New Cluster Validation with Input-Output Causality for Context-Based Gk Fuzzy Clustering

Keun-Chang Kwak
Dept. of Control and Instrumentation Engineering
Chosun University, 375 Seosuk-Dong
Gwangju, Korea

***Abstract*—In this paper, a cluster validity concept from an unsupervised to a supervised manner is presented. Most cluster validity criterions were established in an unsupervised manner, although many clustering methods performed in supervised and semi-supervised environments that used context information and performance results of the model. Context-based clustering methods can divide the input spaces using context-clustering information that generates an output space through an input-output causality. Furthermore, these methods generate and use the context membership function and partition matrix information. Additionally, supervised clustering learning can obtain superior performance results for clustering, such as in classification accuracy, and prediction error. A cluster validity concept that deals with the characteristics of cluster validities and performance results in a supervised manner is considered. To show the extended possibilities of the proposed concept, it demonstrates three simulations and results in a supervised manner and analyzes the characteristics.**

Keywords—Cluster Validation; Fuzzy clustering; Gustafson-Kessel clustering; Fuzzy covariance; Context based clustering; Input-output causality

I. INTRODUCTION

Intelligent systems that optimize using learning schemes without strict mathematical constraints are a very useful approach to construct modeling in complex environments[3][4]. A clustering approach [1-4][8][11-12] is one of the generic methods for determining the structure and parameters of an initial intelligent system. Once the initial structure and parameters are determined, the system can use various learning mechanisms for optimization. However, the method by which a system performs clustering is an interesting issue in itself [2][8][11]. Pattern recognition is one of the most interesting applications of intelligent systems, especially clustering method is useful approach of them. Clustering is a process in which groups of objects with high similarity, as compared to the members of other groups, are collected as clusters. The concept is highly similar to pattern classification or recognition. Generally, clustering methods perform well in an unsupervised manner to divide input spaces and extract useful information from data sets. This helps to construct intelligent systems [5]. [10] [11] such as neural networks and fuzzy systems that divide an input space into several local spaces, in turn allowing for ease of interpretation. In a clustering algorithm, selecting an appropriate number of clusters is a critical problem. A simple method to identify the proper number of clusters is to select the result that provides best performance. Another approach is to apply a cluster validation [6][7][14][17-19] using cluster parameters after the clustering algorithm is terminated. This method only needs clustering results and does not need any additional information such as performance results. Because of this property, many cluster validations have been proposed by researchers in the field of pattern recognition and widely used. In prior work, a semi-supervised clustering method [9][16] and a supervised clustering approach [10-12] have made use of output information. Additionally, context-based clustering methods [11-13] have used a context membership function, which was generated by a context term as output, and contained an input-output causality. This characteristic provides more quantitative information to perform the clustering. Conventional cluster validity methods induce a fixed value on the cluster validity. The cluster validity, including input-output causality such as the cluster validity of the output, has not yet been studied in a supervised manner. Any proposed cluster validity concept can obtain more flexible criterions when it uses the input-output causality or context information such as a context membership function. This means that when the cluster validity uses more than one cluster validity result, it can attempt to induce more flexible values for the cluster validity to adapt the input-output causality, or it can introduce a performance-dependent criterion. To achieve this, it proposes two combined cluster validity concepts that use the classification accuracy of a classification problem and a cluster validity of the context membership function. Among the cluster validity values, the proposed concept can choose a relative ratio to adjust the importance between the cluster validity of the input-output causalities, such as input/output CV, and performance accuracy. The proposed concept extends the cluster validity criterion to the supervised manner in the context-based clustering. The rest of this paper proceeds as follows. Section 2 describes related research, including clustering methods and cluster validity methods. In section 3, a new cluster validity concept that can be applied in a supervised manner is proposed. Section 4 then presents the results of experimental comparisons between our new cluster validation and previous approaches. In Section 5, the conclusion with a summary is given.

II. THE RELATED WORKS

In this section, it briefly describes existing clustering methods and cluster validity methods. These methods based on new cluster validity. A context-based clustering method is introduced after our explanations of general clustering. Then,

three cluster validity criterions will be used to briefly explicate cluster validities.

