A Simultaneous Design of TSK - Linguistic Fuzzy Models with Uncertain Fuzzy Output

Keun-Chang Kwak*, and Dong-Hwa Kim**

* Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
(Tel : +1-780-437-3462; E-mail: kwak@ece.ualberta.ca)

**Department of Instrumentation and Control Engineering, Hanbat National University, Daejeon, Korea
(Tel : +82-42-000-0000; E-mail: kimdh@hanbat.ac.kr)

Abstract: This paper is concerned with a simultaneous design of TSK (Takagi-Sugeno-Kang)-linguistic fuzzy models with uncertain model output and the computationally efficient representation. For this purpose, we use the fundamental idea of linguistic models introduced by Pedrycz and develop their comprehensive design framework. The design process consists of several main phases such as (a) the automatic generation of the linguistic contexts by probabilistic distribution using CDF (conditional density function) and PDF (probability density function) (b) performing context-based fuzzy clustering preserving homogeneity based on the concept of fuzzy granulation (c) augment of bias term to compensate bias error (d) combination of TSK and linguistic context in the consequent part. Finally, we contrast the performance of the enhanced models with other fuzzy models for automobile MPG predication data and coagulant dosing process in a water purification plant.

Keywords: TSK-linguistic fuzzy model, linguistic context, context-based fuzzy clustering, bias term, automobile MPG data, water purification plant

1. INTRODUCTION

Recently, fuzzy modeling is a popular computing framework on the basis of the concepts of fuzzy rules and fuzzy reasoning [1]. The essence of fuzzy modeling is concerned with the development of relationships between information granules regarded as fuzzy sets or fuzzy relations. Over the past few decades, it has found successful application in a wide variety of fields, emerging as interesting, attractive, and powerful modeling environment. There have been a substantial number of various schemes of fuzzy modeling along with specific algorithmic variation that help eventually capture some characteristics of the problem [2-3]. Transparency and accuracy of the models are the two essential and equally important pillars of fuzzy models. While the accuracy has been addressed in many methods, the issue of transparency and interpretability is still quite open. Interpretability implies a certain level of granularity of basic constructs on the basis of information granules. In particular, this was used in case of Linguistic Models (LM) proposed by Pedrycz [4] which ultimately dwells on the concept of information granules. The goal of this linguistic model is to design a new category of fuzzy models that focus on designing meaningful linguistic labels in the space of experimental data [4]. For this purpose, fuzzy clustering plays an important role in the design of linguistic models. Although the effectiveness of linguistic model has demonstrated, this model has a poor approximation and generalization ability, biased error, and a difficulty of design due to uniform form of linguistic contexts constructed in the output domain. Therefore, we develop a TSK (Takagi-Sugeno-Kang)-based Linguistic Fuzzy Model (TSK-LFM) with uncertain model output to solve several problems mentioned above. For this purpose, we use several main stages such as (a) the automatic generation of the contexts by probabilistic distribution (b) fuzzy rule extraction based on Context-based Fuzzy C-Means (CFCM) clustering [5][6] (c) addition of bias term to reduce approximation error (d) combination of TSK and linguistic context in the consequent part. Finally, the experimental results reveal that the proposed model yields better performance in comparison with linear regression, conventional LM, and Radial Basis Function Neural Networks (RBFNN) for coagulant dosing process in a water purification plant [7] and automobile MPG prediction data.

2. TSK-LINGUISTIC FUZZY MODEL

In this section, we briefly describe the underlying concept and architectural fundamentals of linguistic models as originally introduced by Pedrycz in [4]. In contrast to the currently existing schemes of neuro-fuzzy models, which are in essence nonlinear numeric models, linguistic modeling revolves around information granules with fuzzy sets constructed in input and output spaces. The emphasis is on the formation of these granules while the linkages between them are intuitively straightforward as being the result of the construction of the information granules themselves. The conditional (context-based) fuzzy clustering forms a backbone of the linguistic model. Let us concentrate on the enhanced architecture and relate it to the development of the linguistic models. For simplicity, we assume that the TSK-LFM under consideration has two input x and y and one output z. The TSK-LFM architecture is shown in Fig. 1. Here the number of context "p", t=1,2,…, p, and the number of cluster per context "c", i=1,2,…,c. Moreover, we assume that the number of the cluster center in each context is equal.

