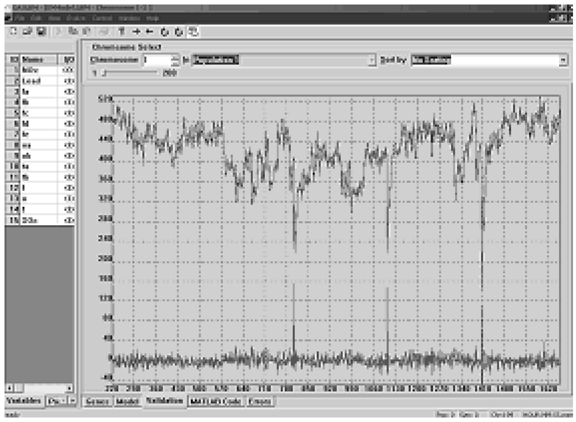

Fig 3. Modelling results – real data, model prediction and error, etc.

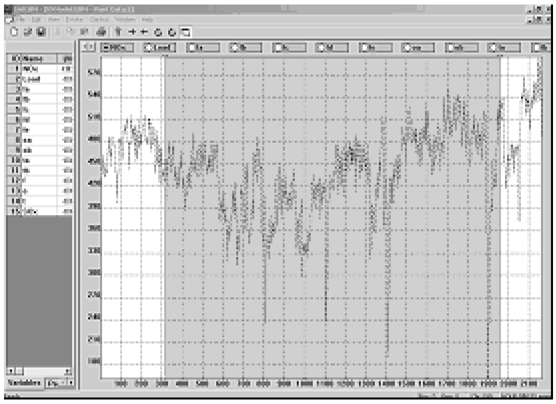

Fig 4. Selection of data segments by sliding window

3.3 Extracting patterns

In human supervisory control of coal-fired power generation plants, different software block is developed and used to perform each specific task, such as plant simulation and operation optimisation. Optimised operation solutions under various conditions (on both normal and abnormal situations) and best practices of operation experts will be also stored in a knowledge base. Inference mechanism will interpret operator’s demands, extract and aggregate solutions from the knowledge base and/or from the simulation and optimisation block, and then translate and pass the aggregated solutions to the human machine interface where the solution to human operators will be appropriately displayed basing on the cognitive point view and other man-machine system design principles.

Although operation conditions in TPPs may change from time to time because of the change of electricity demands and change of coal types, etc. However, plant will operate at limited set of typical working points/operation modes. Therefore, once plant models have been accurately developed using this data mining software, they can be used to obtain the optimal operation conditions for each typical working point/operation mode. As fundamental grey-box modelling will produce a model carrying rich operation information, operation patterns and values will be easily derived. For example, in modelling the NOx emission of a coal-fired power station in Northern Ireland, the following model is derived,

$$\begin{aligned} y(t)(1+a_1q^{-1}) &= b_0 + (b_1q^{-1} + b_2q^{-2} + b_3q^{-3} + b_4q^{-4})((m_{sa}(t) \\ &+ m_{pa}(t))/m_f(t) + b_5)^{c5} + b_5q^{-1}e^{(c4/(\theta(t)+b4))} + \varepsilon(t) \end{aligned} \tag{2}$$

where y(t) is the NOx, q^{-1} is the time lag, m_f is mass flow of coal, m_{pa} and m_{sa} are mass flow of primary and secondary air, θ is the burner tilting position. This model has been proven to give better prediction performance than other regression models, and each term in the model is a reflection of physical reality, therefore patterns can be easily derived from (2).

The optimal operation parameters together with expert experience will form an important part of the knowledge base. A knowledge base can be constructed by software, which consists of a rule-base and a database [13]. The rules in the rule base take the Mamdani form:

IF x_1 is A_1 AND x_2 is A_2 AND ... AND x_p is A_p
THEN
y_1 is B_1 AND y_2 is B_2 AND ... AND x_q is B_q

where x_i are advisory system input variables, y_i are output variables of the advisory system, A_i and B_i are fuzzy sets representing some real-world meanings. As a result, a rule can describe a condition-action statement that in principle can be clearly interpreted by human operators. A requirement for good interpretation is that the rule is transparent for the operator, and that the operator should be able to understand how the advisory system comes to the recommendation.

Obviously, Mamdani type structure will not only be able to deal with real-valued inputs and outputs, it will also provide a natural framework to include expert knowledge in the form of linguistic rules. Therefore this form of rule structure will be used in the proposed advisory system. In the database, the definitions of the linguistic labels are stored. These labels are defined according to some functions that define the patterns, and they are able to determine the membership of a current value of a variable to a label. These labels are applied in the antecedent and consequent proposition of the rules.

4 Conclusion

In this paper, data overload in supervisory control of critical complex industrial system is addressed. Data mining together a new modelling methodology have been shown to be an effective tool to help human supervisory control of a coal-fired power generation plant in reducing pollutant emissions. The data mining software has been developed and introduced.

References

[1] P. Adriaans, and D. Zantinge. *Data Mining*, Addison Wesley, Massachusetts, (1996).

[2] R. Agrawal, T. Dimielinski and A. Swami. Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering,* **5**, pp. 914–925 (1993).

[3] T. Bohlin. Interactive system identification: prospects and pitfalls, Springer: Berlin, (1991).

[4] A. Copado, *et al*. Boiler efficiency and NO_x optimisation through advanced monitoring and control of local combustion conditions, *Sixth International Conference on Technologies and Combustion for a Clean Environment*, **1**, pp: 903-908, Porto, 9-12 July, 2001.

[5] W.J. Frawley, G. Piatetsky-Shapiro and C.J. Matheus, Knowledge discovery in databases: an overview. *AI Magazine* **13**, pp. 57–70, (1992).

[6] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From data mining to knowledge discovery: an overview. *Advances in Knowledge Discovery and Data Mining*, MIT Press, Massachusetts, Chap. 1, (1996).

[7] J. Gleick, Chaos, making a new science. Penguin Books, (1998).

[8] K. Holmes, and L.W. Mayes. Progress report on the development of a generic NO_x control intelligent system (GNOCIS). *Coal R & D Program*, Project Profile 103, ETSU, (1994).

[9] K. Li, and P.A. Wieringa. Understanding perceived complexity in human supervisory control. *Cognition, Technology & Work*, Springer-Verlag London, **2**: 75-88, (2000).

[10] K. Li, S. Thompson, G. Duan, J. Peng (2002). A case study of fundamental grey-box modeling, *15th IFAC World Congress on Automatic Control*, Barcelona, July 2002.

[11] G.P. Liu, and S. Daley. Adaptive predictive control of combustor NO_x emissions, *Control Engineering Practice*, **9**, pp. 631-638, (2001).

[12] J. Peng, K. Li, S. Thompson. GA based software for power generation plant NO_x emission modelling, *6th International Conference on Technologies and Combustion for A Clean Environment*, **2**, 881-887, Porto, Portugal, 2001.

[13] O.G. Smink, P.A. Wieringa and J.C.J. van de Wiel. On genetic programming of fuzzy knowledge-bases, *Honeywell Industrial Automation & Control*, June, 2001.

[14] D.D. Woods, E.S. Patterson and E.M. Roth. Can we ever escape from data overload? A cognitive systems diagnosis. *Cognition, Technology & Work*, Springer-Verlag London, **4**, pp. 22-36, (2002).