LETTER

A Design of Genetically Optimized Linguistic Models

Keun-Chang KWAK[†a)], ***Member***

SUMMARY In this paper, we propose a method for designing genetically optimized Linguistic Models (LM) with the aid of fuzzy granulation. The fundamental idea of LM introduced by Pedrycz is followed and their design framework based on Genetic Algorithm (GA) is enhanced. A LM is designed by the use of information granulation realized via Context-based Fuzzy C-Means (CFCM) clustering. This clustering technique builds information granules represented as a fuzzy set. However, it is difficult to optimize the number of linguistic contexts, the number of clusters generated by each context, and the weighting exponent. Thus, we perform simultaneous optimization of design parameters linking information granules in the input and output spaces based on GA. Experiments on the coagulant dosing process in a water purification plant reveal that the proposed method shows better performance than the previous works and LM itself.

key words: *linguistic model, context-based fuzzy c-means clustering, genetic algorithm, coagulant dosing process*

1. Introduction

During the past decades, a considerable number of studies have been conducted on Fuzzy Model (FM), due to the rapid growth in the variety of applications. The diversity of existing models is well documented in the literature with various methodologies and architectures. While FM accuracy has been target of many methods, the issue of transparency and interpretability is still quite open. Interpretability implies a certain level of granularity of basic constructs, the so called information granules. By changing their level of specificity, interpretability and accuracy requirements in FM can be developed [1]. From this point view, fuzzy clustering performs a central role in the design of FM. For this purposes, various clustering techniques have applied to structure identification in neural networks and fuzzy modeling [2]–[5].

However, these clustering techniques are performed by context-free clustering method without considering the homogeneity between input and output spaces. In contrast to these context-free clustering methods, the objective of context-based clustering is to generate clusters preserving homogeneity of the clustered patterns in connection with their similarity in the input variables as well as in the output variable based on linguistic contexts. The effectiveness of this context-based fuzzy clustering technique has been demonstrated in previous works [6]–[11]. These models represented a nonlinear and complex characteristic more effectively than conventional models based on context-free clustering. However, it is difficult to optimize the number of context, the number of cluster generated by each context, and weighting exponent.

Therefore, the objective of this study is to pursue the systematic development of genetically optimized linguistic models with the use of fuzzy granulation. The Genetic Algorithm (GA) is a derivative-free stochastic optimization method based on the concepts of natural selection and evolutionary processes. Based on GA, we perform simultaneous and parallel optimization of parameters that are the cause of the design problem in the conventional LM (Linguistic Model) [12]. The performance of GA-based LM when applied to the coagulant dosing process in a water purification plant [11] is contrasted with that of LR (Linear Regression), MLP (Multilayer Perceptron), RBFN (Radial Basis Function Networks), TSK (Takagi-Sugeno-Kang)-LFM [11], and LM itself [8].

2. Linguistic Models with the Use of Fuzzy Granulation

2.1 Context-Based Fuzzy C-Means (CFCM) Clustering

CFCM clustering is an effective approach to estimate the cluster centers preserving homogeneity on the basis of fuzzy granulation. In contrast to the context-free clustering methods, the CFCM clustering method is performed with the aid of the contexts produced in output space. By taking into account the contexts, the clustering in the input space is focused by some predefined fuzzy sets of contexts. Let us introduce a family of the partition matrices induced by the t-th context as follows

$$\boldsymbol{U}(W_t) = \left\{ u_{ik} \in [0,1] \,\middle|\, \sum_{i=1}^{c} u_{ik} = w_{tk} \ \forall k \right\} \tag{1}$$

where w_{tk} denotes a membership value of the k-th data point included by the t-th context. The underlying objective function can be expressed as follows

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 \tag{2}$$

where $\|\ .\ \|$ is the Euclidean distance and $\boldsymbol{v}_i$ denotes the i-th cluster. Here we perform the separate clustering tasks implied by the corresponding context. The minimization of objective function is realized by iteratively updating the values of the membership matrix and the prototypes. The update of the membership matrix is computed as follows

Manuscript received April 5, 2012.
Manuscript revised July 27, 2012.

