LETTER

A Design of Incremental Granular Model Using Context-Based Interval Type-2 Fuzzy C-Means Clustering Algorithm

Keun-Chang KWAK[†a)], ***Member***

SUMMARY In this paper, a method for designing of Incremental Granular Model (IGM) based on integration of Linear Regression (LR) and Linguistic Model (LM) with the aid of fuzzy granulation is proposed. Here, IGM is designed by the use of information granulation realized via Context-based Interval Type-2 Fuzzy C-Means (CIT2FCM) clustering. This clustering approach are used not only to estimate the cluster centers by preserving the homogeneity between the clustered patterns from linguistic contexts produced in the output space, but also deal with the uncertainty associated with fuzzification factor. Furthermore, IGM is developed by construction of a LR as a global model, refine it through the local fuzzy if-then rules that capture more localized nonlinearities of the system by LM. The experimental results on two examples reveal that the proposed method shows a good performance in comparison with the previous works.

key words: *incremental granular model, context-based type-2 fuzzy c-means clustering, linear regression, linguistic model*

1. Introduction

Fuzzy set provides a linguistically systematic calculus to deal with the imprecise and uncertain information and performs numerical computations by using linguistic labels specialized by membership functions with triangular or various shapes. Furthermore, fuzzy if-then rules are used as the important component of a Fuzzy Inference System (FIS) that can effectively model human expertise in a real-world application [1]. Over the past few decades, a considerable number of studies have been conducted on the effectiveness of such FIS with a structured knowledge representation in the form of fuzzy rules [2]. However, in order to solve the adaptability to deal with changing external environments and optimization of FIS, studies have focused on hybrid intelligent systems including FIS, neural networks, data clustering, and several optimization methods [3].

On the other hand, clustering techniques are used for knowledge representation and extraction in conjunction with a fuzzy modeling primarily to determine initial locations for fuzzy if-then rules from numerical input-output data points. The most representative fuzzy clustering frequently used in conjunction with fuzzy modeling is Fuzzy C-Means (FCM) clustering algorithm introduced by Bezdek [4]. Recently Rhee [5] proposed FCM clustering with Interval Type-2 (IT2) for uncertain fuzzy clustering. The detailed reasons of adopting IT2 is that the management of uncertainty by an IT2 fuzzy approach deals with the uncertainty in the cluster memberships by incorporating IT2 fuzzy sets and aids cluster prototypes to converge to a more desirable location than a type-1 fuzzy approach [4]. Several examples have been demonstrated the effectiveness of IT2 fuzzy approach methods by improving the uncertainty problem of type-1 fuzzy approach [4].

However, these clustering algorithms estimate cluster centers by only input variables without considering the association between input and output characteristics in the design of FIS. In contrast to these clustering, Context-based Fuzzy C-Means (CFCM) introduced by pedrycz [6]–[8] clustering estimates cluster centers preserving homogeneity of the clustered patterns in connection with their similarity in both the input and output variable based on contexts. However, this clustering involves the uncertainty of fuzzification factor that controls the amount of fuzziness.

Therefore, this paper is concerned with a design of Incremental Granular Models (IGM) as hybrid intelligent system based on Context-based Interval Type-2 Fuzzy C-Means (CIT2FCM) Clustering. Here IGM is followed the model proposed by Pedrycz and Kwak [9]. First, this model is constructed by Linear Regression (LR) model which can be treated as a preliminary design capturing the linear part of the entire data. Next, all modeling errors are compensated by fuzzy if-then rules that capture more localized nonlinearities of the system to be considered. The model output is expressed by upper and lower bounds with uncertainty of predicted values. This CIT2FCM clustering has unique characteristics preserving the homogeneity and covering on the uncertainty associated with fuzzification factor.

This paper is organized in the following fashion. In Sect. 2, the procedure steps of FCM, CFCM, and CIT2FCM clustering methods are described. The entire design process of IGM based on CIT2FCM clustering is presented in Sect. 3. The experimental results are performed and discussed in Sect. 4. Concluding comments are covered in Sect. 5.

2. FCM and CFCM Clustering Algorithms

2.1 Fuzzy C-Means (FCM) Clustering

Fuzzy C-Means (FCM) partitions a collection of input data points into c fuzzy groups and finds cluster centers in each group such that an objective function is minimized [5]. The membership matrix consists of elements with values be-

Manuscript received March 30, 2015.
Manuscript revised September 17, 2015.
Manuscript publicized October 20, 2015.