A. Unsupervised clustering methods

FCM [3][4] is a representative fuzzy clustering method that uses a partition matrix of the membership function between cluster centers and data sets. It measures similarity as follows:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\left(\frac{2}{m-1}\right)}} \quad (1)$$

where d_{ik} is the distance between a center c_i and kth data z_k. An m is a fuzzifier and the similarity μ_{ik} is the element of the partition matrix of the membership function. In the process, center c_i is updated by the similarity until a termination criterion is satisfied, as follows:

$$c_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m x_k}{\sum_{k=1}^{N}(\mu_{ik})^m} \quad (2)$$

Most cluster validity methods primarily use the partition matrix to evaluate the cluster validity.

Gustafson-Kessel (GK) [1][2] clustering uses the fuzzy covariance matrix to adapt elliptical shape cluster sets that use fuzzy covariance information, as shown in following equation:

$$F_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m (x_k - c_i)(x_k - c_i)^T}{\sum_{k=1}^{N}(\mu_{ik})^m} \quad (3)$$

The matrix A_i is combined by equation (4),

$$A_i = [\rho_i det(F_i)]^{1/n} F_i^{-1} \quad (4)$$

where ρ_i is a predefined constant to set to one. Then, the distance between center c_i and data x_k are measured by the following equation:

$$d_{ik}{}^2 = \left(x_k - c_i^{(l)}\right)^T A_i \left(x_k - c_i^{(l)}\right) \quad (5)$$

An updated GK cluster center is calculated as a weighted average by equation (2).

B. Supervised clustering methods

Context-based clustering [11] in a supervised manner uses a context membership function that regards input and output data as causally connected. When a context term, such as output space, can be grouped, connected input spaces are also meaningfully clustered. In the context term, the brief concept of context clusters is shown in Fig. 1. Different shapes are shown because of differences in measurement between simple Euclidean and fuzzy covariance metrics.

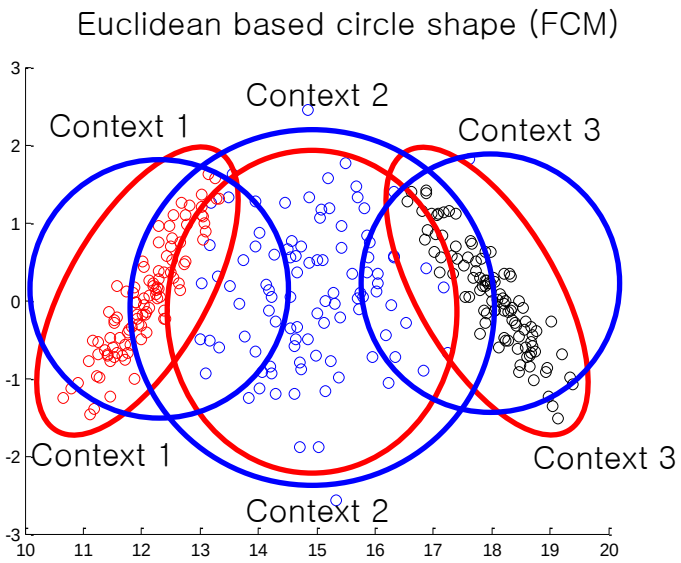

Fig. 1. A concept of context based clustering with FCM and GK

In the unsupervised manner, general similarity is calculated by equation (1). However, a similarity measure of the context clustering in the supervised manner is calculated by equation (6), adding context variable f_k which is induced by data x_k and context membership functions, as shown in Fig. 2.

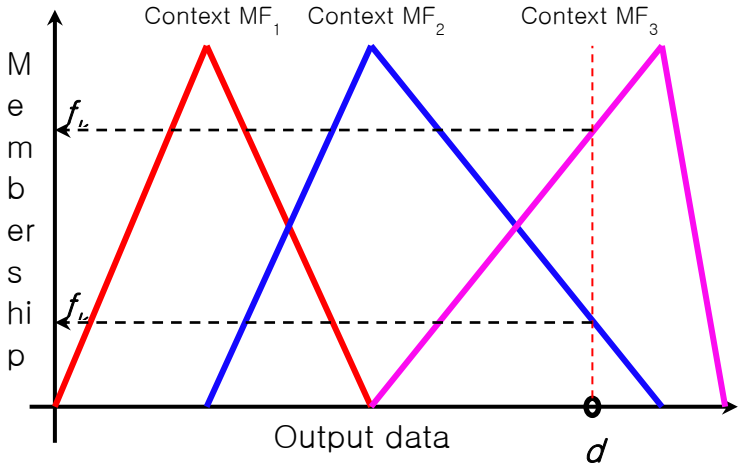

Fig. 2. The concept of context membership function

$$\mu_{ik} = \frac{f_k}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\left(\frac{2}{m-1}\right)}} \quad (6)$$

As shown Fig. 1, the f_k is induced by the context membership function when kth data is obtained by context membership functions two and three. Then, the equation (6) contains context information using f_k that assumes influencing input-output causality in the supervised manner.

