International Journal of Fuzzy Logic and Intelligent Systems, vol. 7, no. 4, December 2007 pp. 279–284

On-line Diagnosis System with Learning Bayesian Networks for fsEBPR

Seong-Pyo Cheon and Sungshin Kim

Department of Electrical Engineering, Pusan National University, Busan, Korea 609-735

Abstract

Nowadays, due to development of automatic control devices and various sensors, one operator can freely handle several remote plants and processes. Automatic diagnosis and warning systems have been adopted in various fields, in order to prepare an operator's absence for patrolling plants. In this paper, a Bayesian networks based on-line diagnosis system is proposed for a wastewater treatment process. Especially, the suggested system is included learning structure, which can continuosly update conditional probabilities in the networks. To evaluate performance of proposed model, we made a lab-scale five-stage step-feed enhanced biological phosphorous removal process plant and applied on-line diagnosis system to this plant in the summer.

Key words : On-line Diagnosis System, Learning Bayesian Networks, fsEBPR Plant

1. Introduction

When the industrial process is born breakdown suddenly, fault detection and faulty cause analysis techniques are important. Fast and accurate diagnosis is not only capable of restarting and repairing process but also guaranteed system's long-life and good health. Therefore, it is a core technology in fault diagnosis that fast finding of faulty cause and prevention of damage extension. Figure 1 shows relations among Wastewater Treatment Plant(WWTP), monitor/control, prevention/diagnosis and consult/advice. Only about ten years ago, one WWTP was maintained by several workers who included an operator, electrical/mechanical engineers and a chemical analyst. Recently, due to drastic development of computer power, sensor technologies, and wideband broadcast communications, one operator handles several remote WWTPs simultaneously in a central control room. However, an automated WWTP has various kinds of additional hardware devices which are electrical/mechanical machinery and complicated sensors. Therefore, to fully utilize these hardware devices, automatic control strategies and data acquisition/analysis softwares are also necessary. In this paper, we suggested a kind of useful software system which is the on-line diagnosis system with a Bayesian network for unmanned WWTPs. In order to study more practically, the lab-scale five-stage step-feed enhanced biological phosphorous removal (fsEBPR) plant had been built, and our on-line diagnosis system was installed on this plant.

2. Learning Bayesian Networks

A Bayesian network is a graphical model that has several advantages for real-world data analysis and finding relationships among variables. Most of all, a Bayesian network based algorithm is regarded as an ideal approach for combining background knowledge and sensory data because of the causal and probabilistic semantics. Therefore, both the knowledge extraction and the rule generation with a Bayesian network approach have been studied and reported on in many papers in various fields [1-7].

A Bayesian network model is usually sketched by directed acyclic graphs (DAGs), in which nodes represent random variables. A Bayesian network is used to estimate target nodes using probabilistic relations of other variables or observed data. The way to find probability distribution for the target variable is said to be *inference*. In general, the computation of a probability of a constructed model is known as *probabilistic inference*. Since 1960s many researchers have illustrated the usefulness and performance of Bayesian probabilistic inference in various fields. Bayesian related studies in applications were categorized into following two topics.

- The powerful graphical presentation ability - It is well-known to real-world system designers.

Manuscript received Sep. 29, 2007; revised Dec. 17, 2007
Corresponding Author : sskim@pusan.ac.kr (Sungshin Kim)
This study was supported by the Second-stage of the Brain Korea 21 proj. in 2007 and ETRI in Korea.

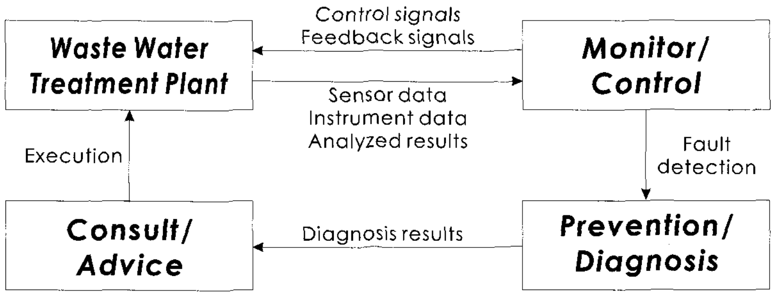


Figure 1: The relations among WWTP, monitor/control, diagnosis, and consult/advice

- The reasonable and flexible inference ability - It is well-suited for development of estimation and decision-support systems.

