Preprints of the 10th IFAC International Symposium on Dynamics and Control of Process Systems
The International Federation of Automatic Control
December 18-20, 2013. Mumbai, India

A New Cluster Validity Index for Fuzzy Clustering

Sreeram Joopudi*, Suraj S. Rathi, S. Narasimhan[a] and Raghunathan Rengaswamy[b]

** Senior Development Engineer, GyanData Private Limited, Email:sreeram@gyandata.com*

a,b- Dept. of Chemical Engineering, Indian Institute of Technology Madras, Email: naras@iitm.ac.in, raghur@iitm.ac.in

Abstract
Performance of any clustering algorithm depends critically on the number of clusters that are initialized. A practitioner might not know, *a priori*, the number of partitions into which his data should be divided; to address this issue many cluster validity indices have been proposed for finding the optimal number of partitions. In this paper, we propose a new "Graded Distance index" (GD_index) for computing optimal number of fuzzy clusters for a given data set. The efficiency of this index is compared with well-known existing indices and tested on several data sets. It is observed that the "GD_index" is able to correctly compute the optimal number of partitions in most of the data sets that are tested.

Keywords: Fuzzy c-mean clustering, Cluster validity index, GD_index.

1. Introduction

Clustering is an unsupervised learning approach for classifying data into groups such that points in the same group have higher resemblance to each other than to those of the other groups [8].Clustering is one of the important data mining techniques that enable the discovery of hidden relationships from data. Clustering finds its application in diverse areas such as computer science (image processing), business and marketing (recommending products to customers) and pattern recognition. One of the major problems in clustering algorithms is that one has to specify the number of clusters *c*, a-priori, so that the algorithm can partition the data into *c* clusters. The quality of the classification and separation of data into partitions depends on the value of the parameter *c* that is provided to the algorithm. For 2-dimensional data, it may be possible to guess the correct number of clusters (by plotting it); but, specifying the number of clusters is a difficult task for higher dimensional data.

The number of cluster partitions that are obtained are the same as the number of cluster centers *c* that is provided as input to the clustering algorithm[12]. As a result, if a smaller value of *c* than the optimal cluster number is initialized, the data will get under-classified (less groups than optimal), whereas, if a larger value of c than the optimal cluster number is used then the data will get over-classified (more number of groups than optimal). The identification of the optimal number of clusters is an important problem with relevance in several application areas. Without the correct number of cluster centers, obtaining accurate final outcomes/results, particularly for higher dimensional data is difficult. Further, with no further knowledge, extensive tests (numerical or otherwise) might have to be conducted to determine if the results are dependable. Therefore, an accurate value of parameter *c* is extremely significant in order to maximize the benefits out of clustering algorithms.

In order to overcome the above mentioned problem, several authors have proposed indices which reach an optimal value at the natural partitions of the data. For computing the value of these indices, the clustering algorithm is executed multiple times by varying the parameter *c* (number of clusters)and then one selects that cluster number which satisfies certain predefined criteria. Many indices have been proposed based on this principle for determining the optimal number of clusters. The compactness (of a cluster) and separation (between any two clusters) are the two major characteristic for cluster validity. If all the data points in a cluster are very close to each other, then one can say that that cluster is highly compact, whereas if the distance between two cluster centers is high, then those clusters are considered to be well separated.

2. Fuzzy C- Means (FCM) Clustering

Clustering can be divided into hard clustering and soft clustering. In hard clustering, clusters are separated by sharp boundaries; whereas in soft clustering there is overlap between clusters.

Fuzzy clustering is a soft clustering technique for classifying data into groups. In fuzzy clustering

each data point belongs to all the clusters with varying memberships and these membership values range between zero and one. For each data pointx_j, its membership value u_{ij} represents how strongly it belongs to i^{th} cluster. Therefore the farther a data point x_j is from the cluster $i$when compared with the distances between that point to remaining clusters, closer is its membership value u_{ij} to zero. The sum of membership values of a data point to all the clusters will be equal to one.