2.1 Linguistic contexts in the output domain

The linguistic contexts are used to extract fuzzy rules in the CFCM clustering. In the conventional LM [4], these contexts were generated through a series of triangular membership functions with equally spaced along the domain of an output variable [4]. However, we may encounter a data scarcity problem due to small data included in some linguistic context. Thus, this problem brings about the difficulty to obtain fuzzy rules from the CFCM clustering. Therefore, we use probabilistic distribution of output variable to produce the flexible linguistic contexts.

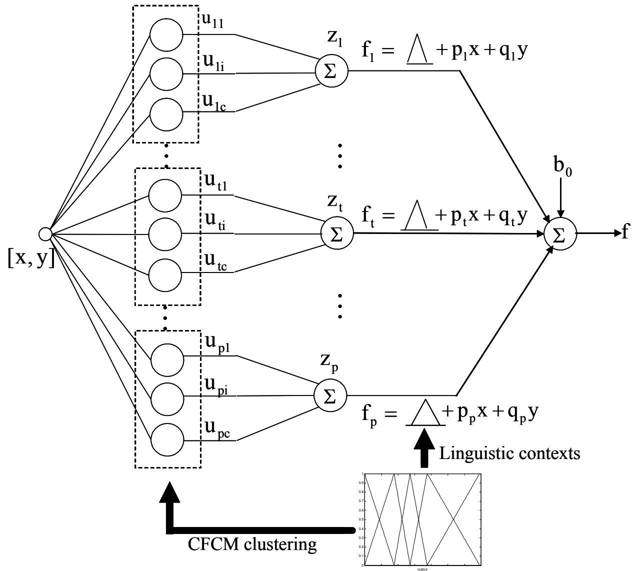

Fig. 1 Architecture of TSK-LFM

Fig. 2 and 3 show an example of output data and its histogram, respectively. We shall explain the well-known automobile MPG to be considered in the experimental part. The output variable to be predicted in terms of the preceding six input variables is the automobile's fuel consumption in MPG.

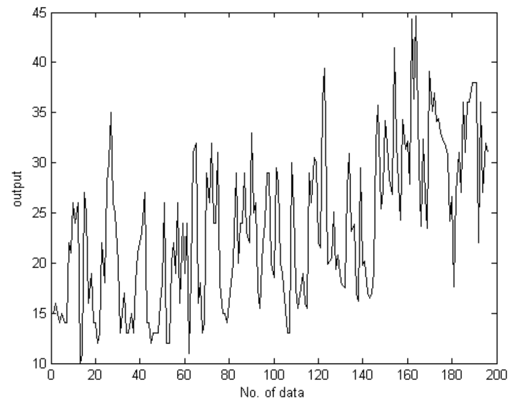

Fig. 2 output data (MPG)

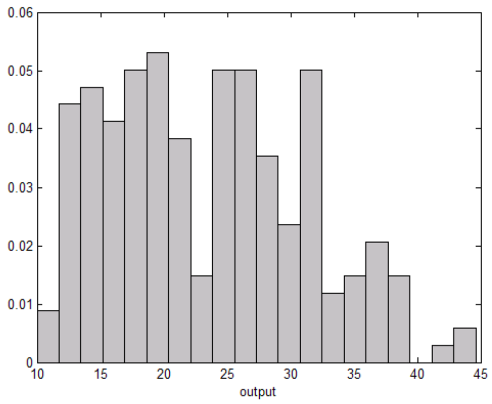

Fig. 3 histogram of MPG data

As shown in Fig. 3, we may encounter a data scarcity problem in the right area when we use equally spaced linguistic contexts along the domain of an output variable. To solve this problem, we construct flexible linguistic context based on probability distribution. Fig. 4-6 visualize the probability density function (PDF), conditional density function (CDF), and the linguistic contexts, respectively. As shown in Fig. 6, the linguistic contexts are produced by probabilistic distribution when p=6. Therefore, we can solve the difficulty to obtain fuzzy rules using the CFCM clustering.