Some new areas have been suggested for researching the use of Bayesian networks: learning structures and update algorithms, calculation and performance improvement methods, various sampling methods, fast and accurate reasoning methods, and other decision-making methods.

In order to briefly explain a general probabilistic inference procedure, some notations are introduced. Here, let X be the query variable, $\mathbf{E}$ be the set of evidence variables, $\mathbf{e}$ be the observed values, $\mathbf{Y}$ be the remaining unobserved variables, and α be the normalization constant. The query of $\mathbf{P}(X|\mathbf{e})$ can be evaluated as

$$\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X,\mathbf{e}) = \alpha\sum_{\mathbf{y}}\mathbf{P}(X,\mathbf{e},\mathbf{y}), \qquad \mathbf{y}\in\mathbf{Y},\mathbf{e}\in\mathbf{E} \tag{1}$$

where the summation is over all possible $\mathbf{y}$s. Notice that together the variables X, $\mathbf{E}$, and $\mathbf{Y}$ constitute the complete set of variables for the domain; thus $\mathbf{P}(X,\mathbf{e},\mathbf{y})$ is simply a subset of probabilities from the full joint distribution. Therefore, once the a priori probability of a number of variables is specified, it is possible to calculate the priori probabilities for all the nodes in the network. This can be done by utilizing a basic probability calculus and Bayes' Theorem. Once the conditional probabilities of linked variables are specified and the priori beliefs of observable variables are decided, it is possible to calculate the priori probabilities for all the other nodes in the network by Eq. (1). A priori belief is modified as new knowledge about the system is obtained, in the form of an observation of the values assumed by one or more variables. Hence, priori beliefs are substituted by the observation values for these variables. In addition, the beliefs about the others are updated through *belief propagation*[8]. Russell et al[9]. suggested the gradient ascent rule which maximizes $P(D|h)$ by following the gradient of $lnP(D|h)$ with respect to the parameters that define the conditional probability tables of the Bayesian network. Let ω_{ijk} denote a single entry in one of the conditional probability tables. In particular, let ω_{ijk} denote the conditional probability that the network variable Y_i will take on the value y_{ij} given that its immediate parents U_i take on the values given by u_{ik}. The gradient of $lnP(D|h)$ is given by the derivatives $\partial\ln P(D|h)/\partial\omega_{ijk}$ for each of the ω_{ijk}. As we show below, each of these derivatives can be calculated as

$$\frac{\partial\ln P(D|h)}{\partial\omega_{ijk}} = \sum_{d\in D}\frac{P(Y_i=y_{ij},U_i=u_{ik}|d)}{\omega_{ijk}} \tag{2}$$

We require that as the weights ω_{ijk} are updated they must remain valid probabilities in the interval [0,1]. We also require that the $\sum_j\omega_{ijk}$ remains 1 for all i,k. These constraints can be satisfied by updating weights in a two-step process. First, we update each ω_{ijk} by gradient ascent

$$\omega_{ijk} \leftarrow \omega_{ijk} + \eta\frac{\partial\ln P(D|h)}{\partial\omega_{ijk}} \tag{3}$$

where η is a small constant and called the learning rate. Second, we renormalize the weights ω_{ijk} to assure that the above constraints are satisfied. As discussed by Russell et al., this process will converge to a locally maximum likelihood hypothesis for the conditional probabilities in the Bayesian network.

3. Lab-scale Five-stage Step-feed EBPR Plant

To diagnosis WWTP, we made a lab-scale fsEBPR plant at first hand. Figure 2 shows the schematic diagram of the fsEBPR plant. As shown in Fig. 2, our fsEBPR plant has two additional controlable influent water injection flows, typical WWTPs have just one influent water injection flow, so it is not easy to operate the whole processes. However, the fsEBPR is able to be maximized its treatment capacity by control of two additional injection flow rates. This WWTP functionally consists of pre-anoxic process, anaerobic process, first anoxic (dPAO) process, second anoxic process, aerobic process and settler. But physically it is composed of total 15 tanks and 1 settler as shown in Figs. 3 and 4. We have summarized operation conditions of our fsEBPR in table 1.