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

DOI: 10.1587/transinf.2015EDL8076

tween 0 and 1 to perform fuzzy partitioning. The summation of belongingness degrees for a data set is equal to unity as following expression

$$\mathbf{U} = \left\{ u_{ik} \in [0,1] \,\middle|\, \sum_{i=1}^{c} u_{ik} = 1 \;\; \forall k \;\; and \;\; \forall i \right\} \tag{1}$$

The objective function for FCM clustering is expressed as Eq. (2)

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2} \tag{2}$$

where d_{ik} is the Euclidean distance between i'th cluster center and k'th data point. FCM clustering estimates the cluster centers and the membership matrix by the following steps.

[**Step 1**] Initialize the membership matrix with random values between 0 and 1.

[**Step 2**] Compute c fuzzy cluster centers using Eq. (3). Here, fuzzification factor is generally used as fixed value.

$$\mathbf{c}_i = \frac{\sum_{k=1}^{N} u_{ik}^{m} \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^{m}} \tag{3}$$

[**Step 3**] Compute the objective function according to Eq. (2). Stop if it is below a certain tolerance value.

[**Step 4**] Compute a new membership matrix using Eq. (4). Go to [Step 2].

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{c}_i\|}{\|\mathbf{x}_k - \mathbf{c}_j\|} \right)^{\frac{2}{m-1}}} \tag{4}$$

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

The CFCM clustering is an effective algorithm to determine the cluster centers preserving homogeneity on the basis of fuzzy granulation [6]. In contrast to FCM clustering, 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 performed by some predefined fuzzy sets of contexts. Let us introduce a family of the partition matrices induced by the t-th context as follows

$$\mathbf{U} = \left\{ u_{ik} \in [0,1] \,\middle|\, \sum_{i=1}^{c} u_{ik} = w_{tk} \; \forall k \;\; and \;\; \forall i \right\} \tag{5}$$

where w_{tk} denotes a membership value of the k-th data point included by the t-th context. The underlying objective function is equal to Eq. (2). Here it performs the separate clustering tasks implied by each context. The minimization of objective function is achieved by iteratively updating the values of the membership matrix and the prototypes. The membership matrix is computed as follows

$$u_{tik} = \frac{w_{tk}}{\sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{c}_i\|}{\|\mathbf{x}_k - \mathbf{c}_j\|} \right)^{\frac{2}{m-1}}} \quad i = 1, 2, \ldots, c, \;\; k = 1, 2, \ldots, N \tag{6}$$

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.

3. Context-Based Interval Type-2 Fuzzy C-Means (CIT2FCM) Clustering

In this section, the CIT2FCM clustering is proposed. The estimation method of cluster center is similar to the procedure of CFCM clustering except for adding uncertainty of fuzzification factor m on the basis of Interval Type-2 (IT2) fuzzy approach. The CIT2FCM clustering is performed by the following steps.

[Step 1] Select the number of context and cluster per context, respectively. For simplicity, it assumes that the number of cluster per context is equal. Initialize the membership matrix with random value between 0 and 1.

[Step 2] Generate linguistic contexts with triangular membership function using statistical probabilistic distribution in the output space [10]. Each context is generated by an overlap of 1/2 point between successive fuzzy sets.

[Step 3] Compute upper and lower partition matrices as Eq. (7). The fuzzifiation factor m is replaced by m_1 and m_2 which represent different fuzzifier value.

$$\overline{u}_{ik} = \max\left(f_k \Big/ \sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{c}_i\|}{\|\mathbf{x}_k - \mathbf{c}_j\|} \right)^{\frac{2}{m_1-1}}, f_k \Big/ \sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{c}_i\|}{\|\mathbf{x}_k - \mathbf{c}_j\|} \right)^{\frac{2}{m_2-1}} \right)$$
$$\underline{u}_{ik} = \min\left(f_k \Big/ \sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{c}_i\|}{\|\mathbf{x}_k - \mathbf{c}_j\|} \right)^{\frac{2}{m_1-1}}, f_k \Big/ \sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{c}_i\|}{\|\mathbf{x}_k - \mathbf{c}_j\|} \right)^{\frac{2}{m_2-1}} \right) \tag{7}$$

[Step 4] Update the cluster center. The individual values of the left and right cluster boundaries in each dimension can be computed by sorting the order of patterns in particular dimension and then applying Karnik-Mendel (KM) iterative procedure [1]. Here KM algorithm is used to update the interval set of cluster centers. The new cluster center is computed by a defuzzification method as follows

$$\mathbf{c} = \frac{\mathbf{c}_L + \mathbf{c}_R}{2} \tag{8}$$

[Step 5] Compute distance measure between the updated clusters and the previous ones. Stop if the improvement over previous iteration is below a certain threshold.