Let $X \in R^N$ denote the multi-dimensional data, then the FCM clustering algorithm [13,14] partitions the data set into c fuzzy groups. For the number of clusters (c) given by user, the objective is to obtain 'c' fuzzy partitions that minimize the sum of membership weighted intra-cluster variance among the partitions. The fuzzy partitions are obtained by minimizing the cost functionJ_n

$$Minimize \;\; J_n(U,\theta) = \sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}^{q} \left|\left|x_j - \theta_i\right|\right| \quad (1)$$

where N is the number of data points, i and j are the indices used to represent the cluster centres and data points respectively, and q is termed as fuzzifier [In all of our simulations, the value of fuzzifier is taken as 2]. The effect of the fuzzifier on the resulting cluster partition is described in [10]. q greater than one controls the overlap in the cluster regions. The value of q close to one results in hard clustering and if $q \to \infty$, clustering becomes totally fuzzy. $||x_j - \theta_i||$is the Euclidean distance of the data object x_j to the cluster centre θ_i. The solution to this optimization problem in Eq. (1) is obtained by iteratively updating the cluster centres and the respective membership functions through the following equations:

$$\theta_i = \frac{\sum_{j=1}^{N} u_{ij}^{q} X_j}{\sum_{j=1}^{N} u_{ij}^{q}}, \qquad 1 \le i \le c \quad (2)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{N}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{q-1}}}, \qquad 1 \le i \le c, 1 \le j \le N \quad (3)$$

Here,d_{ij} is the Euclidean distance of j^{th} data point to i^{th} cluster. Eq.2 and Eq. 3 help in iteratively improving clusters until the improvement in U (Membership matrix)is less than a specified tolerance value. The execution of FMC clustering algorithm is summarized in Table 1 [12].

Table 1 Fuzzy c-means algorithm

a)Fix number of clusters (c) and tolerance value(tol)
b) Generate initial membership matrix of the data with 'c' clusters
c) Let r be the iteration index. $r = 1, 2, 3, \ldots$
d) Update cluster centres (θ)and then membership values (U)by using

$$\theta_i^r = \frac{\sum_{j=1}^{N}(u_{ij}^{(r-1)})^q X_j}{\sum_{j=1}^{N}(u_{ij}^{(r-1)})^q}$$

$$u_{ij}^r = \frac{1}{\sum_{k=1}^{N}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{q-1}}}$$

e) Repeat the above step till

$$||U^r - U^{(r-1)}|| \ge tol$$

3. Cluster Validity indices

The problem of finding optimal number of clusters is addressed by several cluster validity indices. Wang et al. [1] have classified cluster validity algorithms into two types of categories where the first category of algorithms use only the membership matrix and the second category uses cluster centers and data points along with the membership matrix. In section 3.1.1, we will discuss the first category of algorithms and in 3.1.2 we will discuss the second category of algorithms.

3.1.1 Indices using only Membership Values:

1. Bezdek indices:

Bezdek proposed two cluster validity indices: a) Partition Coefficient (PC)[2]and b) Partition Entropy (PE) [3]. Bezdek suggested that the optimal number of partitions are obtained by minimizing the overall content of pairwise fuzzy intersection in the membership matrix (U). As the number of clusters is increased from 2 to n-1,the optimal number of clusters will be obtained at the maximum value ofPC or at the minimum value ofPE.

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} {u_{ij}}^2$$

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} * \log_e u_{ij}$$

Hereu_{ij}represents the membership of j^{th} data point towards i^{th} cluster, n – number of data points and c-number of clusters.

2.Dave index (DI):

In order to overcome the monotonic tendency of Bezdek's PC with number of clusters c, Dave proposed [4] a modified index as

$$D = 1 - \left(\frac{c}{c-1} * (1 - PC)\right)$$

wherePC is Partition coefficient and c is number of clusters

3.1.2 Indices using membership values, cluster centers and data

1. Kwon index (KI):

Kwon proposed an index, represented by K, [6] as an extension to Xie and Bini's (XB) index [5] to overcome the monotonically decreasing nature of XBas the number of cluster points approach the number of data points. Compared to XB index, Kwon index contains a penalizing function for higher number of clusters in the numerator, whereas the denominator represents the separation measure between clusters. A low value of K indicates high compactness and more separation between the clusters.

$$K = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} {u_{ij}}^2 \parallel x_j - v_i \parallel^2 + \frac{1}{c}\sum_{i=1}^{c} \parallel v_i - \bar{v} \parallel^2}{min_{j\neq k} \parallel v_j - v_k \parallel^2}$$

wherev_i represents thei^{th} cluster center and $\bar{v}$is the mean of cluster centers.