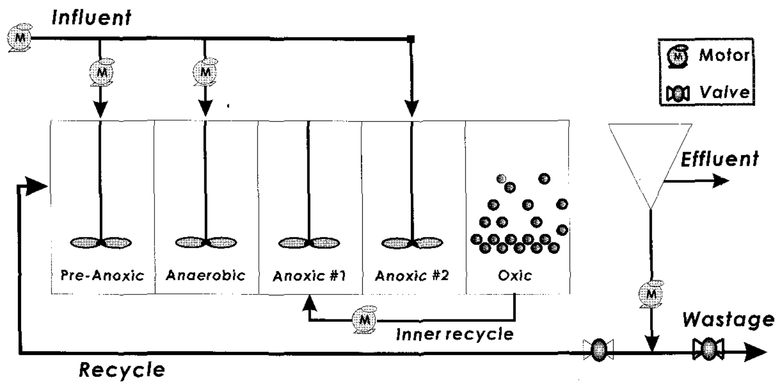


Figure 2: Schematic diagram of five-stage step-feed EBPR process

4. Diagnosis Learning Bayesian networks model for WWTP

The proposed Bayesian network model is expected to have two roles. One role is an effluent water quality predictor and the other role is an artificial WWTP diagnostician. Because effluent water quality is only reported once a day at the most loaded time, proposed Bayesian network is designed to predict phosphate, NH_4 and nitrate concentrations, which are major items included in a daily effluent water quality report, and to diagnose current status of each process with sensor data and real/predicted water quality. In order to achieve more accurate prediction and diagnosis of WWTP, DO, ORP, pH and temperature sensors were installed in each tank. Figure 4 shows the installed sensors, their locations, and measurable data.

In order to design Bayesian networks for WWTP based on related chemical reaction processes and expert knowledge, we designed three Bayesian networks which were related to *Phosphates*, *Nitrates* and *NH_4*, respectively.

The designed three networks are as follows [10-13]:
(i) In the case of *Effluent Nitrates*, *Effluent Nitrates* are directly related with the nitrification process. Thus, *DO*, *Carbon*, *Alkalinity*, and *Ax mixing* are parent nodes of *Effluent Nitrates*.
(ii) In the case of *Effluent* NH_4, *Effluent* NH_4 is inversely proportional to a denitrification reaction. Thus, considering the denitrification process, the NH_4 *Load* is source material. Furthermore, *DO*, *Aero pH*, *Temperature*, *Ox mixing*, and *SRT* are components of reaction conditions for denitrification.
(iii) In the case of *Effluent P*, it is closely related with the EBPR process. The EBPR process has two kinds of reactions, one occurring under aerobic conditions and the other under anaerobic conditions. Thus, the design of an *Effluent P* related network is very complicated. So, instead of directly imitating the EBPR process, we try to create node relations by combination of indirect indices. Finally, an *Effluent P* node is only connected with *MLSS*, *DO*, and *SRT* nodes.

Node's relations are shown in Fig. 5. Of course, initial probabilities are calculated by their correlated node's conditional probabilities and Bayes' Theorem. This model is presented in related expert knowledge, chemical process