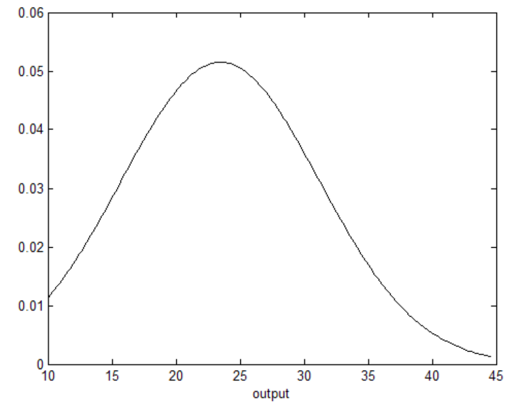

Fig. 4 PDF of MPG data

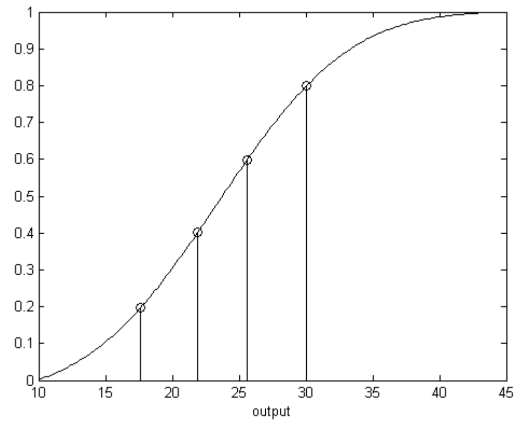

Fig. 5 CDF of MPG data

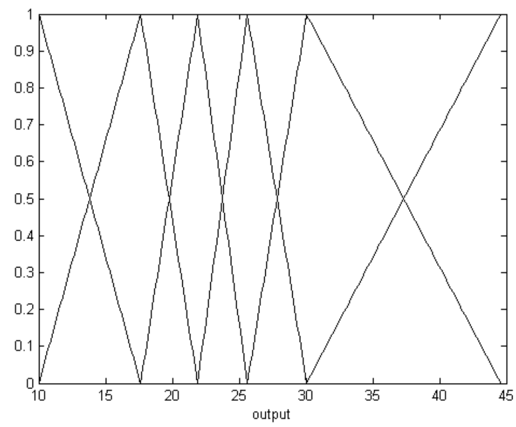

Fig. 6 linguistic contexts

2.2 Context-based FCM (CFCM) clustering

The CFCM clustering method, as proposed by Pedrycz [4][5][6], is an effective approach to estimate the cluster centers. As shown in Fig.1, CFCM clustering is performed in the second layer. The optimization completed by the CFCM clustering is realized iteratively by updating the partition matrix and the prototypes. The update of the partition matrix is completed as follows

$$u_{ik} = \frac{w_{tk}}{\sum_{j=1}^{c}\left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|}\right)^{\frac{2}{m-1}}} \quad i = 1,2,\ldots,c,\ \ k = 1,2,\ldots,N \tag{1}$$

where u_{ik} represents the element of the partition matrix induced by the i-th cluster in the t-th context. Here w_{tk} denotes a membership value of the k-th data to the t-th context. The prototypes are calculated in the form

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^m} \tag{2}$$

where the fuzzification factor "m" is taken as 2.0. For further details on the CFCM clustering, see [5][6]. When applying the CFCM clustering to numerical input-output data pairs, each of the cluster centers presents a prototype that exhibits certain characteristics of the system to be modeled. Fig. 7 displays the blueprint of the linguistic model equipped with three contexts and two clusters per context.

