Hybrid Intelligent Systems for Industrial Data Analysis

Arthur K. KORDON, *Member, IEEE*

Abstract—**A novel approach for industrial data analysis based on integration of three key computational intelligence approaches (genetic programming, analytical neural networks, and support vector machines) is proposed. The developed empirical models have good generalization capabilities, explicit input/output relationships, self-assessment capabilities, and low implementation and maintenance cost. The proposed approach has been implemented in several industrial applications in The Dow Chemical Company.**

Index Terms—**Hybrid intelligent systems, genetic programming, support vector machines, neural networks, inferential sensors**

I. Introduction

HYBRID intelligent systems are based on the effective synergy among several AI approaches like symbolic knowledge based systems, fuzzy systems, neural networks, and genetic algorithms [1]. Although the various approaches have their own strengths and specific application areas, very often they are insufficient to resolve real industrial problems. A typical case is one of the popular industrial applications of neural networks as soft sensors. Soft (or inferential) sensors assume that there is an empirical relationship between some easily measured and continuously available process variables and some critical parameters related to process quality like molecular distribution. Since the early 90s thousands of soft sensors have been applied in different areas of manufacturing [2]. However, along with the benefits that soft sensors have shown in different industrial conditions, several performance and long-term operation issues have appeared. Most of the problems are related to some limitations that are typical for soft sensors based on neural nets. Due to their sometimes ineffective, non-parsimonious structure and poor generalization capability outside the range of training data, their performance is very sensitive to specific process conditions. As a result of this reduced robustness there is a necessity of frequent re-training. The final effect of all of these problems is an increased maintenance cost and gradually decreased performance and credibility.

This example illustrates the negative impact of some limitations, specific to a selected approach, when applied to diverse industrial problems. It also shows the need for effective integration of various intelligent systems techniques in order to deal with the increased complexity of real world applications.

The first wave of hybrid intelligent systems, developed in the early 90s, is based on the key ingredients of soft computing (expert systems, fuzzy logic, neural networks, and genetic algorithms). The different mechanism for fusion, transformation, and integration of these techniques as well as the benefits of the hybrid intelligent systems are discussed in [1], [3], and the contemporary state of the art is given in [4].

Recently, several new intelligent systems approaches have shown remarkable theoretical growth and potential for solving complex industrial problems. Stacked analytical neural networks (internally developed in The Dow Chemical Company) allow very fast model development of parsimonious black-box models with confidence limits. Genetic programming (GP) can generate explicit functional solutions that are very convenient for direct on-line implementation in the existing process information and control systems [5]. Support vector machines (SVM) give tremendous opportunities for building empirical models with very good generalization capability [6].

These approaches are the basis of the second wave of hybrid intelligence systems. A novel methodology for integration of stacked analytical neural networks, GP, and SVM into a hybrid intelligent system is proposed in the paper. The integrated methodology amplifies the advantages of the individual techniques, significantly reduces the development time, and delivers robust empirical models with low maintenance cost. The advantages of the proposed methodology have been demonstrated in several successful applications in The Dow Chemical Company.

II. Requirements for Successful Industrial Data Analysis

If the goal of purely academic data analysis can be simplistically defined as "to transfer data into knowledge", the objective of industrial data analysis is "to transfer data into value". Since the economics is explicitly included in the objective function, the strategy for industrial data analysis is based on factors like minimizing modeling cost and maximizing data analysis efficiency under broad range of operating conditions. An obvious result of this strategy is the increased efforts in robust empirical model building, which is very often at the economic optimum. Another consequence of

A. K. Kordon is with The Dow Chemical Company, Freeport, TX 77566 USA (telephone: 979-238-5149, e-mail: akordon@ dow.com).

0-7803-7134-8/02/$10.00 © 2002 IEEE

the economically-driven industrial data analysis is the tendency to accelerate fundamental model building process by reducing the hypothesis search space with symbolic regression or high throughput design of experiments.

