AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Enhanced Validity for Fuzzy Clustering Using Microarray data

[1]V. Kumutha and [2]S. Palaniammal

[1]*Assistant Professor, Department of Computer Science, D.J. Academy for Managerial Excellence, Coimbatore. India.*
[2]*Professor and Head, Department of Science & Humanities, Sri Krishna College of Technology, Coimbatore. India.*

ARTICLE INFO

Article history:
Received 25 January 2014
Received in revised form 8 March 2014
Accepted 10 March 2014
Available online 2 April 2014

Keywords:
cluster validity indices, uncertainty, fuzzy clustering, fuzzy c-means, hesitation degree

ABSTRACT

Background: The goal of clustering algorithms is to expose the integral partitions of the data set. The important measures of **c**lustering involve at identifying right number of clusters and evaluating the quality of the partitions. **Objective:** In this paper a new validity index for fuzzy clustering is introduced, which assesses the average compactness and separation of fuzzy partitions produced by the fuzzy c-mean algorithm. Fuzzy clusters uses membership matrix to place the objects in all the partitions simultaneously. Fuzzy C-means (FCM) is one of the most widely used fuzzy clustering algorithms. One of the important parameter in FCM algorithm is the number of clusters(c) which has high influence over the resulting partition. Intuitionistic fuzzy set theory is apparently used in medicine, characterized by the values namely membership (belongingness), non-membership (non-belongingness) and hesitation (indeterminacy or uncertainty) of an element to that set. The proposed cluster validity index called as compactness and partition coefficient with hesitation degree (CPCHD) index includes the hesitation degree along with membership value in order to determine the optimal number of clusters for uncertain data. **Results:** Experimental results shows the efficacy of the proposed index with reliable cluster number with c=2(colon cancer and leukemia data set) or c=6(yeast and splice data set). **Conclusion:** Hesitation degree assesses the validity of the produced partitions from the FCM algorithm obtained by minimizing the value of c generating optimal results.

To Cite This Article: V.Kumutha, S.Palanimmal, Enhanced Validity for Fuzzy Clustering Using Microarray data. ***Aust. J. Basic & Appl. Sci.,*** *8(3): 7-15, 2014*

INTRODUCTION

Clustering is an unsupervised learning technique, which organize similar objects into groups, mostly applied to situations where prior knowledge of data is not available. The goal of clustering is to ascertain new set of categories (Rokach, Maimon, 2005). Clustering techniques are broadly classified into hard and soft partitions. The traditional hard partitioning methods allow one object to lie in only one cluster at a time. The hard partition gives undesirable results, i) while fixing an object that almost lie between two clusters and ii) while placing an outlier. This adverse situation can be fixed by fuzzy clustering. Fuzzy clustering allows one data item to belong to several clusters concurrently with different membership degrees. The assigning to a partition is determined by the membership degree that lies between 0 and 1(Babuska, 2009).

Fuzzy c-means (FCM) is the most common fuzzy clustering algorithm. In order to obtain good cluster it is important to set the parameters of the algorithm right. It highly depends on the initial parameters and needs estimation of the number of clusters. The problem of finding an optimal c is usually called cluster validity (Bezdek, 1974a,1974b). It is essential to validate each of the fuzzy partition generated, since different number of initial clusters produces different clustering partitions.

Several cluster validity indices have been proposed in the literature with categories such as i)using only the membership values and ii) involves both the membership value and the data set itself. The frequently used validity indices in recent research are Bezdek's partition coefficient(PC) (Bezdek, 1974a) and classification entropy(CE)(Bezdek, 1974b, 1981), partition index(SC)(Bensaid, Hall, Bezdek, Clarke, Silbiger, Arrington, Murtagh, 1996) separation index(S)) (Bensaid, Hall, Bezdek, Clarke, Silbiger, Arrington, Murtagh, 1996), Xie-Beni's index(XB)(Xie, Beni, 1991), Dunn's index(DI)(Dunn, 1974) and alternative dunn index(ADI)(Halkidi, Batistakis, Vazirgiannis, 2001). Compactness (closeness of cluster elements) and separation (distance between 2 different clusters) are the major criteria proposed for evaluation and selection of the optimal clusters. The real-world clustering applications are stuck with the uncertainty in the localization of the feature vectors.

Corresponding Author: V. Kumutha, Assistant Professor, Department of Computer Science, D.J. Academy for Managerial Excellence, Coimbatore, TN, India
E-mail: kumuthav@gmail.com

Uncertainty, fuzziness and vagueness are the major elements in fuzzy clustering, that again adds a hesitation in defining the membership function of the object.