C. Cluster validity

Cluster validity (CV) [6][7][14][18][19] is used to find the optimal number of clusters in a given data set. Bezdek proposed two CVs: the Partition Coefficient (V_{PC}), which minimizes an index value, and Partition Entropy (V_{PE}), which maximizes an index using a partition matrix as follows [6]:

$$V_{PC} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}{\mu_{ij}}^2}{n} \tag{7}$$

$$V_{PE} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}log_a\left(\mu_{ij}\right) \tag{8}$$

Xie and Beni [19] also proposed a CV index (VXB) that utilizes compactness and separation to find a minimized validity index, as follows:

$$V_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}{\mu_{ij}}^2\left\|x_j - c_j\right\|^2}{n\left(\min\limits_{i\neq k}\|c_i - c_k\|^2\right)} \tag{9}$$

Kim [6] proposed a CV index (VK) for GK clustering that also finds a minimized validity index, as follows:

$$V_K = \frac{2}{c(c-1)}\sum_{p\neq q}^{c}\sum_{j=1}^{n}\left[c\left[\mu_{\widetilde{F_p}}(x_j)\cap\mu_{\widetilde{F_q}}(x_j)\right]h(x_j)\right] \tag{10}$$

Although there are many interesting extensions to the concept, a full explanation is not our present concern; thus, it limits the discussion to our extension of current CVs in a supervised manner using input-output causality.

III. The Proposed Cluster Validity Method

The proposed cluster validity (CV) concept, which it calls context-based cluster validity (CCV), uses more than two CV considerations, such as a CV of the input space clustering and performance results, or a CV of the context clustering. This means that it extends the conventional CV concept in the unsupervised manner to a supervised CV concept. In the clustering process, it assumes that the output information of the data is already known because clustering based on supervised learning uses the output data, as recognized by the context term.

Throughout the causality, the output is causally correlated with the input. To construct the input clusters, context-based clustering serves advanced information of the causality using f_k that includes a causality degree of input and output clusters, as shown in Fig. 3. There are two criterions of the CV that exist in the model as an input and an output side, respectively.