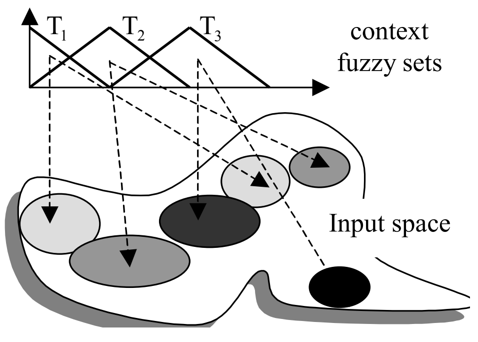

Fig. 7 A blueprint of the linguistic model

2.3 TSK- linguistic type in the consequent part

The conventional LM was designed by linguistic contexts in the consequent part. Although these contexts give meaningful linguistic labels, the obtained results did not show a good performance. Meanwhile, TSK type is by far the most popular candidate for fuzzy modeling and effective to develop a systematic approach [8]. Based on these two complementary approaches, we propose the TSK-based linguistic type in the consequent part as shown in Fig.1. Thus, the TSK-LFM can possess the intensive computation ability together with meaningful linguistic labels. The t-th linguistic context is visualized in Fig. 8.

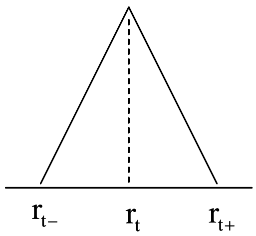

Fig. 8 The t-th linguistic context in consequent part

where $[r_{t-}, r_t, r_{t+}]$ is a 3-element vector that determines the break points of this membership function. Thus the t-th consequent part combined with TSK-type is expressed as follows

$$f_t = [r_{t-}, r_t, r_{t+}] + p_t x + q_t y \tag{3}$$

Here the parameters of linguistic contexts are obtained by probabilistic distribution as mentioned before. The linear coefficients $\{p_t, q_t\}$ of TSK-type are estimated by Least Square Estimator (LSE) [10].

2.4 Uncertain fuzzy output

The results obtained by conventional LM have showed a biased prediction error. This problem brings about a poor approximation and generalization ability. Therefore, we add bias term to conventional LM so that the TSK-LFM can be obtained unbiased prediction error. The bias term is computed in a straightforward manner by the difference between target output and predicted output as follows

$$b_0 = \frac{1}{N}\sum_{k=1}^{N}\left(t\arg et_k - predict_k\right) \tag{4}$$

where $predict_k$ denotes a modal value of fuzzy number produced for k-th input data point. The resulting fuzzy number with bias term is expressed as the following form

$$f = z_1 \otimes f_1 \oplus \cdots \oplus z_p \otimes f_p \oplus \cdots \oplus z_t f_t + b_0 \tag{5}$$

We denote the algebraic operations by $\otimes$ and $\oplus$ to emphasize that the underlying computation operates on a collection of fuzzy numbers. Given the multiplication and addition for two operations, the final fuzzy number (model output) is computed as follows

$$f = \sum_{t=1}^{p} z_t f_t + b_0 = \sum_{t=1}^{p} z_t (r_t + p_t x + q_t y) + b_0 \tag{6}$$

Furthermore, the lower and upper bound of model output are computed by the following form

$$f_{-} = \sum_{t=1}^{p} z_t (r_{t-} + p_t x + q_t y) + b_0 \tag{7}$$

$$f_+ = \sum_{t=1}^{p} z_t (r_{t+} + p_t x + q_t y) + b_0 \quad (8)$$

Based on these bounds, we can represent the uncertain fuzzy output represented by fuzzy number.