The key issue to implement productively the strategy is to develop a consistent methodology that effectively combines different modeling approaches to deliver high quality models with minimal efforts and maintenance. The main requirements toward a successful industrial data analysis can be defined as follows:

1) Robust, fast, and cost effective development process

The assumption is that the derived models from the data analysis have to be more effective than the alternative approaches (hardware sensor design or fundamental model building). Of special importance is the requirement to significantly reduce the development time while improving the consistency and performance of delivered empirical models. Another critical factor is to make the development process user-friendly with minimal tuning parameters and specialized knowledge.

2) Low sensitivity to process changes

Process changes driven by different operating regimes, equipment upgrades, or product demand fluctuations are more of a rule than an exception. It is unrealistic to expect that all the variety of process conditions will be captured by the training data and reflected in the developed empirical or fundamental models. The potential solution is in modeling approaches with better extrapolation capabilities at least 20 % outside the training range.

3) Performance self-assessment capability

Usually the models derived from the industrial data analysis infer the most critical parameters in industrial processes and as such require estimates with a very high level of reliability. It is necessary to include elements of self-assessment of prediction quality. A prospective approach is to use combined predictors [7] and their statistics as a confidence indicator of the model's performance.

4) Low cost of ownership and maintenance

The experience from "classical" neural net-based soft sensors shows that the lion share of maintenance cost is in frequent-re-training and especially in model re-design. The expectation is that by using non-black-box models with increased robustness the need for re-training will be significantly reduced. Another factor that contributes to cost of ownership reduction is the ease of on-line implementation. Of special interest are the explicit functional models, generated by GP. They are well understood by process engineers, directly applicable in the control system, and do not require specialized knowledge for maintenance.

III. SELECTED APPROACHES FOR DEVELOPMENT OF HYBRID INTELLIGENT SYSTEMS

It is very difficult to satisfy the defined requirements by a specific soft computing technique only. However, several intelligent systems approaches can effectively resolve some specific issues and become the building blocks of an integrated methodology for hybrid intelligent systems development. Of special interest are the following three approaches – analytical neural nets, support vector machines (SVM), and genetic programming (GP).

A. Analytical Neural Networks

Analytical neural networks are based on a collection of individual, feedforward, single layer neural networks where the weights of the input to hidden layer have been initialized according to a fixed distribution such that all hidden nodes are active. The weights of the hidden to output layer can then be calculated directly using least squares. Advantages of this method are: it is fast and each neural network has a well defined, single, global optimum. Each of these networks have a known Vapnik-Chernovenkis (VC) dimension, so collections with a given complexity can be developed and optimum use can be made of statistical learning theory. Time delays between inputs are handled through convolution functions. In addition, the use of a collection of networks gives more robust models that include confidence limits based on the standard deviation of stacked neural nets.

Analytical neural networks contribute to the hybrid intelligent systems development process by allowing an extensive nonlinear sensitivity analysis and input feature selection. They allow for a fast feasibility test of the model development process and they deliver models that have confidence limits associated with predicted outputs.

B. Support Vector Machines

Support vector machines have become an active field of research in recent years. This type of learning machine implements the Structural Risk Minimization principle, which has its foundation in statistical learning theory and is particularly useful for learning with small sample sizes [6]. One of the key features is the use of kernel functions. This enables the method, not only to use non-linear mappings of the input data, but also overcomes the curse of dimensionality. Furthermore, through the introduction of a special loss function, the ε-insensitive loss, the model is defined in terms of a subset of the learning data, called the support vectors. Varying the size of ε influences the number of support vectors and therefore allows direct control over the complexity of the model.

The SVM method is a very robust method and has a unique contribution to the hybrid intelligent systems development by means of automatic outlier and novelty detection. The fact that the SVM model is a sparse representation of the learning data allows the extraction of a condensed data set based on the support vectors. Finally, by using certain types of kernels, the extrapolation capabilities of the model can be increased dramatically, especially by incorporating prior information [8]. All these features combined pave the way to the development of robust empirical models.