2.Zhang index (ZI):

Zhang computes the variation and separation measure between clusters and once both of these measures are computed for a range of clusters from 2toc_{max}, these measures are normalized individually. The optimal number of clusters corresponds to a minimum value of Zhang index ($Z(V,U)$) [7].

Variation measure, Var(U,V):

$$Var_c(U,V) = \left(\sum_{i=1}^{c}\sum_{j=1}^{n}\frac{u_{ij}d^2(x_j,\vartheta_i)}{n(i)}\right) * \left(\frac{c+1}{c-1}\right)^{\frac{1}{2}}$$

$n(i)$represents the number of samples in ith cluster.

$d(x,y)$is defined as

$$d(x,y) = [\,1 - \exp(-\beta \parallel x - y \parallel^2)]^{1/2}$$

Here β represents the inverse of the sample covariance of data.

Separation measure, Sep(c,U):

$$\text{Sep(c, U)} = 1 - \max_{i\neq j}[\max_{x_k \in X} \min(u_{ik}, u_{jk})]$$

Zhang's final validity index, $Z_c(V,U)$, is computed as

$$Z_c(V,U) = \frac{Var_{N,c}(V,U)}{Sep_N(c,U)}$$

Where $Var_{N,c}(V,U)$ and $Sep_N(c,U)$ represent the normalized variation and separation measures for 'c' clusters that are computed by dividing with their respective maximum values.

$$Var_{N,c}(V,U) = \frac{Var_c(V,U)}{Var_{max}} \;;$$

$$Var_{max} = \max\big(Var_c(V,U)\big), \qquad \forall\, c = 2\ to\ c_{max}$$

3.2 Motivation:

When all the points similar to each other are grouped towards one cluster and points which are dissimilar to each other are separated into different clusters, then one can say that the data has been divided into optimal number of partitions. When points similar to each other are grouped in one cluster, then the average distance between that cluster center and the data points is likely to be low and such clusters are referred as compact clusters. Xie and Bini [5] defined compactness of a fuzzy cluster as the weighted summation of distance between all the data points to that cluster center with weights being their membership values raised to fuzzifier power. With an increase in the number of clusters, it can be observed that the average distance between the cluster centers and the data points will decrease and in this process, the separation between cluster centers will also decrease. With the increase in the number of clusters, along with the compactness of fuzzy clusters the overlap between them will also increase. Thus, both compactness (high compactness) and overlap (low overlap) between cluster groups are two important qualities that are to be considered for optimal classification of data.

In FCM clustering, the association of a point towards various clusters can be compared by using the membership values of that point with respect to those clusters. Among all these membership values, the cluster (say A) which has the highest membership value can be chosen as the cluster that strongly possesses the data point. This maximum membership value can be used for interpreting the strength of association of the data point towards the cluster. Similarly, the second maximum membership value of the data point will indicate the strength of association to its next nearest cluster (say B). This second maximum membership value can be used to judge the strength of overlap

between these clusters (A and B) over this data point.

In general, before reaching the optimal number of clusters, the process of increasing the number of clusters will result in the splitting of big clusters. Consider a fuzzy cluster, say D, which is split into two smaller clusters, say E and F, when the number of clusters is increased by1.Now, consider the points which share their maximum membership value with the fuzzy cluster D. A majority of these points are likely to be closer to their cluster centers (E and F) when compared with their distance to the earlier cluster D. In most cases, if the number of clusters is less than the optimal number, splitting of clusters is likely to result in an increase in the strength of association of these data points towards fuzzy clusters E and F and this will be reflected in an increase in the *maximum membership value* for most of these points. Once the number of clusters exceeds the optimal number, an increase in the number of clusters will result in fuzzy clusters competing with each other on their strength of association over the data points. This will lead to an increase in the overlap of new fuzzy clusters. Consider a fuzzy cluster G that is split into clusters H and J with an increase in the number of clusters by 1. Now, the new clusters H and J will be competing with each other over the data points which share their maximum membership value towards the earlier cluster G. This leads to an increase in the overlap between the fuzzy clusters and this will be evident in an increase in the *second maximum membership value* for a majority of these points.

