International Journal of Fuzzy Logic and Intelligent Systems, vol. 5, no. 3, September 2005 pp. 200-205

Datamining Roadmap to Extract Inference Rules and Design Data Models from Process Data of Industrial Applications

Hyeon Bae, Youn-Tae Kim, Sungshin Kim, and George J. Vachtsevanos*

School of Electrical and Computer Engineering, Pusan National University, Busan, Korea

***School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA**

Abstract

The objectives of this study were to introduce the easiest and most proper applications of datamining in industrial processes. Applying datamining in manufacturing is very different from applying it in marketing. Misapplication of datamining in manufacturing system results in significant problems. Therefore, it is very important to determine the best procedure and technique in advance. In previous studies, related literature has been introduced, but there has not been much description of datamining applications. Research has not often referred to descriptions of particular examples dealing with application problems in manufacturing. In this study, a datamining roadmap was proposed to support datamining applications for industrial processes. The roadmap was classified into three stages, and each stage was categorized into reasonable classes according to the datamining purposed. Each category includes representative techniques for datamining that have been broadly applied over decades. Those techniques differ according to developers and application purposes; however, in this paper, exemplary methods are described. Based on the datamining roadmap, nonexperts can determine procedures and techniques for datamining in their applications.

Key Words : Knowledge & Rule Extraction, Datamining Roadmap, Industrial Applications

1. Introduction

A number of factors are driving corporations toward short product life cycles, lower total production volumes, and higher levels of product mixes due to customization. Therefore, there is a crucial need to develop tools and methodologies that assist in the reduction of time and effort spent on planning, specifying, designing, validating, and deploying facilities for quickly changing product models. The need for businesses to have timely, quality, and dynamically changing information available to executives and mid-level management is not only critical to winning a competition but necessary for survival.

In the field of Datamining (DM), knowledge is mined, extracted, from data. In modern manufacturing environments, vast amounts of data are collected in database management systems and data warehouses from all relevant areas such as product and process design, assembly, materials planning and control, scheduling, and maintenance. This study aims to determine the overall technical architecture and framework by which the Knowledge Management (KM) techniques can be implemented. The goal is to apply, in the manufacturing arena, DM that will provide an understanding of the existing databases and their integration into decision-making solutions.

A commonly used definition of datamining is a non-trivial process of extracting potentially useful, previously unknown, and ultimately understandable information from data. Datamining, also referred to as knowledge discovery in databases, uses sophisticated statistical and/or intelligent analysis, as well as modeling techniques, to uncover patterns and relationships hidden in organizational databases.

Datamining is the core of the knowledge discovery process. However, datamining explains only 15% to 25% of the entire knowledge-discovery process [1]. It is required to identify specific manufacturing problems and determine suitable datamining tools for the data preprocessing and datamining stages. According to Pawlak, a great master of datamining, there is no single technique to solve real-world problems, because industrial data usually consist of hundreds of dimensions and types. That is, datamining is not a panacea for all manufacturing problems. There are various datamining techniques available and each tool has its advantages and disadvantages depending on the type of task. Therefore, after reviewing the features of various datamining methods, a method is selected and applied to target applications. To support this process, a datamining roadmap, based on survey information from much of the literature, was developed. Berry and Linoff classified mining tasks into six areas, including classification, estimation, prediction, affinity grouping, clustering, and description [2]. Their useful scheme in matching various datamining tools with various tasks was examined and a new scheme based on the datamining roadmap is here proposed.

Manuscript received July 27, 2005; revised Aug. 30, 2005. This work was supported by "Research Center for Logistics Information Technology (LIT)" hosted by the Ministry of Education & Human Resources Development in Korea.

2. Data Mining and Knowledge Extraction

2.1 Datamining Procedure

The evolution of datamining (DM) has been going on since business people began storing data on computers. Through

continued improvements in data access, users are now better able to navigate their data storage systems. Datamining provides information crucial in helping businesses and industries that want to improve their products, marketing, sales, and customer service. Datamining allows users to analyze large databases to solve business decision problems. That is, the ever-increasing quantity of data in every computing environment presents both an opportunity to extract useful information and a challenge to process the massive volume of data effectively.

Datamining is an extension of statistics, but it has more of a technology focus; it is another way of recognizing, investigating, and tracking patterns of data information for specific purposes. Analyzing and generating models from data used to be in decades the pattern-recognition and machine-learning communities have greatly expanded their areas of application and the kind of information to be extracted, as well as the variety of models. In most cases, standard database results are presented to users as something they already knew existed in the database, whereas datamining extracts information from the database that users did not know was there. Typical problems are usually how to address or classify data, cluster data, or find relationships and analysis among data.