C. Genetic Programming

The third approach of interest to hybrid intelligent systems development is GP with its capability for symbolic regression

[5]. GP-generated symbolic regression is a result of simulation of the natural evolution of numerous potential mathematical expressions. The final results is a list of "the best and the brightest" analytical forms according to the selecting objective function. Of special importance to industry are the following unique features of GP[9]:

- no *a priori* modeling assumptions
- derivative-free optimization
- few design parameters
- natural selection of the most important process inputs
- parsimonious analytical functions as a final result.

The last feature has double benefit. On one hand, a simple empirical model often has better generalization capability, increased robustness, and needs less frequent re-training. On the other hand, process engineers and developers prefer to use non-black box empirical models and are much more open to take the risk to implement models based on functional relationships. An additional advantage is the low implementation cost of such type of models. It can be applied directly into the existing Distributed Control Systems (DCS) avoiding additional specialized software packages, typical for neural net-based solutions.

At the same time there are still significant challenges in implementing empirical models generated by GP: function generation with noisy industrial data, dealing with time delays, sensitivity analysis of large data sets, to name a few. Of special importance is the main drawback of GP – the slow speed of model development due to the inherent high computational requirements of this method. For real industrial applications the calculation time is in order of days, even with the current high-end PCs.

IV. Integrated Methodology for Hybrid Intelligent Systems Development

The objectives of the proposed integrated methodology are to satisfy the defined criteria for successful industrial data analysis, i.e., to reduce development time, to deliver a model with the best generalization capability, and to minimize the implementation and maintenance cost. The main blocks of the methodology and the related process of data reduction are shown in Fig 1.

The main purpose of the first main block is to reduce the number of inputs to those with the highest sensitivity toward the output. Another objective is to test via simulation the hypothesis whether some form of nonlinear relationship between the selected inputs and the output exists. This is a critical point in the whole methodology, because if a neural net model cannot be built, the empirical model development process stops here. The conclusion in this case could be that if

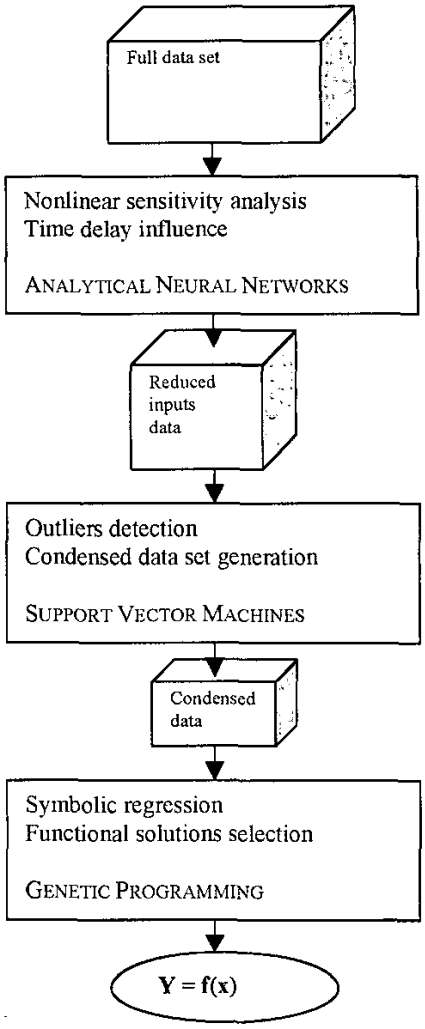

Fig. 1 Main blocks of an integrated methodology for hybrid intelligent systems development.