[Step 6] Compute a new membership function based on average of lower and upper bound as type-reduce step Eq. (9). Go to Step 3.

$$u_{ik} = \frac{\overline{u}_{ik} + \underline{u}_{ik}}{2} \tag{9}$$

4. Incremental Granular Model (IGM) Using CIT2FCM Clustering

In this section, the design of IGM based on the proposed CIT2FCM clustering algorithm is presented. The IGM deals with localized nonlinearities of complex system so that the modeling error can be compensated. After performing the design of Linear Regression (LR) as the first global model, we refined it through a series of local fuzzy if-then rules in order to capture the remaining localized characteristics [9]. Figure 1 shows the main design process of the IGM. Firstly, we determine the granularity information to be used in the development of the model such as the number of contexts and the number of cluster estimated by each context. The design procedure of IGM is as follows

[Step 1] Design LR with input-output data points. LR fits a nonlinear relationship between the value of $\mathbf{x}_k$ and the corresponding output. On the basis of the original data set, a collection of input-error pairs, $(\mathbf{x}_k, \mathrm{e}_k)$ is obtained. Here the error is obtained by LR modeling.

[Step 2] Generate linguistic contexts in the error space $(\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_p)$ of the LR. These contexts are automatically generated by histogram, probability density function, and conditional density function in order [10]. The linguistic contexts is characterized by fuzzy sets $(\mathrm{W}_1, \mathrm{W}_2, \ldots, \mathrm{W}_p)$.

[Step 3] Perform CFCM clustering in the input-output space from the linguistic contexts produced in the error space.

[Step 4] Calculate the activation levels of the clusters estimated by the corresponding contexts and their overall aggregation through weighting by the contexts in output.

[Step 5] The IGM's output is combined with the output of the linear part. The result is obtained by $\mathrm{Y} = \mathrm{z} \oplus \mathrm{E}$. Here triangular fuzzy number E is expressed as

$$\mathrm{E} = \mathrm{W}_1 \otimes \xi_1 \oplus \mathrm{W}_2 \otimes \xi_2 \oplus \ldots \mathrm{W}_n \otimes \xi_n \tag{10}$$

It denotes the algebraic operations by $\otimes$ and $\oplus$ to emphasize that the underlying computing operates on a collection of fuzzy numbers.

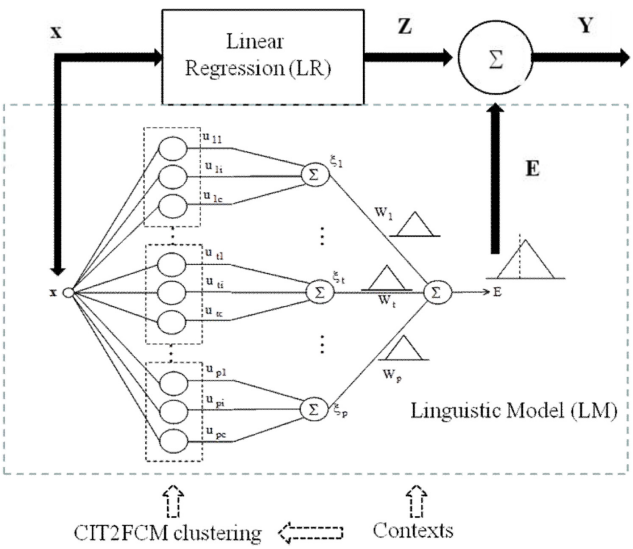

Fig. 1 Main design process of IGM based on LR and LM.

5. Experimental Results

In what follows, the experiment is performed with two selected datasets originating from the Machine Learning repository: automobile MPG (miles per gallon) data and Boston housing data. First, the automobile's MPG predication problem is a typical nonlinear regression problem, in which several attributes are used to predict another continuous attribute. This data set includes six attributes consisting displacement, horsepower, weight, number of cylinder, model year, and acceleration. Second, the Boston housing data concerns prices of real estate in the Boston area. The MEDV (median value of the price of the house) depends on 12 continuous attributes except for 1 binary attribute. All experiments were performed in the 10 fold cross-validation mode with a typical 60–40% split between the training and testing data subsets in each experiment those were randomly selected from the entire dataset. Tables 1 and 2 summarize the prediction performance of the proposed IGM for the automobile MPG data and Boston housing checking data sets. In the design of MLP (Multilayer Perceptron), we used 30 and 60 nodes as five times of the input's number to determine the number of hidden node, respectively. For illustrative purposes, the experiment is compared with

Table 1 Comparison results of RMSE for automobile MPG data set (∗: num. of hidden node).