Intuitionistic Fuzzy Sets (IFS) are generalized fuzzy sets used to handle the problem of uncertainty, coping with the hesitancy originating from the imprecise information (Atanassov, 1986, 1999). Membership and non-membership value are elements involved in this sets. The membership value represents the trueness of element to the set, non-membership value denotes the falseness of the element to the set. According to fuzzy set theory, a value between zero and one is assigned for membership and 1 minus the degree of membership is assigned as degree of non-membership of an element, which may not be always certain in reality.

The existing fuzzy validity index involves only the membership value and the data set in determining the optimality of the cluster number. The proposed index involves the hesitation degree along with the membership value in order to overcome the uncertainties in the real world application.

The rest of this paper is organized as follows. Section 2 briefs about the background of FCM fuzzy clustering algorithm and section 3 recalls few well known validity indices. Section 4 describes the formulation of the proposed validity index. Experimental results on data sets are given in section 5 and section 6 summarizes the conclusions of this study.

MATERIALS AND METHODS

Fuzzy c-mean algorithm:

The Fuzzy C-Means algorithm (FCM) is an iterative algorithm that finds clusters in data and which uses the concept of fuzzy membership, instead of assigning an object to a single cluster, each object will have different membership values on each cluster. It partitions set of n objects in R^d dimensional (Atanassov, 2003) space into c ($1 < c < n$) $O=\{o_1,o_2,\ldots o_n\}$ fuzzy clusters with $Z=\{z_1,z_2,\ldots.z_n\}$ cluster centers or centroids. The fuzzy clustering of objects is described by a fuzzy matrix μ with n rows and c columns in which n is the number of data objects and c is the number of clusters, $\mu_{ij,}$ the element in the i^{th} row and j^{th} column in μ, point out the degree of association or membership function of the i^{th} object with the j^{th} cluster. The characters of μ are as follows:

$$\mu_{ij} \epsilon [0,1] \forall\, j = 1,2, \ldots, n \tag{1}$$

$$\sum_{i=1}^{c} \mu_{ij} = 1 \,, \forall\, j = 1,2, \ldots, n \tag{2}$$

The objective function of FCM algorithm is to minimize the Eq. 3:

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^m \; d_{ij}^2 \;\; 1 \leq m < \alpha \tag{3}$$

where $d_{ij} = |o_i - z_j|$, where *m(m>1)* is a scalar termed the weighting exponent and controls the fuzziness of the resulting clusters and d_{ij} is the Euclidean distance from object o_i to the cluster center z_j. The z_j, centroid of the jth cluster, is obtained using Eq. (4).

$$Cj = \frac{\sum_i [\mu_j(x_i)]^m x_i}{\sum_i [\mu_j(x_i)]^m} \tag{4}$$

Algorithm 1. Fuzzy c-means:

1. Select m (m>1) and initialize the membership function values , μ_{ij} i=1,2,…c, j=1,2…n
2. Compute the cluster centers Z_j, j = 1, 2,..., n by using Eq. (4)
3. Compute Euclidean distance, d_{ij}, i = 1, 2,..., c; j=1,2,..., n
4. Update the membership function, μ_{ij} i = 1, 2... c; j=1,2,..., n by using below equation μ_{ij}

$$\mu_j(x_i) = \frac{[\frac{1}{d_{ji}}]^{1/m-1}}{\sum_{k=1}^{c} [\frac{1}{d_{ki}}]^{1/m-1}} \tag{5}$$

5. If not converged, go to step 2.

1. Validation indices for the fuzzy c-mean:

After finding a partition of data by a fuzzy clustering algorithm such as FCM, the objective is to determine whether the partition has presented the data structure correctly or not. The cluster validity problem is to determine the optimal number of clusters. Most of the fuzzy clustering methods assume an initial cluster number, c to describe the data structure completely. Cluster validity index method performs the validation of the generated fuzzy c-partition. c_{min} and c_{max} are the minimum and maximum number of partitions defined

respectively, where each c [c_{min} , c_{max}]. The optimal cluster number can be determined by calculating all partition indexes with all cluster numbers and compare by selecting a minimal of maximal index obtained. The several validity indices available are reviewed as follows.

a) Bezdek proposed the validity index partition coefficient(PC)(Bezdek, 1074a) associated with FCM defined as