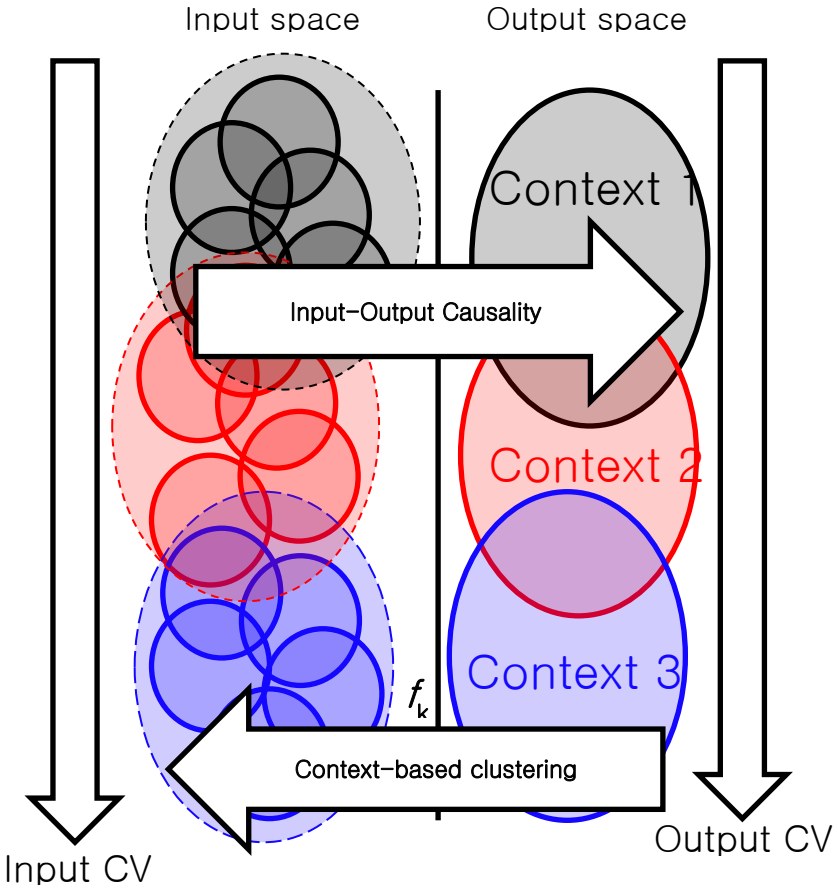

Fig. 3. The concept of input-output causality and context based clustering

The two types of context information are presented. The first is the accuracy (error) of the classification problems. The second is the CV of the partition matrix of the context membership in equation (12). In the classification problems, the context-based clustering method often does not obtain a context membership degree between zero and one. It only includes zero or one. Therefore, it cannot directly obtain the CV of the context membership function and then replace a classification error for adapting the causality. However, the classification error can be estimated easily by comparing the clustering results and the output data, such as class labels in the supervised manner. In case of very small values less than one, it amplifies the error to affect the CV result, with amplification ratio manually decided by minimum error value. This amplification helps to ensure an observed change in the CV curves. Eq. (11-1) contends that an induced new CV includes the CV of the input spaces and the classification error results in the context term. This CV concept influences the new CV result with the error. Despite getting a good input CV result, the proposed concept can have a bad CV value when classification error increases on the context term. In addition, Eq. (11-2) is the form of applying influence parameter α. It can influence an effect ratio of the context term such as error.

$$Proposal = CV\ of\ input \times (1 - error) \times (Amplification) \quad (11\text{-}1)$$

$$new\ CV = \alpha \times (CV\ of\ input) + (1 - \alpha)(proposal) \quad (11\text{-}2)$$

$$ew\ CV = \alpha \times (CV\ of\ input) + (1 - \alpha)(CV\ of\ context) \quad (12)$$

In Eq. (12), a new cluster validity concept that uses the CV of the context term and adjusts the relative ratio using the variable α is proposed. The parameter α can adjust the influence ratio of the input-output relativity emphasis. Conventional CVs generally calculate a criterion to induce a value that has no possibility of adjustment. In this paper, the variable α is important as it allows us to adjust the influence of the context information. It extends the CV concept from a fixed value of the CV to a choice preference in the scope of the input-output relativity emphasis. When the output data have continuous values and do not have a label index, generating the CV of the context membership function easily allows for the application of the causality. In this case, the proposed CV concept can apply an extended CV evaluation using the input and output CV. In the context-based clustering during the supervised learning, the clustering algorithm generally optimizes the input clusters using an advanced similarity metric with input-output causality. Then, the cluster validity also needs to extend the validity criterions at that environment. It specifies that the first characteristic is input-output causality in supervised settings. The input characteristic is already in existence as the CV. When the context-based clustering algorithm cannot obtain the context membership degree, such as in classification problems that do or do not only belong to the class, it assumes that classification error can replace the context membership function to represent the input-output causality. To apply the context CV, the classification accuracy is used to estimate the context CV of the classification problem. However, when it can obtain the context CV, the proposed concept easily adapts the criterion through an Eq. (12) such that a regression problem is used by the context membership degree, alongside other information to influence the final result.

IV. Experimental Results

In this Section, it used two computer simulations to show the characteristics of the proposed concept. The simulations using MATLAB 12, which was run on a Windows 7 machine with an i7 2.80 GHz CPU and 16 GB of DDR3 RAM is performed. The three simulation data sets, including two synthetic classification problems and one real data set are used. The two synthetic data sets were generated by a random selection method that intentionally forced shapes to obtain the elliptical geometric structure. The outputs were composed of three and five class labels. The real data set was downloaded from the UCI machine learning repository. This data set has 506 instances and fourteen attribute numbers, including an output that comprises the median value of owner-occupied homes in $1000. Here it used two input attributes: the weighted distance to five Boston employment centers, and the lower status of the population. The synthetic data distribution is shown in Fig. 4. It has five groups with various shapes, distributions, and densities. The three class problem is also from the same data set where two central classes are merged into a new class and two-sided small classes are also merged into a new class.