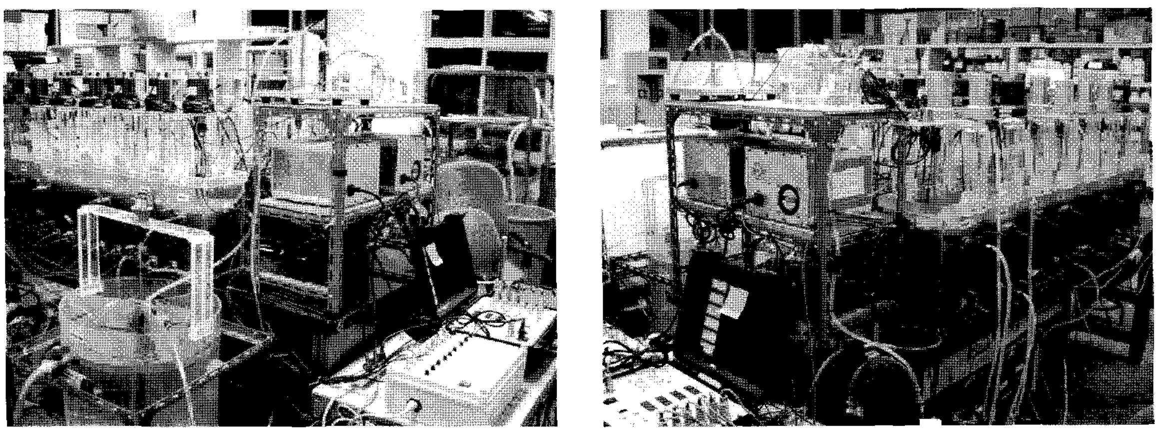


Figure 3: Lab-scale Five-stage Step-feed EBPR Process

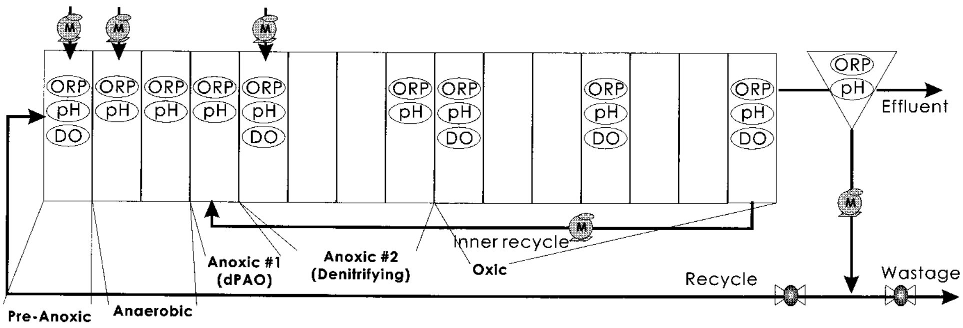


Figure 4: Installed sensors' location and measured data in fsEBPR

reactions, and detail causes and effects among whole equipment and devices. Initial data for the conditional probabilities setting was obtained from a pilot-scale plant, which had been operated as fsEBPR process for 655 days from February, 2000 to November, 2001. This data was also acquired from a similar fsEBPR process but its operational conditions had been changed by its influent water characteristics, degree of process stability and activated sludge status etc,. Therefore, we should take two-step probabilities setting procedures: definitely related nodes had probabilities calculated from initial data; The other uncertain nodes were decided by operator's experience and common knowledge.

5. Experiment Results

On-line diagnosis system for fsEBPR process is implemented as shown in Fig. 6. In Fig. 6, white nodes are just estimated nodes but patterned nodes are estimated and checked ones. Because patterned nodes have their own sensor(s), we can make a check and diagnosis status on each sensor value. Therefore, conditional probability tables about white nodes cannot be updated using learning strategy. We evaluate our suggested Bayesian network model using lab-scale plant data for about three months from July to September, 2006. This period is a summer in Korea and it is generally hard to operate the plant because high temperatures seriously influence a small size lab-scale plant. Table 2 shows influent characteristics of the lab-scale plant. We try to predict and diagnose with two kinds of Bayesian network. One is a BN model which is a Bayesian network without learning structure and the other is a LBN model which is a learning Bayesian network with the gradient ascent algorithm which has a learning rate = 0.2. According to our lab-scale operational data, our plant had 18 days of at least one component abnormal effluent water quality, and 51 days with at least one sensor data out of the normal range. After we applied the prediction and diagnosis procedures to our evaluation data, we summarized the prediction and diagnosis results from two models as shown in table 3. Table 4 shows monthly diagnostic results. Both the BN model and the LBN model are set up using other large-scale plant operation data. Thus, their diagnosis results are similar in the beginning of July. However, in August, each model's diagnosis results are very different because LBN