The process of datamining is not limited to inductive learning, and the data must be preprocessed before a learning algorithm is applied. This usually includes such steps as data cleaning and data reduction. Data cleaning may include accounting for missing values, and determining how to deal with noisy and inconsistent data. Data reduction may be done along both the attribute and instance dimensions, that is, reducing either the number of variables (attributes) or data points (instances). Once the data have been prepared, inductive learning most commonly involves learning one of the three concepts: classification, clustering, or association rule discovery. Classification involves learning a model that can discriminate between the values (classes) of a particular target attribute called the class attribute. This is a supervised learning task, since the training data from which the algorithm learns is labeled, that is, the class values for the target attributes are known. If there is no such class attribute, two types of unsupervised learning are most commonly used. Data-clustering algorithms aim to discover natural groupings, or clusters, of instances, whereas association rule discovery aims to discover relationships between the attributes. Once the patterns have been obtained, they must be validated and then the knowledge learned can be implemented (Fig. 1)

2.2 Problems in Applying Datamining

Many techniques and tools have been applied in data mining. Each method has a different performance and implementation according to applications.

In past studies related to datamining, datamining procedures for knowledge discovery have been roughly described, but detailed procedures of stages and methods have not dealt with classifying the categories of data mining. Especially, most datamining applications have been applied in marketing areas, so it is difficult to find adequate application examples for manufacturing processes. In addition, how to achieve goal and what to do to achieve it have been determined by experienced knowledge; it is very difficult to support a decision for improving quality and yield without sufficient experience of the target case.

In manufacturing processes, costs must be reduced with the least trial and error and safe operation of processes must be guaranteed with adequate recipes, thus, the datamining roadmap is necessary to support fast and easy selection categories and datamining methods.

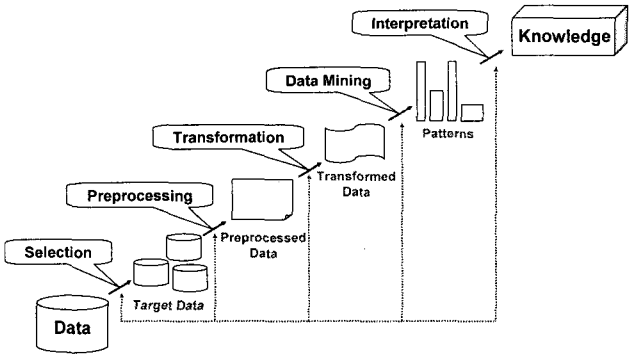

Fig. 1. Processing of datamining for knowledge discovery.

2.3 Functions of Datamining Roadmap

Determination of procedures and selection of techniques play important roles in datamining, because many techniques including statistical methods have been examined and applied in datamining applications. However, that determination has been achieved by experts' knowledge and intuition. Determination based on intuition can cause significant problems in industrial applications.

Datamining has been broadly employed in marketing fields; developers use trial and error to find proper procedures and methods of datamining. This does not lead to significant problems in marketing applications, but it is not an adequate approach in industrial applications. Mistakes will directly affect yield and quality in industrial processes, so using the wrong datamining technique is worse than not using datamining. This is a reason that process operators do not want to implement new techniques, such as datamining, in industrial fields. To save their trouble, a guideline for applications is required for industrial processes, so the datamining roadmap is proposed in this paper. The major function of the datamining roadmap is to give information to developers of the datamining applications. The roadmap will reduce trial and error time, cut down expenses, and support operators in many-sided considering.

Berry and Linoff classified mining tasks into six areas, including classification, estimation, prediction, affinity grouping, clustering, and description [2]. The goal of the proposed datamining roadmap is to support development of a knowledge-based system in manufacturing processes that require fast development and less trial. And, also because the proposed roadmap was constructed after a broad survey of the literature, it can be an effective guideline for datamining and knowledge discovery in manufacturing processes.

3. Development of Datamining Roadmap

In this study, the datamining roadmap was constructed to reduce application troubles and to support easy development. Categories summarized from several studies were considered and classified according to purposes and goals [3-8]. After surveying past research, the roadmap was designed based on a diagram drawn for easy understanding of the roadmap. In past studies, there were some classified categories but the categories, were organized for marketing fields. Therefore, there was a limitation in applying these classified categories in industrial fields. That is, it was not easy to search adequate guidelines for datamining in industrial processes.