[†]The author is with the Dept. of Control, Instrumentation, and Robotic Eng., Chosun University, Gwangju, 501–759 Korea.
a) E-mail: kwak@chosun.ac.kr

DOI: 10.1587/transinf.E95.D.3117

$$u_{tik} = \frac{w_{tk}}{\sum_{j=1}^{c} \left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|}{\|\boldsymbol{x}_k - \boldsymbol{v}_j\|} \right)^{\frac{2}{m-1}}} \quad i = 1, 2, \ldots, c, \; k = 1, 2, \ldots, N \tag{3}$$

where u_{tik} represents the element of the membership matrix induced by the i-th cluster and k-th data in the t-th context. The cluster centers $\boldsymbol{v}_i$ are calculated in the form

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{N} u_{tik}^m \boldsymbol{x}_k}{\sum_{k=1}^{N} u_{tik}^m} \tag{4}$$

2.2 Linguistic Models

For the design of the LM, we consider the contexts to be described by triangular membership functions being distributed in the output space with the 1/2 overlap occurring between two successive fuzzy sets. The linguistic contexts are automatically generated by histogram, probability density function, and conditional density function in order [10]. We denote those fuzzy sets by W_1, W_2, …, W_p as linguistic contexts. Each context generates a number of induced clusters whose activation levels are afterwards summed up as shown in Fig. 1. Assuming the triangular form of the contexts, triangular fuzzy number E is expressed as

$$E = W_1 \otimes \xi_1 \oplus W_2 \otimes \xi_2 \oplus \cdots \oplus .W_n \otimes \xi_n \tag{5}$$

We denote the algebraic operations by $\otimes$ and $\oplus$ to emphasize that the underlying computing operates on a collection of fuzzy numbers. The bias term is computed in a straightforward manner so that it eliminates a potential systematic error

$$w_0 = \frac{1}{N} \sum_{k=1}^{N} (e_k - y_k) \tag{6}$$

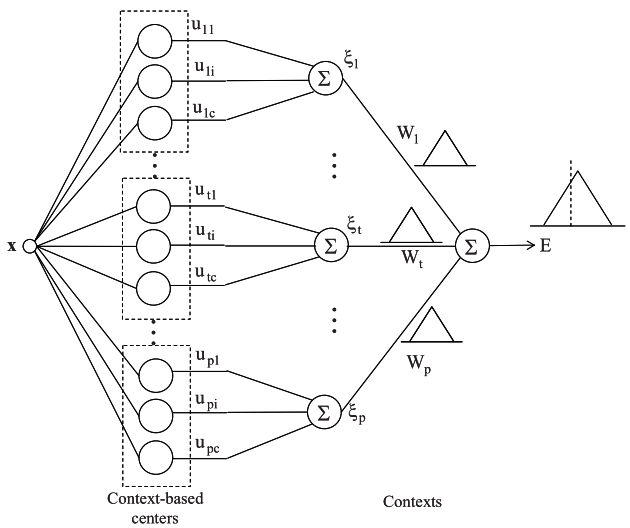

Fig. 1 The general architecture of the linguistic model.

where y_k denotes the actual output for given input $\boldsymbol{x}_k$. The resulting granular output E comes with the following equation

$$E = \sum_{t=1}^{p} \xi_t w_t + w_0 \tag{7}$$

3. Genetically Optimized Linguistic Models

3.1 Genetic Algorithm (GA)

GA encodes each point to be optimized into a binary bit string, and each point is concerned with a fitness value that is equal to the objective function computed at the point. In each generation, GA produces a new population using genetic operators such as crossover and mutation.

GA used in this paper includes encoding schemes, fitness evaluation, parent selection, crossover operator, and mutation operator. The fitness evaluation is to calculate the fitness value of each individual in the population after creating a generation. The fitness value of each individual is computed by the objective function for a maximization problem. The fitness function to be considered in this paper is as follows

$$f = \frac{1}{Q_{tRMSE} + Q_{cRMSE}} \tag{8}$$

$$Q_{tRMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} [e_k - y_k]^2}, \qquad Q_{cRMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} [e_{k'} - y_{k'}]^2} \tag{9}$$

Here we use the root mean square error (RMSE) to express the matching level of the granular output of the model. Q_{tRMSE} and Q_{cRMSE} are RMSE of training and checking data, respectively. For further details refer to [12]. The selection process determines which parents participate in producing offspring for the next generation. Crossover process is to apply to selected pairs of parents with a probability of a given crossover rate. Mutation process is to change the selected bit with a probability of a given mutation rate. Thus, it can prevent the population from converging at any local optima. Furthermore, we choose the elitism principle of always keeping a certain number of best members when each new population is generated.