Method	Num.of rules	Training data(RMSE)	Checking data(RMSE)
LR	-	3.383	3.472
MLP	30*	3.309	3.104
RBFN [7]	36*	2.338	3.177
LM(p=6, c=6)[8]	36	2.802	3.319
IM(p=6, c=6) [9]	36	2.390	3.060
Proposed method	36	2.101	2.989

Table 2 Comparison results of RMSE for Boston housing data set (∗: num. of hidden node).

Method	Num.of rules	Training data(RMSE)	Checking data(RMSE)
LR	-	4.535	5.043
MLP	60*	4.883	5.270
RBFN [7]	36*	3.386	4.490
LM(p=6, c=6)[8]	36	4.563	5.568
IM(p=6, c=6) [9]	36	3.279	4.298
Proposed method	36	3.213	3.865

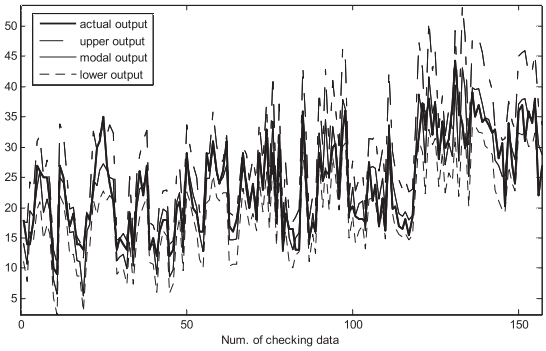

Fig. 2 Prediction performance for automobile MPG data.

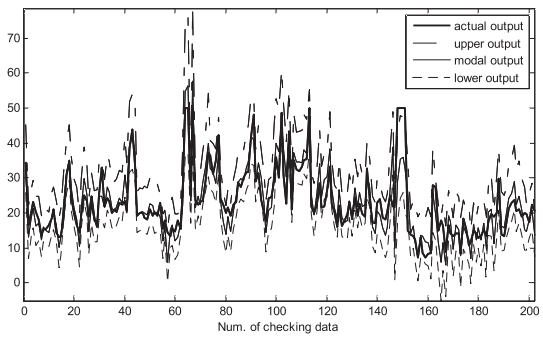

Fig. 3 Prediction performance for Boston housing data.

$p = c = 6$ in which case we have achieved a sound balance between the granularity of information formed in the output and input space in the design of RBFN (Radial Basis Function Network), LM (Linguistic Model), and IM (Incremental Model). While the fuzzification factor associated with these models is used as $m = 2$, the uncertainty range of this factor is used as the values between $m = 1.1$ and $m = 5.0$ in the design of IGM. The number of parameters of the proposed model additionally includes the set of linear parameters for LR and two fuzzification factors in comparison to that of the IM itself in Tables 1 and 2. As listed in Tables 1 and 2, the experimental results revealed that the proposed IGM showed good performance in comparison to the previous works. Figures 2 and 3 show the prediction performance for automobile MPG data and Boston housing data, respectively.

6. Conclusions

The design method of IGM using context-based IT2FCM clustering algorithm with the aid of fuzzy granulation has been proposed. For this, the knowledge extraction of fuzzy if-then rules based on new clustering technique is developed. Thus, it could estimate the effective clusters from linguistic context produced in the output space, and cover the uncertainty of fuzzification factor. Furthermore, the effectiveness and superiority of the proposed IGM has been demonstrated through the experiments. For further research, the uncertainty of both the contexts and fuzzifcation factor with IT2 approach will be considered.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A2012127)

References

[1] J.M. Mendel, Rule-Based Fuzzy Logic Systems: Introduction and new directions, Prentice Hall, 2001.
[2] W. Pedrycz and F. Gomide, Fuzzy systems engineering: Toward human-centric computing, Wiley-Interscience, 2007.
[3] 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.
[4] T.C. Havens, J.C. Bezdek, C. Leckie, L.O. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," IEEE Trans. Fuzzy Syst., vol.20, no.6, pp.1130–1146, 2012.
[5] F.C.H. Rhee, "Uncertain fuzzy clustering: Insights and recommendations," IEEE Comput. Intell. Mag., vol.2, no.1, pp.44–56, 2007
[6] W. Pedrycz, "Conditional fuzzy c-means," Pattern Recognit. Lett., vol.17, no.6, pp.625–631, 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. B, Cybern., vol.29, no.6, pp.745–757, 1999.
[9] W. Pedrycz and K.-C. Kwak, "The development of incremental models," IEEE Trans. Fuzzy Syst., vol.15, no.3, pp.507–518, 2007.
[10] K.-C. Kwak, "A design of genetically optimized linguistic models," IEICE Trans. Inf. & Syst., vol.E95-D, no.12, pp.3117–3120, Dec. 2012.