As shown in Fig. 2, the proposed roadmap consists of a preprocessing stage, a datamining and knowledge stage, and a system construction stage. In each stage, proper roles are performed, so stable application can be achieved in industrial fields. Because field data are sometimes dirty, preprocessing is required. This task is accomplished in the preprocessing stage. Some techniques have been considered for preprocessing, such as transform, missing data treatment, and others. However, those methods were not included as an important stage of datamining, but as a sub-treatment for performance improvement. In the proposed roadmap, they are classified as one stage of the datamining roadmap. The data mining and knowledge discovery stage consists of five categories that have been dealt with over decades. In this study, the categories were defined for industrial applications.

Finally, the result of datamining and knowledge discovery is classified into two systems: one is data models and the other is inference rules. These two systems are considered as the final goal of datamining and knowledge discovery. It is possible to disagree with the proposed categories, but with detailed consideration, most datamining results can be included in the two systems. The roadmap proposed in this paper is expected to provide information for datamining in industrial processes.

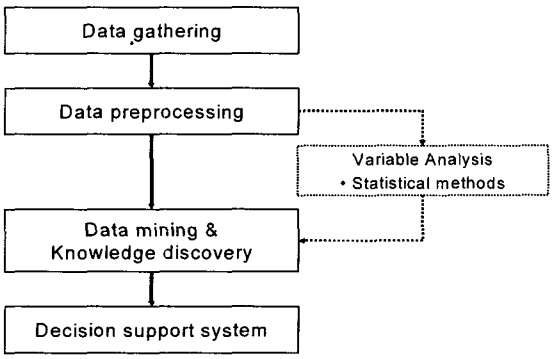

Fig. 2. Main stages of datamining roadmap.

3.1 Stages of Data Preprocessing

As shown in Fig. 3, the preprocessing stage consists of data cleaning, data transform, and data generation. Collected data from industrial fields is not always adequate for the required performance in datamining, and this is a frequently occurring problem. Therefore, proper preprocessing is required for adequate performance. In this study, the three typical categories were classified as data preprocessing, but some more categories can be included.

3.1.1 Data cleaning

Data collected from industrial fields is of high impurity because of the status of sensors or measuring environments. These data can deteriorate the performance of constructed models or rules. Therefore, the impurity of data needs to be reduced by cleaning filters. In this category, data are cleaned and unnecessary data are removed. Several methods of data cleaning have been developed and applied, such as the moving average method, several filters, and others.

3.1.2 Data transform

Industrial data are usually collected in time-series format, but, in accordance with the application, it is difficult to analyze data in the time domain. In such a case, data are transformed onto the frequency domain. Especially, time-series datamining requires data transformation to extract features from time-series data. Transformation expands the scopes of data analysis and extracts several features. Fourier and wavelet transform have been broadly applied for data transformation. Among methods, developers select the suitable one corresponding to the data characteristics or goals.

3.1.3 Data generation

Other problems with data are data missing and data nonunifomity. Collected data from industrial processes can be missed at some columns of a database, which caused by the status of sensors and equipment. In this case, important and principal information can be lost in feature extraction. In general, process data are stored by sampling inspection and under normal conditions; thus, data preprocessing is required. Because users occasionally do not occasionally recognize the problem resulting from this cause, successive faults can occur. Therefore, it is necessary to remove unnecessary data sets and generate filling data.

Many methods have been developed for data generation and there is no one, all-purpose method. Statistical methods and mathematical models have been used. In a case study by the author, data generation was achieved to fill-in insufficient data for a management of wafer process. Collected data from industrial processes also have nonuniformity, because the data are gathered under normal conditions or in specific cases. Therefore, it is necessary to compensate for the weak points. Accordingly, data were generated by the bootstrap method, one of the statistical resampling methods used to check the status of data distribution. This data generation method using statistical and intelligent methods is useful in variable processes.

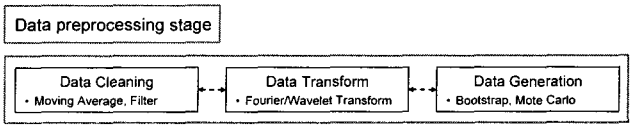

Fig. 3. Data preprocessing stage.

3.2 Datamining and Knowledge Discovery Stage

As shown in Fig. 4, the data mining and knowledge discovery stage for extracting knowledge from preprocessed data is one of the most important procedures, and various techniques have been employed.