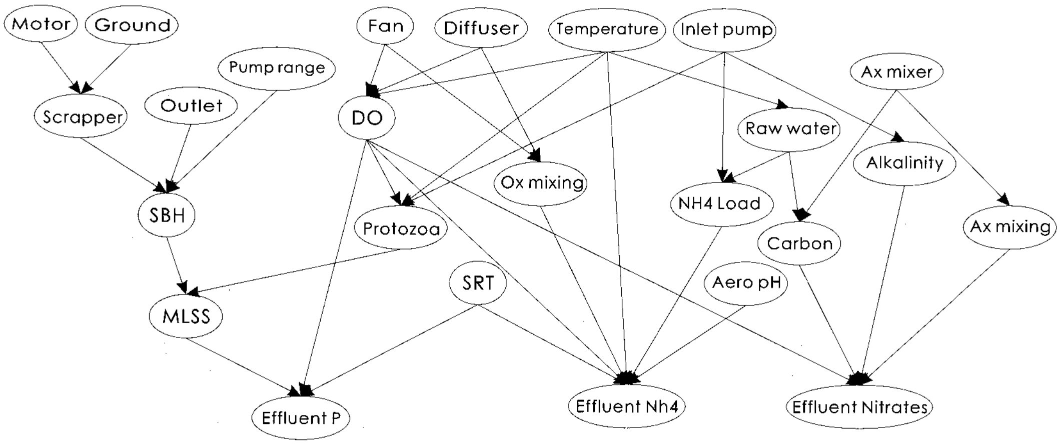


Figure 5: The Diagnosis Bayesian network Model for WWTP

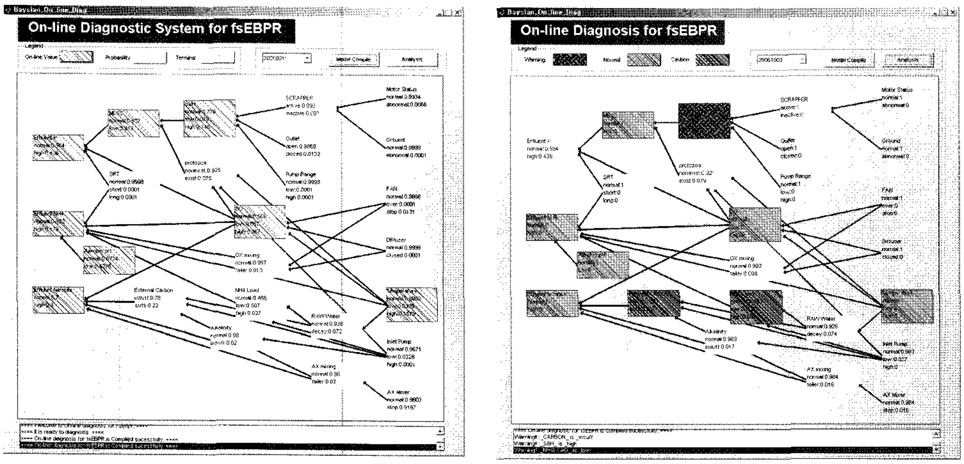


Figure 6: GUI of On-line Diagnosis System for fsEBPR Process

model probabilities are updated during the previous month. As can clearly be seen for *Carbon* and *NH_4* elements, the BN model warns of *Carbon* and *NH_4* states 25 times, while the LBN model never warns of these elements. This is because initial probabilities are biased about *Carbon* and *NH_4*. On the other hand, in considering *SBH, DO, Temperature*, and *pH-level* elements, the LBN model is very stable and reliable. We could not evaluate fault detection ability for the models regarding devices and sensors, because we never detected a device fault or an equipment error in three months. However, according to the process diagnostic performance of the LBN model, its diagnostic ability is also ensured.

6. Conclusion

The lab-scale fsEBPR process had been built at first hand and it has been operated for about 1 year, before we implemented the prediction and diagnostic system. For operation periods, frequent unexpected abnormal conditions and device faults were detected, which is the motivation for our study. In order to suggest a practical and effective diagnostic algorithm for the fsEBPR, we decided to implement a diagnostic model with a Bayesian network. Most importantly, we adopted a Bayesian approach because it is based on a strict foundation of probability theory, it has easy to understand diagnostic results, and it is guaranteed to provide reasonably accurate diagnostic results. In addition, its results are stable and robust against noisy sensor data and disturbance. In this paper, we suggest two kinds of Bayesian network models. One is a Bayesian network model without an additional probability update strategy and the other is a learning Bayesian network model. In fact, we suggested two models because we could not set up the initial probabilities with our own plant data. Because it usually requires a few months to set up and operate desired operational conditions in a WWTP, we had to obtain operation data from other plants. Thus, we had to adjust and optimize diagnostic models using the learning method. However, adjustment and optimization procedures are not extraordinary, and most diagnostic systems should adopt

Table 4: Each monthly frequency of warning/caution elements for the BN model and LBN model

	Bayesian model				**Learning Bayesian model**			
	July	Aug.	Sep.	Sum	July	Aug.	Sep.	Sum
SBH	11	17	2	30	10	5	-	15
DO	5	12	2	19	5	5	-	10
External Carbon	15	25	2	42	3	-	-	3
NH_4 Load	15	25	2	42	3	-	-	3
Temperature	7	7	-	14	7	5	-	12
Aerobic pH	9	-	-	9	7	-	-	7
MLSS	1	-	-	1	1	-	-	1
SUM	63	86	8	157	36	15	-	51

these procedures. In conclusion, the LBN model is superior to BN model and we verified that the LBN model is well-adapted to the diagnosis field. As a modeling method, Bayesian approaches have some basic limitations. First, whole network design procedures entirely depend on operator knowledge and experience regarding WWTPs. Second, if there are insufficient data, reliability of the model drastically decreases. To overcome insufficient data problems, many researchers have suggested various methods like the MCMC (Markov Chain Monte Carlo) method [14] and Particle filter method [15, 16]. In the near future, we will suggest more robust and reliable diagnostic methods and compare the diagnostic performance among various models.