3. EXPERIMENTS AND RESULTS

3.1 Automobile MPG prediction

We shall use the well-known automobile MPG data as an nonlinear regression example. In this example, six input variables consist of cylinder number, displacement, horsepower, weight, acceleration, and model year. The output variable to be predicted in terms of the preceding six input variables is the automobile's fuel consumption in MPG. The data set consists of 392 examples of different car makes after removing instances with missing values. We divide the data set into training (odd number) and test data sets (even number) in the normalized space between 0 and 1, respectively. The training data set is used for model construction, while the test set is used for model validation. Thus, the resultant model is not biased toward the training data set and it is likely to have a better generalization capacity to new data [11][12]. Fig. 9 and 10 show the RMSE versus the number of cluster (2~10) in each context for training and test set, respectively. As shown in these figures, the best model we can achieve occurs when the test error is minimal (p=6, c=2) [10]. Table 1 summarizes the RMSE in comparison with other models. The RBFNN used in Table 1 was designed by the centers of the receptive field functions using FCM clustering. The weights were estimated through LSE method. Fig. 11 shows the RMSE obtained by RBFNN for training and test data. In the design of conventional LM, we encountered a data scarcity problem due to small data included in of right side of linguistic context. Thus, we obtained the best model when p=6 and c=5. As listed in Table 1, we can recognize from the results that the TSK-LFM outperformed the previous works. Fig. 12 and 13 visualize the actual output and the uncertain model output represented by lower and upper bound for training and test data, respectively. As shown in these figures, the TSK-LFM showed the good approximation and generalization ability.

Table 1 Comparison of RMSE (*: no. of hidden nodes)
(Type1: LM with bias term and flexible contexts, type2: LM with bias term, flexible contexts, and TSK-based linguistic type)

		[p, c]	RMSE (training)	RMSE (test)
Linear regression		.	3.452	3.444
RBFNN		39*	3.211	3.293
LM [4]		[6,5]	4.273	4.368
	Type1	[6,2]	3.224	3.080
	Type2	[6,2]	2.717	3.048

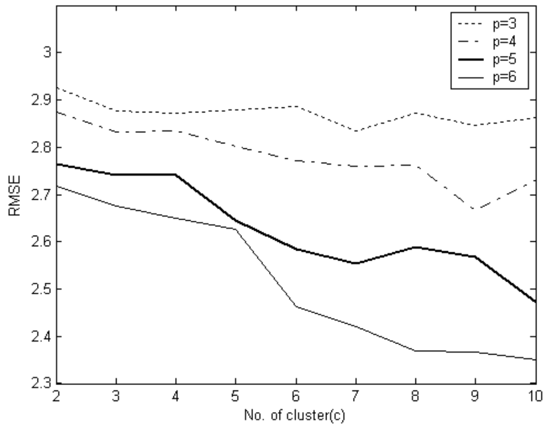

Fig. 9 RMSE by variation of "p" and "c"(training data)

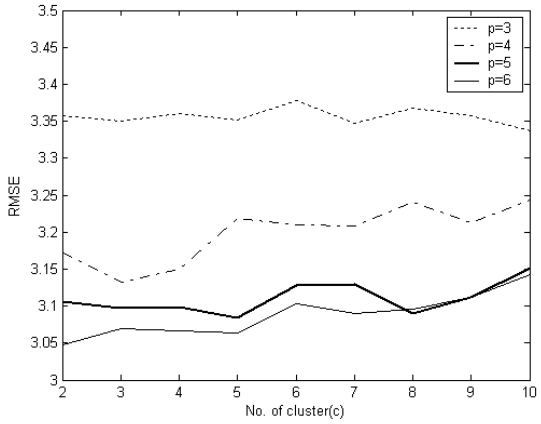

Fig. 10 RMSE by variation of "p" and "c"(test data)