In past studies, the datamining and knowledge discovery stage has been researched in marketing cases for off-line data. Therefore, in industrial applications, this stage has no flexibility and adaptability, and it is difficult to select the proper category for the best procedure. Specially, if the developer does not have much experience, much effort to determine categories and techniques is required. To solve this problem, five categories were classified and proper techniques were included in each category according to datamining purpose in this study. The purpose is that users be able to determine categories and techniques.

Many types of techniques were employed to extract knowledge in this stage. In a case study by the author, first, for management of the wafer fabrication process, decision trees generated inference rules in rule generation and neural networks trained data models in prediction. Second, for fault detection and diagnosis of induction motors, wavelet analysis and fuzzy similarity measure were used for feature extraction and neural network models classified fault types. Finally, for coagulant control of water treatment plants, fuzzy c-means clustering classified data for rule generation to determine a proper coagulant type.

3.2.1 Feature extraction

Feature extraction is one of the most important procedures to extract information or patterns from cleaned or transformed data in datamining. In feature extraction, many statistical and intelligent techniques such as neural networks, PCA, genetic algorithms, decision trees, fuzzy set, and others have been studied over the past decades. This category plays an important role in datamining.

3.2.2 Classification

This category classifies observed data into proper classes using rules or models. In general, rules show the relationship between input and output variables, but the rules in this category are used for classification. Models were also employed for classification. Statistical models and rules such as decision trees, neural networks, Bayesian networks, supporting vector machine (SVM), and others have been used for classification.

3.2.3 Prediction

Prediction is one procedure for mapping inputs to outputs. This has been widely applied to obtain information from data, but it is not usually classified as one of the data mining stages. In this study, prediction was included in the datamining category, because it was considered that prediction is a procedure to obtain necessary information from data. The effects of current inputs in the future can be analyzed, and the status can be estimated with past data in the prediction category. For prediction, many methods such as regression, Bayesian networks, neural networks, dynamic polynomial neural networks (DPNN), and others have been used over decades.

3.2.4 Rule extraction

Rule extraction is one of the important categories in company with feature extraction. From rule extraction, logics can be constructed that define the relationships between input and output variables, and logics have been broadly employed for modeling and control. Rules (logics) can be extracted from raw data, transformed data, or extracted features. Several methods such as nearest neighbor, Kohonen map, decision tree, fuzzy set, and others have been evaluated in rule extraction.

3.2.5 Clustering

Clustering is one of the popular traditional methods for data analysis. Clustering is classified as a statistical method, but today the boundaries between statistics and datamining are collapsing. Therefore, in this study, clustering was categorized as datamining. Especially, since the advent of fuzzy clustering, clustering has been widely used in system design based on fuzzy rules. In this study, clustering was deemed to be an important category. There are two types of clustering methods: hierarchical and nonhierarchical methods. Data can be properly grouped by a combination of both methods. Clustering has been developed based on fuzzy c-means clustering, k-means clustering, hierarchical clustering, statistical methods, and others.

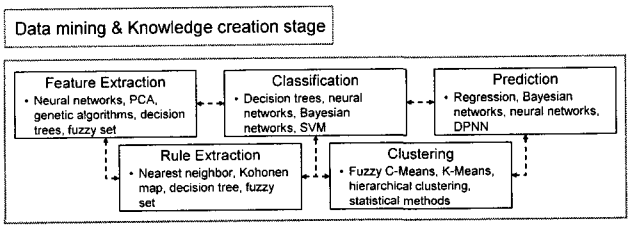

Fig. 4. Datamining and knowledge discovery stage.

3.3 Stage for System Construction

As shown in Fig. 5, the system construction stage is classified into two systems. One is an inference system based on rules, and the other is a prediction system based on models. Research related to unit techniques is part of the traditional literature of datamining, but studies related to the final system have still not been dealt with.

In the datamining roadmap proposed in this paper, the results of datamining and knowledge discovery can be explained by two systems that consist of inference rules and data models.

3.3.1 Data model

The data model is a major result of datamining and knowledge discovery. The model can be usefully used for mapping inputs to outputs and constructing equations for classification. In datamining and knowledge discovery, data models have

functions such as prediction, classification, and others. In the past, if a model was used for classification, it was regarded as a datamining category. However, if a model was applied for prediction, it was not categorized as datamining. In this study, prediction was included in the datamining categories. Datamining and knowledge discovery is a process of obtaining information from data. Prediction is also accomplished to obtain information; thus, it is classified as one of the datamining techniques. Considering the datamining concept, it is reasonable to categorize the prediction model as datamining. Because the data model can be constructed easily compared with constructing inference rules, the data model has been broadly applied in system analysis. However, the performance of the data model cannot be guaranteed if the proper preprocessing is not performed, and it is difficult to evaluate the performance.