a universal approximator, like a neural net, cannot capture a nonlinear relationship, there would be no basis for variable dependence and no need to look for other methods. The sensitivity analysis is based on stacked analytical neural nets. A big advantage of this type of neural nets is the reduced development time. Within a couple of hours, the most sensitive inputs are selected, the performance of the best neural net models is explored, and the data for the computationally intensive symbolic regression (GP-function generation) step is prepared. Typically, thirty stacked neural nets are used to improve generalization and estimate neural net model agreement error. This step begins with the most complex structure of all possible inputs. During the sensitivity analysis, decreasing the number of inputs, gradually reduces the initial complex structure. The sensitivity of each structure is the average of the calculated derivatives on every one of the stacked neural nets. The procedure performs automatic

elimination of the least significant inputs and generates a matrix of input sensitivity vs. input elimination.

Another important task performed by the analytical neural networks is to deal with time delays. The classical approach to handle time series by neural nets is to add additional inputs for the previous time steps [10]. Unfortunately, this technique increases the dimensionality of the neural net significantly. This increase in the dimensionality of the input vectors has a large impact on the number of required data points for proper model identification. The problem is even bigger in the case of GP modeling. Therefore, it would be desirable to include information from previous time-steps without increasing the dimensionality of the input to the network. This can be achieved by performing a convolution on the input using an appropriately shaped function. As a result of the first block of the integrated methodology, the size of the full data set is reduced to the number of the most sensitive inputs.

The purpose of the next block, based on SVM, is to further reduce the size of the data set to only those data points that represent the substantial information about the nonlinear model. Outliers' detection is the first task in this process. For outliers detection, we make use of the fact that the data points containing important information are identified by the SVM method as support vectors. When the weight of a data point is non-zero, it is a support vector. The value of a support vector's weight factor indicates to what extent the corresponding constraint is violated. Non-zero weight factors hitting the upper and lower boundary indicate that their constraints are very difficult to satisfy the optimal solution. Such data points are often so unusual with respect to the rest of the samples, that they might be considered as outliers. An outlier detection tool, using the SVM method, typically constructs several models of varying complexity. Data points with a high frequency of weight values on the boundaries are assumed to be outliers.

One of the main advantages of using SVM as a modeling method is that the user has direct control over the complexity of the model (i.e., the number of support vectors). The complexity can be controlled implicitly or explicitly. The implicit method controls the number of support vectors by controlling the acceptable noise level. To explicitly control the number of support vectors, one can either control the ratio of support vectors or the percentage of non-support vectors. In both cases, a condensed data set that reflects the appropriate level of complexity is extracted for effective symbolic regression.

An additional option in this main block is to deliver an empirical model based on SVM. Some recent results show [8], that SVM models based on mixed global and local kernels have very good extrapolation features. If an empirical model, generated by GP does not have acceptable performance outside the range of training data, the SVM-based inferential model is a viable on-line solution.

The final block of the integrated methodology for hybrid intelligent systems development uses the GP approach to search for potential analytical relationships in a condensed data set of the most sensitive inputs. The previous steps significantly reduce the search space and the effectiveness of GP is considerably improved. The final result from the symbolic regression is a list of several analytical functions and subequations that satisfy the best solution according to a defined objective function. The analytical function selection for the final empirical model is still more of an art than a well-defined procedure. Very often the most parsimonious solution is not acceptable due to specific manufacturing requirements. It is preferable to deliver several potential functions with different levels of complexity and let the final user make the decision. The generalization capabilities of each empirical model are verified for all possible data sets. Of special importance is the performance outside the training range. It is also possible to design a model agreement-type confidence indicator based on stacked symbolic predictors.

V. APPLICATION AREAS

The potential for value creation from effective industrial data analysis based on hybrid intelligent systems is enormous. Some of the key application areas, explored recently in The Dow Chemical Company are as follows:

A. Robust Soft Sensors

Some of the critical parameters in chemical processes are not measured on-line (composition, molecular distribution, density, viscosity, etc.) and their values are captured either by lab samples or off-line analysis. However, for process monitoring and quality supervision the response time of these measurements with low frequency (several hours, even days) is very slow. When the critical parameters are not available on-line in situations with alarm showers due to complex root causes the negative impact could be significant and eventually could lead to shutdown. One of the approaches to address this issue is through development and installation of expensive hardware on-line analyzers. Another solution is by using robust soft sensors developed by the proposed methodology [9]. An example of a robust soft sensor for emission estimation is given in the next section.