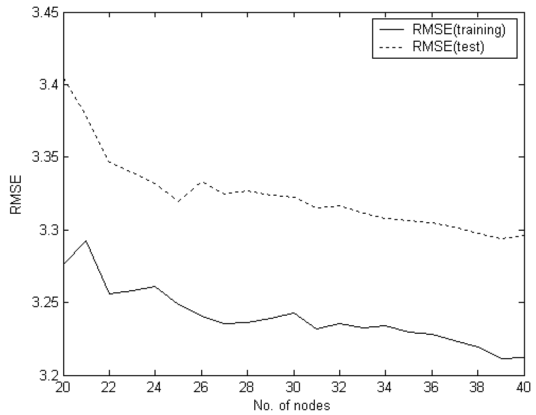

Fig. 11 RMSE obtained by RBFNN

3.2 Coagulant dosing process in a water purification plant

The field test data of a coagulant dosing process to be modeled is obtained at the Amsa water purification plant, Seoul, Korea, having a water purification capacity of 1,320,000 ton/day. We use the successive 346 samples among jar-test data for the past one year. The input variables consist of the turbidity of raw water, temperature, pH, alkalinity, and so on. The output variable to be predicted in terms of the preceding input attributes is PAC (Poli-Aluminum Chloride) widely used as a coagulant. In order to evaluate the resultant model, we divide the data sets into training and checking data sets. Here, we choose 173 training sets for model construction, while the other test sets are used for model validation.

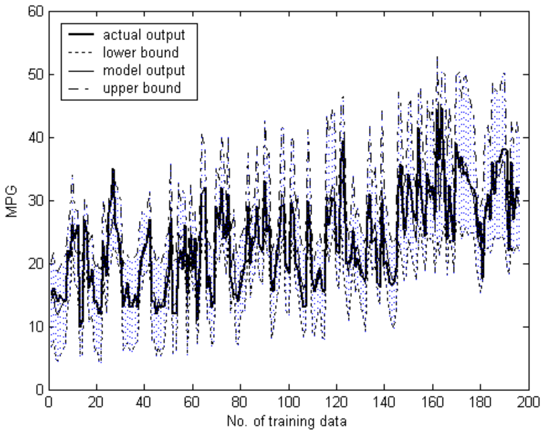

Fig. 12 Approximation ability of the TSK-LFM (training data)

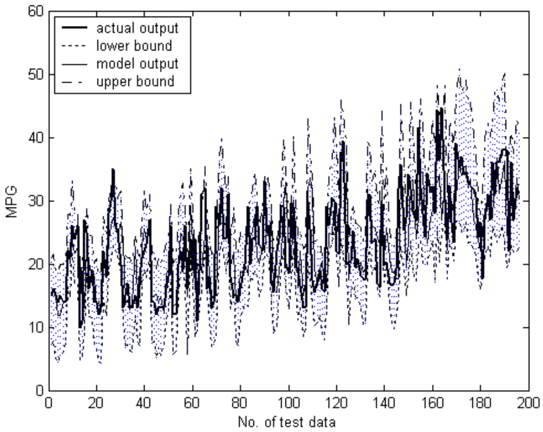

Fig. 13 Generalization ability of the TSK-LFM (test data)

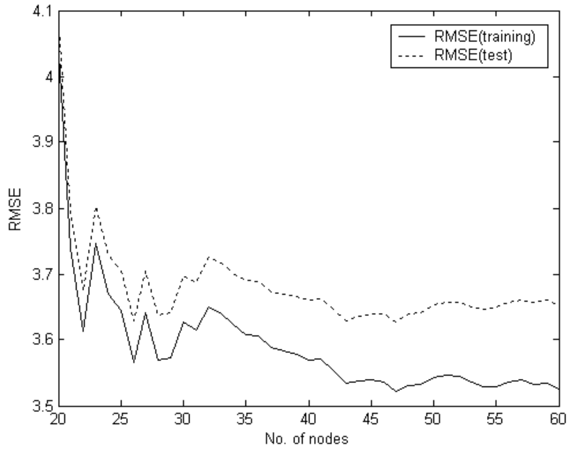

Fig. 14 RMSE obtained by RBFNN

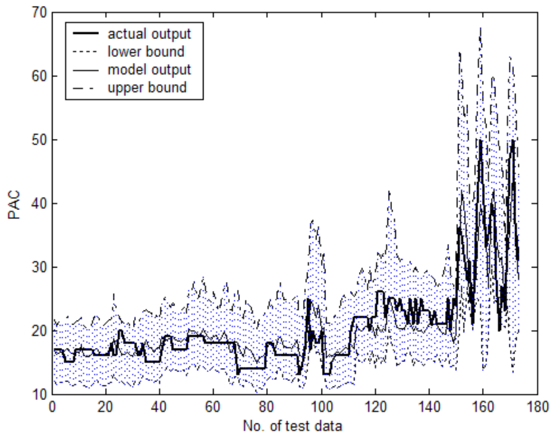

Fig. 15 Generalization ability of the TSK-LFM (test data)