3.2 Optimization Design

Based on the concepts as mentioned above, GA procedure is described as follows

[Step 1] Initialize a population with randomly generated individuals and set to crossover and mutation rate, bit number, and fitness function. And then evaluate the fitness value of each individual. GA simultaneously performs

parallel search through six populations with the number of different contexts.

[Step 2] Select two individuals from the population with probabilities proportional to fitness values in each population. The coding scheme is to arrange the number of cluster generated by each context and weighting exponent into a chromosome such that the representation preserves certain good properties after recombination specified by crossover and mutation operators.

[Step 3] Apply crossover and mutation with a probability of crossover and mutation rate, respectively.

[Step 4] Repeat Step 2 to Step 3 until a stopping criterion is met.

4. Experimental Results

This section is to demonstrate the performance of GA in the optimization design of the LM. For this, we apply the proposed method to coagulant dosing process in a water purification plant. We use the successive 346 samples among jar-test data for one year [11]. The input variable consists of four, including the turbidity of raw water, temperature, pH, and alkalinity. The output variable 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 remaining data sets are used for model validation.

In order to find the optimized parameters using GA, we first confine the search domain such as the number of cluster from 2 to 9 each context and weighting exponent from 1.5 to 3, respectively. Here, we need a new optimization strategy because the number of chromosome varies from the number of context. Thus we perform a parallel GA through six populations of LM with the number of different contexts from 3 to 8. In the design of GA-based LM, we encountered a data scarcity problem due to small data included in side linguistic context when $p = 9$. Thus, we determined $p = 8$ as the maximum number in this experiments. We used 8-bit binary coding for each variable. Each generation in GA implementation contains 30 individuals. Furthermore, we used a simple one-point crossover scheme with the crossover rate equal to 0.97 and uniform mutation with the mutation rate equal to 0.01. Figure 2 shows the best values of the objective function across 30 generations when the number of context varies from 3 to 8. Since we used elitism to keep the best two individuals at each generation, the best curve is monotonically increasing with respect to generation numbers. Here we finally obtained the best parameters ($p = 8$, $c = [9\ 8\ 2\ 7\ 6\ 5\ 5\ 5]$, $m = 1.714$).

Figure 3 shows the comparison between the desired and model output for both training and checking data, respectively. As shown in Fig. 3, it is obvious that the proposed GA-based LM has a good prediction performance. Figure 4 shows the interval prediction performance represented by lower bound, modal output, and upper bound. Table 1 lists the comparison results of RMSE of training and checking data, respectively. The RBFN used in Table 1 was designed by supervised adjustments of the center and shape of the receptive field functions. Here we used 45 receptive fields, 1000 epoch, and learning rate 0.01. The best model of TSK-LFM was obtained when the test error is minimal ($p = 5$, $c = 9$). As listed in Table 1, the experimental results obtained by the proposed method yielded a good performance in comparison to the previous works.

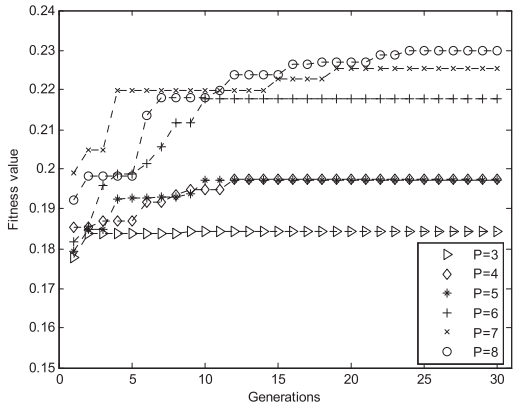

Fig. 2 Performance of GA by generation and context variation.

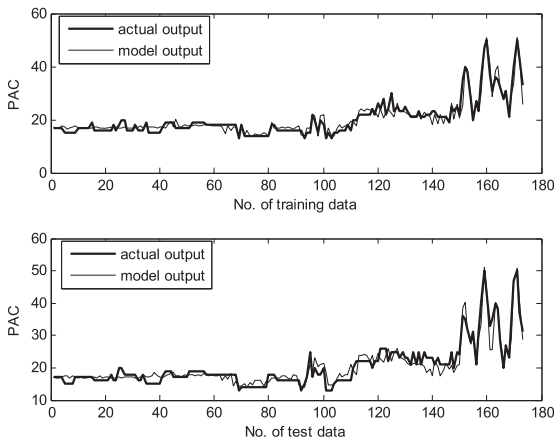

Fig. 3 Generalization and approximation capability.