B. Automated Operating Discipline

Operating discipline is a key factor for competitive manufacturing. Its main goal is to provide a consistent process for handling all possible situations in the plant. It is the biggest knowledge repository for plant operation. However, this documentation is static and is detached from the real-time data of the process. The missing link between the dynamic nature of process operation, and the static nature of operating discipline documents is traditionally carried out by the operating personnel. However, this makes the existing operating discipline process very sensitive to human errors, competence, inattention, or lack of time.

One approach to solving the problems associated with operating discipline and making it adaptive to the changing operating environment is to use real-time hybrid intelligent systems. Such type of a system was successfully implemented in a large-scale chemical plant at The Dow Chemical Company [11]. It is based on integrating experts' knowledge with soft sensors and fuzzy logic. The hybrid system runs in parallel with the process; it detects and recognizes problem situations automatically in real-time; it provides a user-friendly interface so that operators can readily handle complex

alarm situations; it suggests the proper corrective actions via a hyper-book; and it facilitates effective shift-to-shift communication.

C. Accelerated Fundamental Model Building

The large potential of genetic programming (GP)-based symbolic regression for accelerated fundamental model building is demonstrated in a case study for structure-property relationships [12]. The generated symbolic solution is similar to the fundamental model and is delivered in significantly less time. Additional benefits include identifying key variables and transforms, enabling rapid testing of a new physical hypothesis, and the reduction of the number of experiments for model validation. By optimizing the capabilities for obtaining fast and reliable GP-generated functional solutions in combination with the fundamental modeling process, a real breakthrough in the speed of new product development can be achieved.

D. Effective Design of Experiments (DOE)

The integration of GP with DOE has the potential to improve the effectiveness of empirical model building by saving time and resources in situations where experimental runs are quite expensive or technically unfeasible because of extreme experimental conditions. GP was successfully applied to the development of variable transforms that linearize the response in statistically designed experiments for a chemical process in The Dow Chemical Company [13].

E. Empirical Emulators

Empirical emulators mimic the performance of first principle models by using various data-driven modeling techniques. A key feature of empirical emulators is that the training data for empirical model building is generated by design of experiments from first principle models called simulators. This allows a high degree of freedom for development of reliable data-driven models. The most obvious scheme for implementation of empirical emulators is as accelerator of computational time for fundamental models (the gain is $10^3 - 10^5$ times faster). Another possible scheme is to use the empirical emulator as an estimator of fundamental model's performance. Of special importance to on-line optimization is the scheme of the empirical emulator as an integrator of different types of fundamental models (steady-state, dynamic, fluid, kinetic, thermal, etc). The results from a case study of an emulator implementation are given in [14].

VI. INDUSTRIAL APPLICATION

Soft sensors for emission estimation are one of the most popular application areas and a viable alternative to hardware analyzers. Usually an intensive data collection campaign is required for empirical model development. However, during on-line operation the output measurement is not available and some form of soft sensor performance self-assessment is highly desirable. Since it is unrealistic to expect that all possible process variations will be captured during the data collection campaign, a soft sensor with increased robustness is required. Such type of soft sensors, based on the proposed integrated methodology, was developed and implemented in one of The Dow Chemical Company plants in Freeport, TX. The key results from implementation of the main blocks are as follows:

A representative data set from eight potential process input variables and the measured emission as output included 251 data points for training and 115 data points for testing. The test data is 140% outside the range of the training data, which by itself is a severe challenge for the extrapolation capability of the model. As a result of the nonlinear sensitivity analysis based on the analytical neural networks, the data set was reduced to five relevant inputs. The performance of such type of potential model with five inputs, 10 neurons in the hidden layer, and a model disagreement indicator based on the standard deviation of 30 stacked predictors is shown in Fig. 2. The possibility for nonlinear model building and the potential of the model agreement indicator for performance self-assessment are clearly demonstrated.