Table 2 summarizes the RMSE in comparison with other models. In the design of conventional LM, we encountered a data scarcity problem due to small data included in side linguistic context. Thus, we obtained the best model when p=5 and c=5. As listed in Table 2, we can recognize from the results that the TSK-LFM showed a better performance in comparison with the previous works. Fig. 14 shows the RMSE obtained by RBFNN for training and test data. Fig. 15 visualizes the actual output and the uncertain model output represented by lower and upper bound for test data.

Table 2 Comparison of RMSE (*: no. of hidden nodes)

		[p, c]	RMSE (training)	RMSE (test)
Linear regression		.	3.508	3.578
RBFNN		48*	3.522	3.628
LM [4]		[5,5]	3.725	3.788
	Type1	[5,9]	2.965	3.123
	Type2	[5,9]	2.514	2.661

4. CONCLUSIONS

We have developed the comprehensive design framework of linguistic model introduced by Pedrycz. Thus we have proposed the TSK-LFM with uncertain model output represented by fuzzy number. The results obtained by TSK-LFM showed a better performance than other models for coagulant dosing process in water purification plant and automobile MPG data. The TSK-based linguistic type used in this study provided the meaningful linguistic representation of Mamdani fuzzy model as well as intensive computation ability of Sugeno fuzzy model. Furthermore, we confirmed the effectiveness through the bias term and flexible linguistic contexts.

ACKNOWLEDGMENTS

This work was also supported by EESRI (R-2003-0-285).

REFERENCES

[1] L. A. Zadeh, "Toward a theory of fuzzy systems," *In Aspects of Network and System Theory*, R.E Kalman and N. De Claris, Eds. New York: Holt, Rinehart & Winston, 1971.

[2] G. J. Klir and T. A. Folger, *Fuzzy Sets, Uncertainty, and Information*, Englewood Cliffs, NJ: Prentice Hall, 1988.

[3] R. Kruse, J. Gebhardt, and F. Klawonn, *Foundations of Fuzzy Systems*, New York: Wiley, 1994.

[4] W. Pedrycz and A. V. Vasilakos, "Linguistic models and linguistic modeling," *IEEE Trans. on Systems, Man, and Cybernetics - Part C*, Vol. 29, No. 6, pp.745-757, 1999.

[5] W. Pedrycz, "Conditional fuzzy C-Means," *Pattern Recognition Letters*, Vol. 17, pp. 625-632, 1996.

[6] W. Pedrycz, "Conditional fuzzy clustering in the design of radial basis function neural networks," *IEEE Trans. on Neural Networks*, Vol. 9, No. 4, pp. 601-612, 1998.

[7] M. G. Chun, K. C. Kwak, and J. W. Ryu, "Application of ANFIS for a coagulant dosing process in a water purification plant," *IEEE International Fuzzy Systems Conference Proceedings*, Vol. 3, pp. 1743-1748, Seoul, Korea, 1999.

[8] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 15, pp. 116-132, 1985.

[9] D. Wettschereck and T. Dietterich, "Improving the performance of radial basis function networks by learning center locations," *In J. E. Moody, editor, Advances in Neural Information Processing Systems 4*, pp. 1133-1140, San Mateo, CA, 1992.

[10] J. S. R. Jang, C. T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice Hall, 1997.

[11] J. S. R. Jang, "Input selection for ANFIS learning," *Proceedings of IEEE Int. Conference on Fuzzy Systems*, New Orleans, pp.1493-1499, 1996.

[12] J. S. R. Jang, "ANFIS: Adaptive-Network-based Fuzzy Inference Systems," *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 23, No. 3, pp.665-685, 1993.