A. Cluster validity Cluster validity in classification problems

The index values of five and three (5, 3 classes) to represent the cluster validity of the input space and the classification error between the inferred cluster label and the real output label are used.

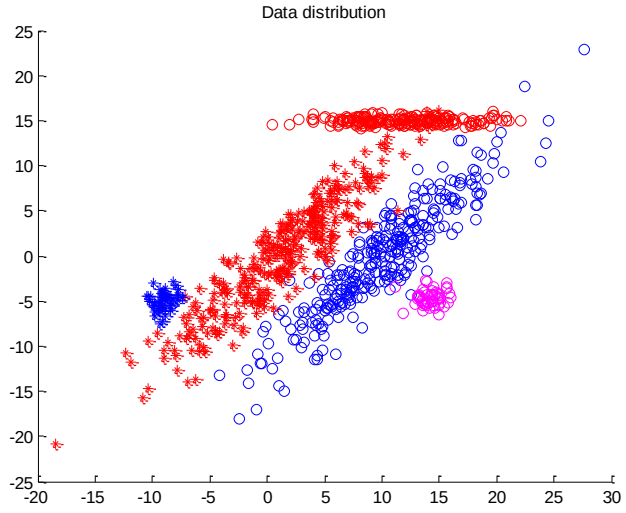

Fig. 4. Synthetic data distribution

To compare the change of the CV, all performance and CV results are normalized in Fig. 5 when the FCM algorithm is performed. The thick black line is a classification result that increases the classification performance when the number of clusters is increased. The thin red line is the cluster validity result of [19]. The dotted red line is the result of Eq. (11-1). The thick red line is a result of Eq. (11-2), which applies the input CV results and classification result with an influence parameter α of 0.5. The blue lines are similar to the CV of the [6]. Regarding the blue lines, the CV of the input and applied CV is a different curve. This means that if it knows the classification error then it can change the number of the clusters to fit the performance.

Figs. 5 and 6 show the CV results when FCM and GK clustering are performed. The cluster number scope is two to fifteen. In the three class problem, the Vk and our proposed concept are more different when the cluster number is increased. It is also possible to see the black line of the classification accuracy that influenced the proposed CV curve. In the five class problem, the cluster number is started from five to twenty. Figs. 7 and 8 show the CV results when FCM and GK clustering are performed.

B. Cluster validity in a regression problem

The CV results of the Boston housing regression [15] problem at the CFCM are shown in Fig. 9. The thick blue line is an input CV and the other lines are influenced by a CV of the context term as output and the influence parameter α in equation (12). The figure shows different results when influence parameter α is changed. As shown in Fig. 9, when the influence parameter α is already 0.5, a criterion value of the proposed concept is less than the input CV value. This means that the final determination including the CCV can change the optimal cluster number.

As illustrated in Fig. 10, it shows the result of the GK clustering when the influence parameter α is changed. It seems to have little effect compared with the FCM.