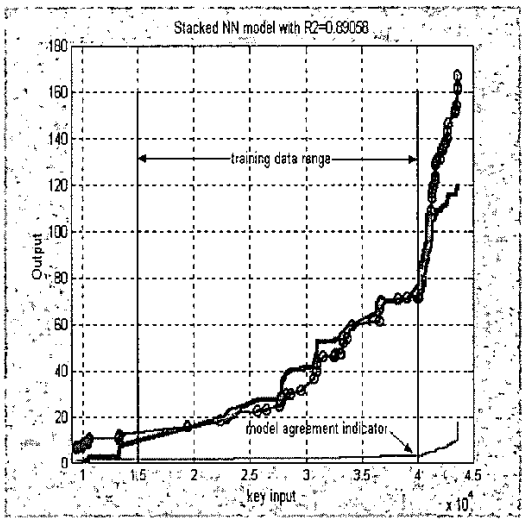

Fig. 2. Performance of a stacked analytical neural net model with model agreement indicator.

The extraordinary extrapolation capability of a potential empirical model based on SVMs is shown in Fig.3. The model is based on a mixture of a second order polynomial global kernel and an RBF local kernel with width of 0.5 in a ratio of 0.95. An additional benefit from this phase of the integrated methodology is that the model is based on 34 support vectors only.

As a result, the representative data set for deriving the final symbolic regression model is drastically reduced to only 8.44% of the original training data set. As it is shown in Fig. 4, the performance of the GP-generated model, based on the condensed data set, is comparable with the other two approaches.

The initial functional set for the GP includes: {addition, subtraction, multiplication, division, square, change sign, square root, natural logarithm, exponential, and power}. Function generation takes 20 runs with population size of 500, number of generations of 100, number of reproductions per

generation of 4, probability for function as next node of 0.6, parsimony pressure of 0.05 and correlation coefficient as optimization criterion. Eight symbolic predictors with different number of inputs and nonlinear functions were selected in a stacked model. The average value is used as the soft sensor prediction and the standard deviation is used as a model disagreement indicator. The soft sensor for emission estimation is in operation in Freeport, TX since August 2001.

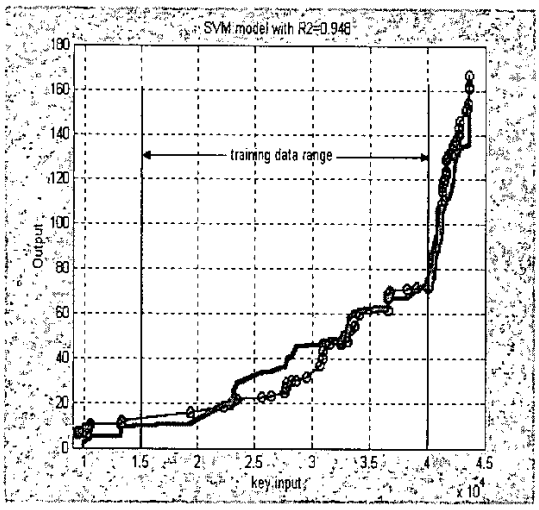

Fig. 3. Performance of an SVM model using a mixture of polynomial and RBF kernels.