Proposed Index

The proposed Graded Distance Index (GD_index) uses both the maximum and second maximum membership values of all data points. The optimal number of clusters is obtained by: i) maximizing the strength of association of all data points towards their respective maximum membership cluster and ii) minimizing the overlap between fuzzy clusters over the data points. GD_index is computed by using the average difference between the first maximum membership and second maximum membership of all data points. A negative term, proportional to number of clusters, is added to this average value for penalizing high number of clusters. Thus, GD_index for 'c' clusters is

$$GD_{index,c} = \frac{\sum_{i=1}^{N}(u_{i,1stmax} - u_{i,2ndmax})}{N} - \left(\frac{c}{N}\right)$$

where

$GD_{index,c}$ – GD_index for 'c' cluster partitions
$u_{i,1stmax}$ – first maximum membership of i[th] point
$u_{i,2ndmax}$ – second maximum membership of i[th] point
N – total number of data points
c – number of clusters

By varying the number of clusters 'c' from 2 to N, Graded Distance index (GD_index) is computed for different 'c' values. The number of clusters c* which corresponds to the maximum value of GD_index is regarded as the optimal number of clusters to which the data should be partitioned**.**

Now, the variation of GD_index with an increase in the number of clusters will be studied for the sample Data set A shown in Fig 1.Fig.2represents the classification of clusters obtained from FCM clustering with cluster numbers ranging from 2 to 5. Even though the obtained clusters are fuzzy (all data points belong to all clusters but with varying memberships), in order to visually differentiate them, a unique color is assigned to all those points which share their highest membership values towards a single cluster. This visual distinction between fuzzy clusters will be helpful in identifying an approximate location of cluster centers and also in studying the qualitative impact of increasing cluster numbers on membership values. As the number of clusters is increased from 2 to4, from Fig. 2, it can be observed that the top and bottom clusters are being split into smaller clusters. Further increase in the number of clusters from 4 to 5 leads to further splitting of top-left cluster. This further splitting of top left cluster results in an increase in the overlap measure of the two new fuzzy clusters. This will be reflected in an increase in the second maximum membership value of majority of the points in top left cluster. As a result, GD_index value will decrease as the number of clusters is increased from 4 to 5. Table-1 lists the variation in GD_index with variation in cluster number from 2 till 7. From Fig.3 and Table-1, the maximum value of GD_index is obtained at **4 clusters** which is indeed the optimal number of clusters for the given data.

Table 1. GD_index for clusters ranging from 2 to 7

Cluster Number	2	3	4	5	6	7
GD index	0.74	0.71	**0.77**	0.68	0.63	0.6

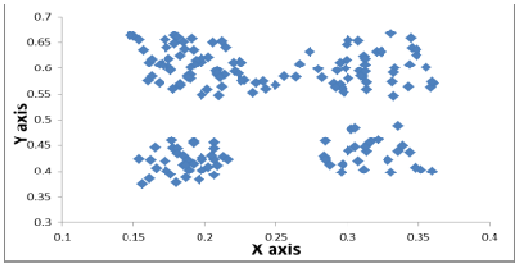

Fig. 1. Sample data set having four clusters

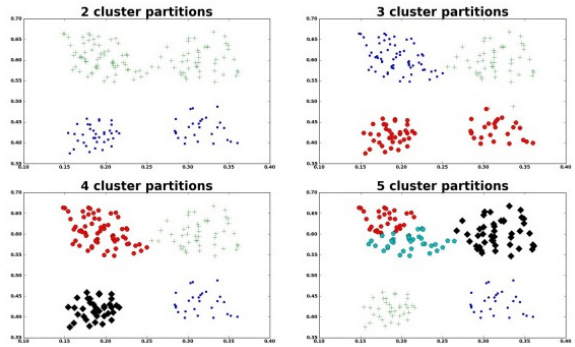

Fig. 2. FCM clustering partitions

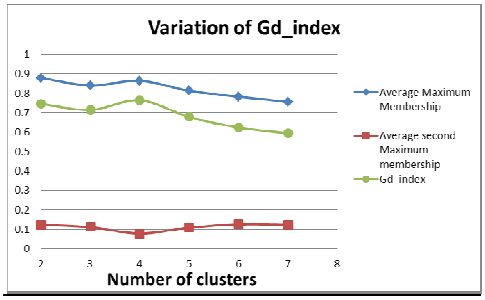

Fig. 3. GD_index variation with number of clusters

4. Comparing efficiency of "GD_index" with other cluster validity indices

For studying the validity of "GD_index" and its performance vis-a-vis the existing cluster validity indices, this index is tested with other validity indices on several two-dimensional data sets. The optimal number of clusters proposed by all these indices is compared with the actual number of clusters. Bezdek's PC, PE, Dave, Kwon and Zhang indices are the other cluster validity indices that are compared with "GD_index". In all these cases, it is found that "GD_index" is able to compute the optimal number of clusters (OC) correctly. A detailed description of the data sets and the results of all these cluster validity indices are described below.