3.3.2 Inference rule

The inference rule is one result of datamining and knowledge discovery that can logically express casual relationships of variables. Inference rules execute functions for classification, diagnosis, and prediction. Especially, the performance of inference rules can be improved by combining them with optimization methods such as genetic algorithms, evolution strategy, and others, so the rule-based system has been developed over a long time in system analysis. In past studies, the result of the category was not classified, but in this study, this category is classified as a target system of datamining. Because inference rules can explain results well compared with data models, they have been widely used in system analysis. However, if the principal inputs are not selected, the correct rules cannot be generated. This indicates that variables affect the performance of rule-based system. Therefore, it is very important to select important variables that can guarantee sufficient performance.

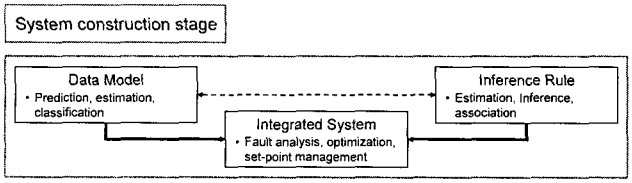

Fig. 7. System construction stage.

3.4 Developed Datamining Roadmap

Figure 6 shows the datamining roadmap proposed in this study. The left side of the roadmap shows datamining procedures and the right side of the roadmap shows each category that includes some proper techniques. The categories were classified according to the objectives of datamining. The function of the roadmap is to help users determine suitable procedures and techniques for knowledge discovery from data. Users who do not have much experience in datamining can attain their goal using the proposed datamining roadmap. However, problems in applying datamining need to be solved by developers. Solutions are not included in the datamining roadmap here. It is difficult to readjust using categories because there are several problems in datamining. Still, the roadmap makes it possible to reduce trial and error.

In this study, three applications were performed to provide solutions to the specific application problems. In the future, if these types of techniques are included in the roadmap, the functions and performance of datamining can be improved.