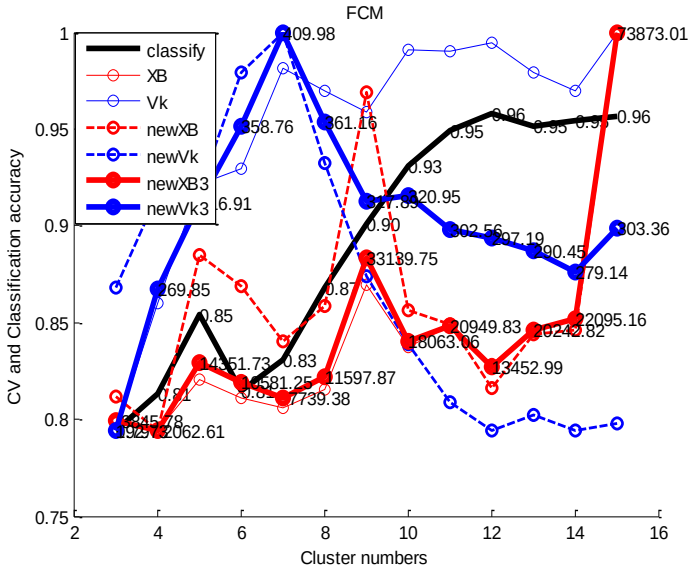

Fig. 5. Cluster validity result on FCM

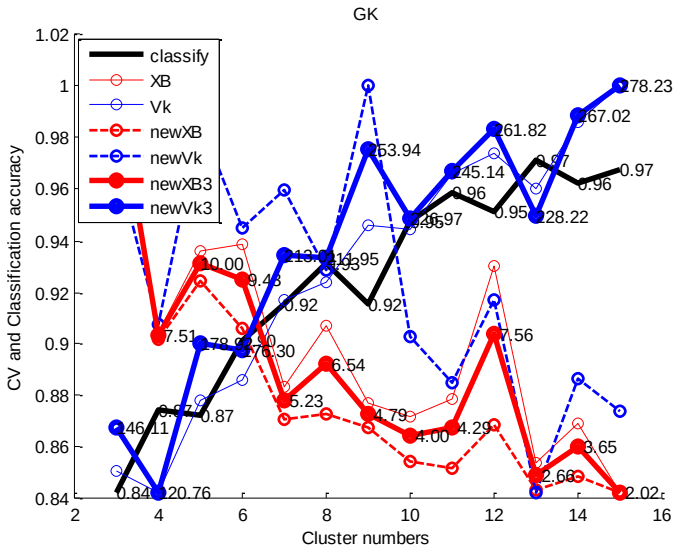

Fig. 6. Cluster validity results on GK in the three class problem

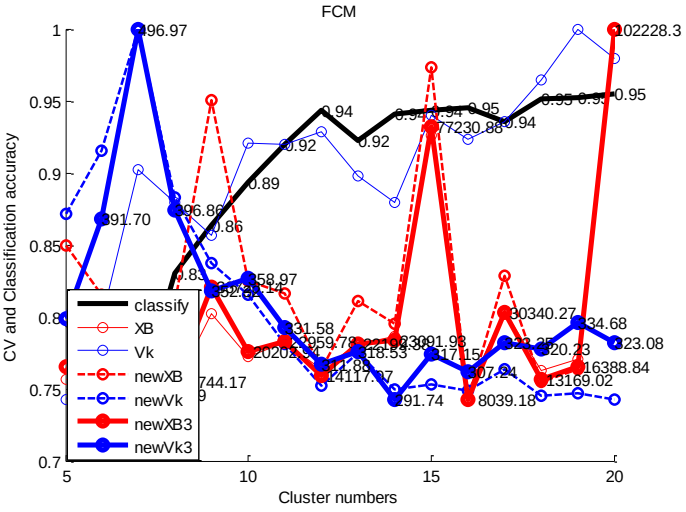

Fig. 7. Cluster validity result on FCM in the five class problem

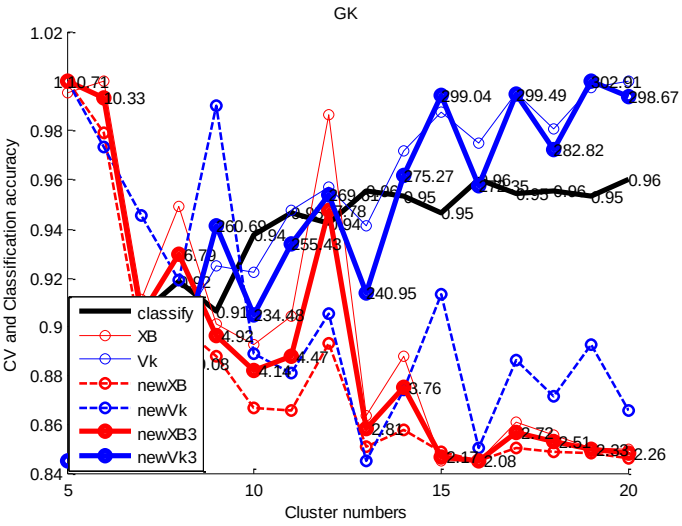

Fig. 8. Cluster validity result on GK in the five class problem

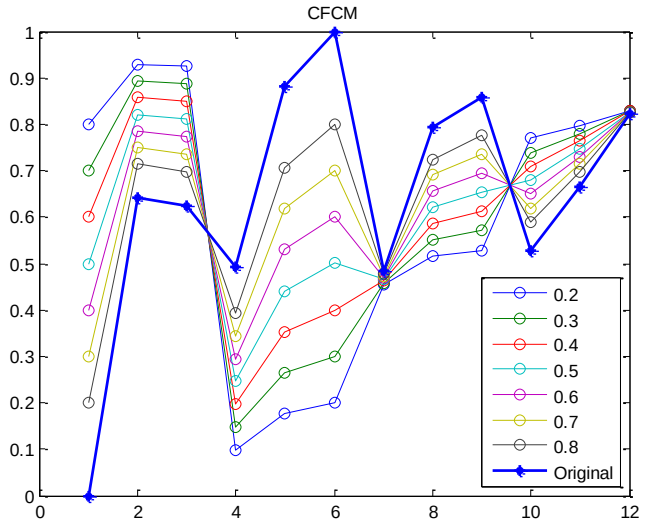

Fig. 9. Cluster validity result on FCM in a regression problem

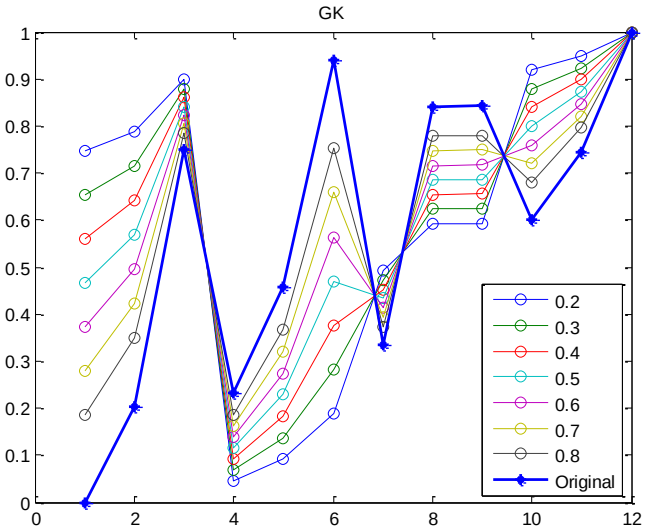

Fig. 10. Cluster validity result on GK in a regression problem

TABLE I. COMPARISON RESULTS OF CV

Case	Context number	Cluster number in a context	Cluster number	Input CV	Proposed CV
1	2	2	4	0	0.4
2	2	3	6	0.6425	0.7855
3	2	4	8	0.6235	0.7741
4	3	2	6	0.4921	**0.2952**
5	3	3	9	0.8825	0.5295
6	3	4	12	1.00	0.6001
7	4	2	8	**0.4849**	0.4698
8	4	3	12	0.7943	0.6554
9	4	4	16	0.8574	0.6933
10	5	2	10	0.5286	0.6492
11	5	3	15	0.6641	0.7304
12	5	4	20	0.8242	0.8266

Comparison of the values in Table 1 indicates that the best optimal cluster number is eight when only the input CV is used. However, in our concept, the best optimal cluster number is six at three context clusters. It has two cases of six clusters with different CV values at cases two and four.

As indicated by the CV results, it attempts to show the difference between conventional CV approaches and our proposed concept. Our approach has two advanced characteristics. First, it extends the cluster validity concept from the unsupervised to the supervised setting. In addition, introducing influence parameter α provides a more varied range of possible extensions.

V. CONCLUSIONS

In this paper, a new cluster validation method for context-based clustering in a supervised manner has developed. By adding more information to the context term, the cluster validation concept extends the possible application from unsupervised to supervised settings. Applying an input-output causality and an influence parameter provide wider choice in the cluster validity. This approach easily adapts to the context-based clustering. Conventional cluster validity values tend to have fixed values or constants and do not consider the input-output causality. Our proposed cluster validity extends this constancy to offer greater flexibility by using various elements and adjustments, such as α. Instead of constancy in the unsupervised settings, the proposed concept has sufficient scope to determine the most suitable number of clusters. In the instruction of an intelligent system using clustering, our approach can provide more marginal choice to determine the best overall parameters. Context-based clustering can adapt various context membership functions to improve performance. Thus, applying various membership functions in context terms and, later, analyzing the results of cluster validity will be very interesting opportunities for further research. Future work should also include applying the semi-supervised clustering and related works.