Artificial Data sets

1. Data Set B:

In order to test the efficiency of the above indices on clusters with contrasting sizes, a dataset which has 3 clusters is simulated and plotted in Fig.4. The optimal number of clusters for this dataset is computed by using the above mentioned cluster validity indices. Only "GD_index" and "Kwon" indices identified 3 as the optimal number of clusters, whereas the remaining indices identified 2 as the optimal number. Fig.5 represents the optimal cluster partitions suggested by these six indices.Table-2 represents the values of these 6 indices as the number of clusters is increased from 2 to 7.

Table 2. Variation in cluster validity indices for dataset B

	2	3	4	5	6	7	OC
PC	**0.79**	0.75	0.63	0.59	0.57	0.54	2
PE	**0.15**	0.21	0.34	0.37	0.38	0.43	2
DI	**0.79**	0.75	0.58	0.58	0.57	0.54	2
GD	0.71	**0.73**	0.51	0.54	0.54	0.51	3
KI	209	**105**	420	228	181	161	3
ZI	**0.70**	1.06	0.88	0.84	0.84	0.88	2

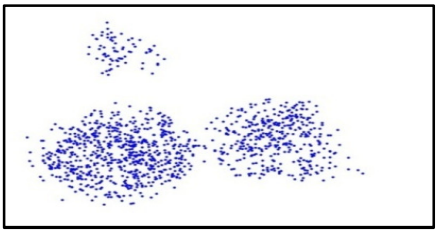

Fig. 4. Synthetic data set with different cluster sizes

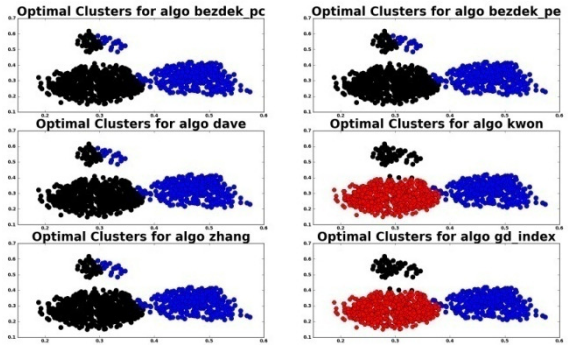

Fig. 5. Optimal clusters suggested by various indices

Table 3. Variation in cluster validity indices for dataset C

	2	3	4	5	6	7	OC
PC	0.91	**0.91**	0.83	0.78	0.72	0.79	3
PE	**0.06**	0.08	0.15	0.19	0.24	0.19	2
DI	0.91	**0.91**	0.82	0.77	0.72	0.78	3
GD	0.84	**0.86**	0.74	0.65	0.67	0.65	3
KI	**5.20**	5.43	47	100	121	80	2
ZI	0.36	**0.35**	0.48	0.64	0.83	1.01	3

2. Data Set C

To study the effect of outliers on these indices, data set which has 3 clusters and some outliers is simulated and plotted in Fig.6. On this data set, except PE and Kwon index, all indices suggested the correct optimal number of clusters as 3. Similar to data set B, here also the optimal cluster

partitions suggested by these indices is plotted in Fig.7 and in Table3the value of these cluster validity indices with an increase in the number of clusters from 2 to 7 are reported.