References

[1] S. Andreassen, M. Woldbye, B. Falck, S. K. Andersen, "MUNIN - A Causal Probabilistic Network for Interpretation of Electromyographic Findings," *Proc. 10th Int'l Joint Conf. Artificial Intelligence,* pp.366-372, Aug. 1987.

[2] C. Chien, S. Chen, and Y. Lin, "Using Bayesian Network for Fault Location on Distribution Feeder," *IEEE Trans. Power Delivery,* vol. 17, no. 13, pp.785-793, July 2002.

[3] I. J. Good, "Amount of Deciding and Decisionary Effort," *Information and Control,* vol. 4, no.2-3, pp.271-281, 1961.

[4] D. Heckerman and J. S. Breese, "Causal Independence for Probability Assessment and Inference Using Bayesian Networks," *IEEE Trans System, Man, and Cybernetics-Part A: Systems and Humans,* vol. 26, no. 6, pp.826-831, Nov. 1996.

[5] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Toward normative expert systems: The pathfinder project," *Technical Report KSL-91-44,* Knowledge Systems Laboratory, Stanford University, 1991.

[6] E. Horvitz, "Lumiere Project: Bayesian Reasoning for Automated Assistance," available at http://research.microsoft.com/ horvitz/lum.htm, *Adaptive Systems &Interaction,* Microsoft Research, Redmond, Washington, 1998.

[7] J. H. Kim and J. Pearl, "CONVINCE: a conversational inference consolidation engine," *IEEE Trans. Systems, Man and Cybernetics,* vol. 17 , no. 2, pp.120 - 132, Mar./Apr. 1987.

[8] D. Heckerman, "A Tutorial on Learning with Bayesian networks," *Technical report Microsoft Research Advanced Technology Division,* Nov. 1996.

[9] S. Russell, P. Norvig, *Artificial Intelligence A Modern Approach,* second ed. Pearson Education, Inc., Prentice Hall, 2003.

[10] M. Henze, C. D. L. Grady, W. Gujer, M. C. M. Van Loosdrecht, and T. Matsuo "Activated Sludge Model No.1," *IAWPRC,* London, 1987.

[11] G. Olsson and B. Newell, *Wastewater Treatment Systems Modelling, Diagnosis and Control,* IWA Publishing, 1999.

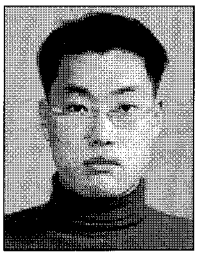
[12] H. Spanjers, P. A. Vanrolleghem, G. Olsson, P. L. Dold "Respirometry in Control of the Activated Sludge Process: Principles," *IAWQ Task Group on Respirometry,* International Association on Water Quality, 1998.

[13] P.F. Strom, "Introduction to phosphorus removal," Invited Presentation for Wastewater Treatment Operators Workshop, 91st Annual Meeting, NJWEA, Atlantic City, NJ, 2006.

[14] C. Yuan, and M. J. Druzdzel, "Importance sampling algorithms for Bayesian networks: Principles and performance," *Mathematical and Computer Modelling,* vol. 43, pp.1189-1207, May 2006.

[15] D. Guo and X. Wang, "Dynamic Sensor Collaboration Via Sequential Monte Carlo," *IEEE Journal on Selected Areas in Communications,* vol. 22, no. 6, pp.1037-1047, Aug. 2004.

[16] D. Nikovski, "Constructing Bayesian Networks for Medical Diagnosis from Incomplete and Partially Correct Statistics," *IEEE Trans. Knowledge and Data Eng.,* vol. 12, no. 4, pp.509-516, July/Aug. 2000.



Seong-Pyo Cheon received the B.S. and M.S. degree in Electrical Engineering from Pusan National University Busan, Korea, in 1999, and 2001, respectively. He had worked for LG CNS in Korea from 2001 to 2004. Mr.Cheon is currently pursuing a doctoral degree in Information and Intelligent System from Pusan National University. His research interests include intelligent system, machine learning, and data mining.

Sungshin Kim received the B.S. and M.S. degrees in Electrical Engineering from Yonsei University Seoul, Korea, in 1984, and 1986, respectively. He received the Ph.D. degree in electrical engineering from Georgia Institute of Technology, Atlanta, in 1996. 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.