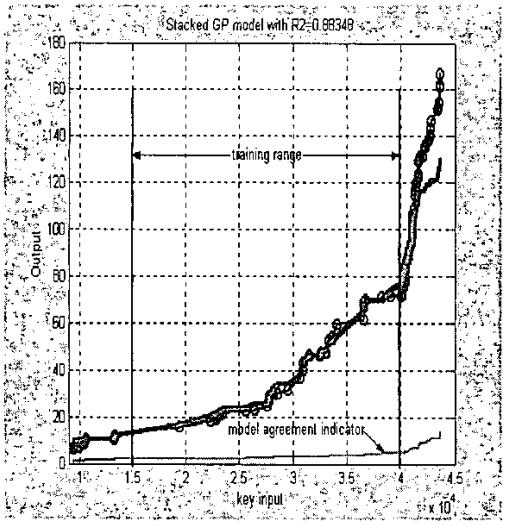

Fig. 4. Performance of a Stacked Symbolic Regression model with model agreement indicator.

VII. Conclusion

A novel integrated methodology for industrial data analysis has been defined and successfully implemented in various applications in The Dow Chemical Company. The proposed hybrid methodology is based on using different computational intelligence components (stacked analytical neural nets, genetic programming, and support vector machines). The driving force behind the need for integration is the requirement of industry for empirical models with increased robustness and reduced development time. The illustrated application shows one of the main advantages of the proposed methodology – significant reduction of the training data set by nonlinear sensitivity analysis and support vector machines. The final on-line solution, generated by GP, is based on a very compact and robust stacked empirical model with self-assessment capability that requires minimal re-training and maintenance cost. The success of this application in a complex industrial environment, as well as similar implementations in the area of automating operating discipline, accelerating fundamental model building, empirical emulators, and effective DOE, demonstrate the great potential of the integrated approach for solving difficult industrial problems.

VIII. References

[1] L. Medsker, *Hybrid Intelligent Systems* Boston, MA: Kluwer, 1995.

[2] R. Neelakantan and J. Guiver, "Applying Neural Networks", *Hydrocarbon Processing*, vol. 9, October 1998, pp. 114-119.

[3] R. Khosla and T. Dillon, *Engineering Intelligent Hybrid Multi-Agent Systems* Boston, MA: Kluwer, 1995.

[4] S. Mitra, S. Pal, and P. Mitra, "Data Mining in Soft Computing Framework: A Survey", *IEEE Trans. Neural Networks*, vol. 13, pp. 3-14, 2002.

[5] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Cambridge, MA: MIT Press, 1992.

[6] V. Vapnik, *Statistical Learning Theory*, New York, NY: Wiley, 1998.

[7] S. Sharkley (Editor), *Combining Artificial Neural Networks*, London, UK: Springer, 1999.

[8] G. Smits and E. Jordaan, "Using Mixtures of Polynomial and RBF Kernels for Support Vector Regression", *In Proceedings of WCCI'2002*, Honolulu, HA: IEEE Press, 2002, pp. 2785 – 2790.

[9] A. Kordon and G. Smits, "Soft Sensor Development Using Genetic Programming", *In Proceedings of GECCO'2001*, San Francisco, CA: Morgan Kaufmann, 2001, pp. 1346 – 1351.

[10] Medsker L. and L. Jain (Editors), *Recurrent Neural Networks: Design and Applications*, Boca Raton, FL:CRC Press, 2000.

[11] A. Kordon, A.Kalos , and G.Smits, "Real Time Hybrid Intelligent Systems for Automating Operating Discipline in Manufacturing", *Artificial Intelligence in Manufacturing Workshop Proceedings of the 17th International Joint Conference on Artificial Intelligence IJCAI-2001*, Seattle, WA, August 2001.

[12] A. Kordon , H. Pham, C. Bosnyak, M. Kotanchek, and G. Smits, "Symbolic Regression in Fundamental Model Building", accepted for GECCO'2002.

[13] F. Castillo, K. Marshall, J. Greens, and A. Kordon, "Symbolic Regression in Design of Experiments: A Case Study with Linearizing Transformations", accepted for GECCO'2002.

[14] P. K. Mercure,, G.F. Smits, and A. K. Kordon, "Empirical Emulators for First Principle Models", *Fall 2001 AIChE meeting*, Reno, NV, 2001.