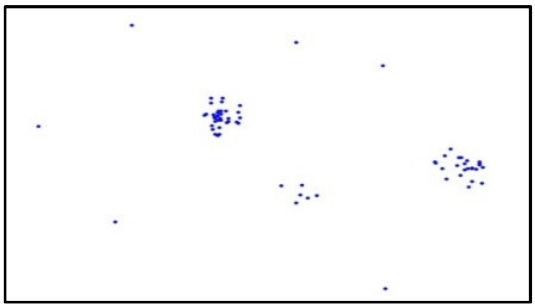

Fig. 6. Data set containing outliers

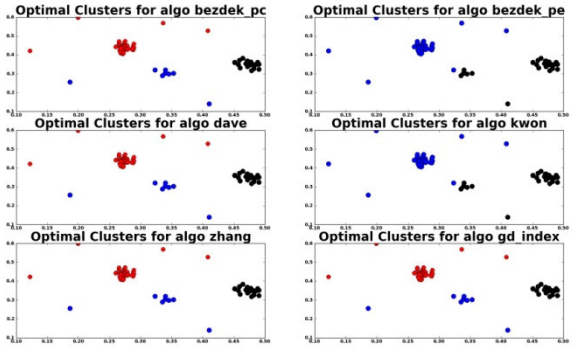

Fig. 7. Optimal clusters suggested for dataset C

3.Iris data set

Table 4. Variation in cluster validity indices for Iris data

	2	3	4	5	6	7	OC
PC	**0.89**	0.78	0.68	0.62	0.61	0.55	2
PE	**0.09**	0.17	0.25	0.31	0.35	0.39	2
DI	**0.89**	0.78	0.68	0.62	0.60	0.55	2
GD	**0.85**	0.72	0.60	0.61	0.55	0.48	2
KI	**8.39**	22	104	70	55	100	2
ZI	0.67	**0.62**	0.79	0.78	1.00	0.96	3

In this example, the performance of these indices is compared on the well-known iris data set [11]. Iris-setosa, Iris-versicolor and Iris-virginica are the three different classes of this data set and it contains information of four features namely a)sepal length b) sepal width c) petal length and d) petal width. Pal and Bezdek [9] mentioned that there is a significant overlap between two of these classes in the feature space and this has led several cluster validity indices to suggest 2 as the optimal cluster number. "GD_index" also suggests 2 as the optimal cluster number for this dataset. In Table 4, the optimal numbers of clusters suggested by the 6 indices are reported.

5. Conclusion

In this paper, we presented a new cluster validity index for fuzzy clustering algorithms. A good clustering method will keep similar points in one group and dissimilar points in different groups. Here, we propose a new "Graded Distance index" (GD_index) which uses only the fuzzy membership matrix (U). As this index: a) maximizes the summation of strength of association of all data points towards their nearest cluster and b) minimizes the cluster overlap over all the data points, a maximum value of this index is obtained at the optimal number of clusters in all the examples that we studied. We also compared our index with other cluster validity indices in the literature on various data sets.

References

1) W.Wang,Y.Zhang, On fuzzy cluster validity indices, Fuzzy Sets Syst. 158(19) (2007) 2095-2117.
2) J.C. Bezdek, Numerical taxonomy with fuzzy sets, J. Math. Biol. 1 (1974) 57–71.
3) J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybern. 3 (1974) 58–78.
4) R.N. Dave, Validating fuzzy partition obtained throughc-shells clustering, Pattern Recognition Lett. 17 (1996) 613–623.
5) X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. PatternAnal. Mach. Intell. 13 (1991) 841–847.
6) S.H. Kwon, Cluster validity index for fuzzy clustering, Electron. Lett. 34 (22) (1998) 2176–2177.
7) Y. Zhang, W. Wang, X. Zhang, Li Yi, A cluster validity index for fuzzy clustering, Inform. Sci. 178 (4) (2008) 1205–1218.
8) A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
9) N.R.Pal,J.C.Bezdek,Onclustervalidityforthefuzzy c-means model,IEEE Trans. FuzzySyst.3(1995)370–379.
10) Klawonn, F., Hoppner, F.: Advances in Intelligent Data Analysis V. Springer, Berlin (2003)
11) Iris data set downloaded from http://archive.ics.uci.edu/ml/datasets/Iris , Website last accessed on 18th April 2013.
12) Vidyashankar k, Rengaswamy R, Evaluation of prediction error based fuzzy model clustering approaches for multiple model learning ;Int J AdvEngSciAppl Math (March–June 2012) 4(1–2):10–21
13) J.C. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. Dissertation, Cornell University, Ithaca, NY, 1973.
14) J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, J. Cybernet. 3 (1974) 32–57.