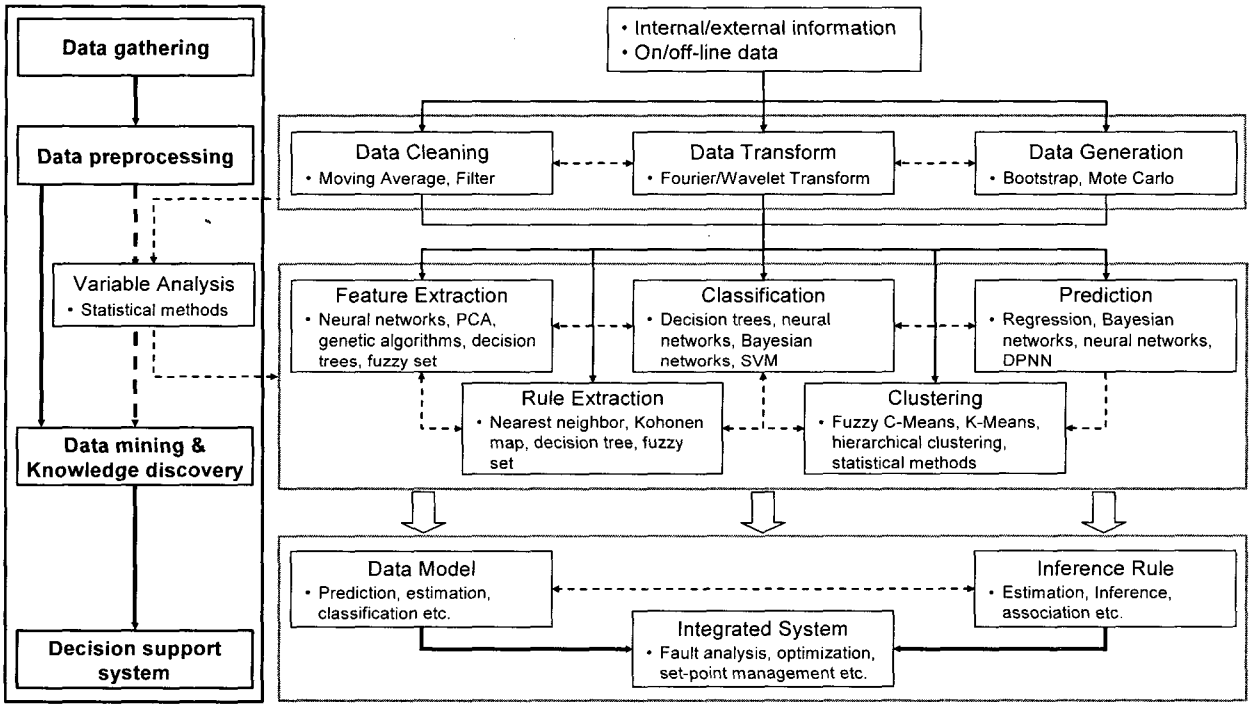

Fig. 6. The proposed datamining roadmap for manufacturing

3.5 Case Studies for Industrial Applications

The roadmap was developed based on several reference books and papers. A suitable category, such as area, can be selected corresponding to goals, and then proper application methods can be determined by checking the techniques in each category. For example, in this study, data generation, prediction, and rule extraction were selected for management of a wafer process; data transform, feature extraction, and classification were selected for diagnosis of induction motors; and data cleaning, clustering, and rule extraction were applied for control of water coagulant.

In the present study, typical application problems were defined and then solutions to those problems by implementation of datamining to industrial processes were proposed. In application to the wafer process, data generation was performed to compensate for insufficient measurement data caused by sampling inspection. The performance of prediction models and inference rules were improved by the proposed data generation. In application to the motor diagnosis, feature selection was performed to determine the adequate features based on fuzzy similarity measure. The fault diagnosis could be implemented under different environmental conditions by the proposed feature selection. Finally, in application to the coagulant control, rule extraction was executed to generate rules from data automatically. A knowledge-based system built by only operators' experience could be supplemented by the proposed rule extraction.

4. Conclusion

In this study, a datamining roadmap was proposed to support datamining applications for industrial processes. The roadmap was classified into three stages and each stage was categorized into reasonable classes according to the purpose of datamining. Each class included representative techniques for datamining that have been broadly applied over decades. The techniques can be different according to developers and application purposes; however, exemplary methods were described in this study. Nonexperts can determine procedures and techniques for datamining based on the datamining roadmap. Thereby, it is possible to reduce trial and error. Still, after determining procedures and methods, there exist application problems.

Reference

[1] R. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, "Mining Business Databases," Communications of the ACM, vol. 39, no. 11, pp. 42-48, 1996.

[2] M. Berry and G. Linoff, Datamining Techniques for Marketing, Sales, and Customer Support, New York: Wiley Computer Publishing, 1997.

[3] J. Han, "Datamining: concepts and techniques," San Francisco: Morgan Kaufmann Publishers, 2001.

[4] P. Adriaans and D. Zantinge, "Datamining," Harlow, England, Reading, Mass.: Addison-Wesley, 1996.

[5] M. A. Bramer (Editor), "Knowledge discovery and datamining," London: The Institution of Electrical Engineers, 1999.

[7] K. J. Cios, "Datamining methods for knowledge discovery," Boston: Kluwer Academic, 1998.

[8] B. V. Dasarathy (Eds.), "Datamining and knowledge discovery: theory, tools, and technology II," Orlando, Florida, Bellingham, Wash.: SPIE, 2000.

Hyeon Bae

He received the M.S. and Ph.D. degree in electrical engineering from Pusan National University in 2001 and 2005, respectively. His research interests include intelligent system and control, data mining, systems bilogy, and bioinformatics.

Youn-Tae Kim

He received the M.S. degree in electrical engineering from Pusan National University in 2005. His research interests include intelligent control.

Sungshin Kim

He received the B.S. and M.S. degrees from Yonsei University, and the Ph.D. degree in electrical engineering from Georgia Institute of Technology, Atlanta, in 1984, 1986, and 1996, respectively. He is currently an Associate Professor in the School of Electrical and Computer Engineering, Pusan National University. His research interests include intelligent control, fuzzy logic control, manufacturing systems, and data mining.

George J. Vachtsevanos

He attended the City College of New York and received his B.E.E. degree in 1962. He received an M.E.E. degree from New York University and his Ph.D. degree in Electrical Engineering from the City University of New York in 1970. His research focused on adaptive control systems. Since joining the faculty at Georgia Tech, he has been teaching courses and conducting research on intelligent systems, robotics and automation of industrial processes and diagnostics/prognostics of large-scale complex systems.