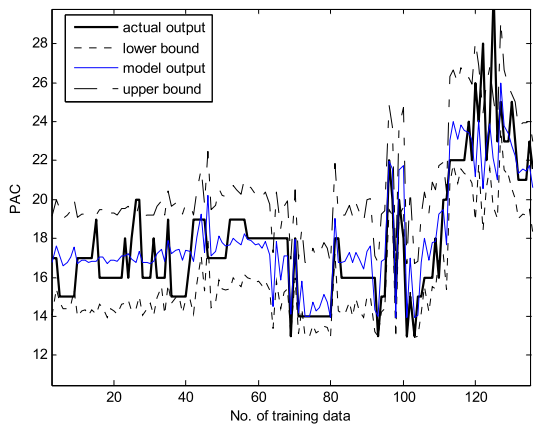

Fig. 4 Performance with interval prediction.

Table 1 Comparison results of RMSE (*: num. of hidden node).

Method	Num.of rules	Training data(RMSE)	Checking data(RMSE)
LR	-	3.508	3.578
MLP	45*	3.191	3.251
RBFN	45*	3.048	3.219
LM(p=8, c=6)[8]	48	2.549	2.820
LM(p=8, c=7)[8]	56	2.526	2.835
LM(p=8, c=8)[8]	64	2.427	2.800
TSK-LFM[11]	45	2.514	2.661
GA-based LM	47	2.060	2.290

5. Conclusions

We have enhanced the design methodology by genetically optimized linguistic model with the aid of fuzzy granulation. For this, we used the parallel GA that is derivative-free stochastic optimization methods based on the concepts of natural selection and evolutionary processes. Thus, we could optimize the number of context produced in the output space, the number of clusters obtained from each context, and weighting factor. To evaluate the performance of the proposed method, we applied it to the coagulant dosing process in a water purification plant. The experimental results revealed that the GA-based LM showed a good performance in comparison with the previous works.

Acknowledgments

This work was supported by research funds from Chosun University, 2009.

References

[1] W. Pedrycz and K.C. Kwak, "Linguistic models as a framework of user-centric system modeling," IEEE Trans. Syst. Man Cybern. A, vol.36, no.4, pp.727–745, 2006.
[2] E. Kim, M. Park, S. Kim, and M. Park, "A transformed input-domain approach to fuzzy modeling," IEEE Trans. Fuzzy Systems, vol.6, no.4, pp.596–604, 1998.
[3] M.L. Hadjuli and V. Wertz, "Takagi-Sugeno fuzzy modeling incorporating input variables selection," IEEE Trans. Fuzzy Systems, vol.10, no.6, pp.728–742, 2002.
[4] J. Abonyi, R. Babuska, and F. Szeifert, "Modified Gath-Geva fuzzy clustering for identification of Takagi-Sugeno fuzzy models," IEEE Trans. Syst. Man Cybern., vol.32, no.5, pp.612–621, 2002.
[5] S.S. Kim and K.C. Kwak, "Development of quantum-based adaptive neuro-fuzzy networks," IEEE Trans. Syst. Man Cybern. B, vol.40, no.1, pp.91–100, 2010.
[6] W. Pedrycz, "Conditional fuzzy C-means," Pattern Recognit. Lett., vol.17, pp.625–632, 1996.
[7] W. Pedrycz, "Conditional fuzzy clustering in the design of radial basis function neural networks," IEEE Trans. Neural Netw., vol.9, no.4, pp.601–612, 1998.
[8] W. Pedrycz and A.V. Vasilakos, "Linguistic models and linguistic modeling," IEEE Trans. Syst. Man Cybern., vol.29, no.6, pp.745–757, 1999.
[9] W. Pedrycz and K.C. Kwak, "Boosting of granular modeling," Fuzzy Sets and Systems, vol.157, pp.2934–2953, 2006.
[10] W. Pedrycz and K.C. Kwak, "The development of incremental models," IEEE Trans. Fuzzy Systems, vol.15, no.3, pp.507–518, 2007.
[11] K.C. Kwak and D.-H. Kim, "TSK-based linguistic fuzzy model with uncertain model output," IEICE Trans. Inf. & Syst., vol.E89-D, no.12, pp.2919–2923, Dec. 2006.
[12] K.C. Kwak and W. Pedrycz, "A design of genetically oriented linguistic model with the aid of fuzzy granulation," WCC2010 IEEE World Congress on Computational Intelligence, pp.52–57, Barcelona, Spain, 2010.