ACKNOWLEDGEMENTS

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] I. Gath, A. B. Geva, "Unsupervised optimal fuzzy clustering", *IEEE Trans on Pattern Analysis and Machine Intelligence* Vol. 11, No. 7, pp. 778-780, 1989.

[2] D. E. Gustafson, W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix", *IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, Vol. 17, pp. 761-766, 1978.

[3] S. Haykin, Neural Networks: A Comprehensive Foundation 2nd. Prentice Hall, 1999.

[4] 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.

[5] S. S. Kim, H. J. Choi, K. C. Kwak, "Knowledge extraction and representation using quantum mechanics and intelligent models", *Expert System with Applications*, Vol. 39, No. 3, pp. 3572-3581, 2012.

[6] Y. I. Kim, D. W. Kim, D. H. Lee, K. H. Lee, "A cluster validation index for GK cluster analysis based on relative degree of sharing", *Information Sciences*, Vol. 168, No. 4, pp. 225-242, 2004.

[7] S. H. Kwon, "Cluster validity index for fuzzy clustering", *Electronics Letters*, Vol. 34, No. 22, pp. 2176-2177, 2002.

[8] R. Krishnapuram, J. Kim, "A Note on the Gustafson-Kessel and Adaptive Fuzzy Clustering Algorithms", *IEEE Trans on. Fuzzy Systems* , Vol. 7, No. 4, pp. 453-461, 1999.

[9] M. H. C. Law, A. Topchy, A. K. Jain, "Clustering with Soft Group Constraints. Structural, Syntactic, and Statistical Pattern Recognition", *Lecture Notes in Computer Science*, Vol. 3138, pp. 662-670, 2004.

[10] W. Lu, W. Pedrycz, X. Liu, J. Yang, P. Li, "The modeling of time series based on fuzzy information granules", *Expert Systems with Applications*, Vol. 41, No. 8, 3799-3808, 2014.

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

[12] W. Pedrycz, "Conditional fuzzy clustering in the design of radial basis function neural networks", *IEEE Trans. on Neural Networks*, Vol. 9, No. 4, pp.745-757, 1999.

[13] W. Pedrycz, K. C. Kwak, "Linguistic models as a framework of user-centric system modeling", *IEEE Trans. on Systems, Man, and Cybernetics-Part A*, Vol. 36, No. 4, pp.727-745, 2006.

[14] B. Rezaee, "A cluster validity index for fuzzy clustering.", *Fuzzy Sets and Systems*, Vol. 161, No. 23, pp. 3014-3025, 2010

[15] D. A. Belsley, E. Kuh, R. E. Welsh, *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*, John Wiley & Sons, Inc, 1980.

[16] K. Wagstaff, C. Cardie, S. Rogers, S. Schroedl, "Constrained K-means Clustering with Background Knwledge", *Proceeding of the Eighteenth International Conference on Machine Learning*, pp.577-584. 2001.

[17] W. Wang, Y. Zhang, "On fuzzy cluster validity indices", *Fuzzy Sets and Systems* , Vol. 158, No. 19, pp. 2095-2117, 2007.

[18] K. L. Wu, M. S. Yang, "A cluster validity index for fuzzy clustering", *Pattern Recognition Letters*, Vol. 26, No. 9, pp. 1275-1291, 2005.

[19] X. L. Xie, G. Beni, "A validity measure for fuzzy clustering", *IEEE Trans on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 8, pp. 841-847, 1991.

AUTHOR PROFILE

Keun-Chang Kwak received the B.Sc., M.Sc., and Ph.D. degrees from Chungbuk National University, Cheongju, Korea, in 1996, 1998, and 2002, respectively. During 2003–2005, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. From 2005 to 2007, he was a Senior Researcher with the Human–Robot Interaction Team, Intelligent Robot Division, Electronics and Telecommunications Research Institute, Daejeon, Korea. He is currently the Associative Professor with the Department of Control & Instrumentation, Engineering and Department of Electronics Engineering, Chosun University, Gwangju, Korea. His research interests include human–robot interaction, computational intelligence, biometrics, and pattern recognition. Dr. Kwak is a member of IEEE, IEICE, KFIS, KRS, ICROS, KIPS, and IEEK.