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{2} \quad (6)$$

where $\frac{1}{c} \leq V_{PC} \leq 1$. The PC index indicates the average contents of pairs of fuzzy subsets in U, by combining into a single number. Most favorable cluster number c* can be obtained by solving $max_{2\leq c\leq n-1}V_{PC}$to produce the best clustering performance for the data set X.

b) Classification entropy(CE)(Bezdek, 1974b, 1981) was defined by Bezdek as

$$V_{CE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\ log_a\mu_{ij} \quad (7)$$

where a is the base of the logarithm. CE measures the fuzziness of the cluster partition similar to the Partition Coefficient. An optimal c* is obtained by minimizing $V_{CE}$to produce the best clustering performance for the data set X.

c) Partition Index(SC): it indicates the relative amount of the sum of compactness and separation of the clusters. It takes the division of fuzzy cardinality of each partition to find the sum of the individual cluster validity measures (Bensaid, Hall, Bezdek, Clarke, Silbiger, Arrington, Murtagh, 1996).

$$SC(c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}(\mu_{ij})^{m}||x_j-v_i||^2}{N_i\sum_{k=1}^{c}||v_k-v_i||^2} \quad (8)$$

Equal number of clusters produces valuable results for different partitions. Better separation of SC can be obtained by taking minimum value.

d) Separation Index(S): It uses a minimum-distance separation for partition validity (Bensaid, Hall, Bezdek, Clarke, Silbiger, Arrington, Murtagh, 1996).

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{2}||x_j-v_i||^2}{N min_{i,k}||v_k-v_i||^2} \quad (9)$$

e) Xie and Beni's Index (XB): XB involves compactness and separation between clusters(Xie, Beni, 1991).

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{m}||x_j-v_i||^2}{N min_{i,j}||x_j-v_i||^2} \quad (10)$$

The optimal number of clusters should minimize the value of the index.

f) Dunn's Index (DI): Dunn proposed to identify compact and well separated clusters(Dunn, 1974). So the result of the clustering has to be recalculated as it was a hard partition algorithm.

$$DI(c) = min_{i\in c}\left\{min_{j\in c,i\neq j}\left\{\frac{min_{x\in C_i,y\in C_j}d(x,y)}{max_{k\in c}\{max_{x,y\in C}d(x,y)\}}\right\}\right\} \quad (11)$$

The main drawback of Dunn's index is computational since calculating becomes computationally very expensive as c and N increase.

g) Alternative Dunn Index (ADI)(Halkidi, Batistakis, Vazirgiannis, 2001): to make calculation simple Dunn's index was altered. The dissimilarity function between two clusters $min_{x\in C_i,y\in C_j}d(x,y)$is rated in value from beneath by the triangle-non equality:

$$d(x,y) \geq |d(y,v_j) - d(x,v_j)| \quad (12)$$

where v_j is the cluster center of the j-th cluster.

$$ADI(c) = min_{i\in c}\left\{min_{j\in c,i\neq j}\left\{\frac{min_{x_i\in C_i,x_j\in C_j}d(y,v_j)-d(x_j,v_j)}{max_{k\in c}\{max_{x,y\in C}d(x,y)\}}\right\}\right\} \quad (13)$$

2. Proposed Validity Index:

Compactness and separation are the two measures that a good validation index should possess for a fuzzy c-partition. Let A={a_1,a_2,….,a_n}be a data set in R^s. Assume that $\mu = \{\mu_1, \ldots\ldots, \mu_c\}$ is a fuzzy c-partition based on a fuzzy clustering algorithm(eg. FCM). Figure 1 shows the framework of the model. In this paper a reliable validation functional is proposed which provides a solution to the problem of uncertainty that exists with the membership degree of an element to that set. Hence a hesitation degree used as a standard error for the membership degree to obtain the exact membership of an element to that set. This can help in obtaining an optimal cluster c for the data set. The proposed index involves two factors in determining the validity. The first factor indicates the compactness and the second factor finds the partition coefficient with hesitation degree which validates each cluster. These two factors united together to create a new validity index, called a compactness and partition coefficient with hesitation degree (CPCHD) index.

$$V_{CPCHD} = \sum_{I=1}^{N} \frac{e^{C-PCHD}}{n} \tag{14}$$

The first factor compactness(C) is defined as

$$C = \frac{\min\left(||x_i - v_j||^2\right)}{n} \tag{15}$$

C measures the average minimum square distance between data points and cluster centers. PCHD indicates the average difference of the overall context of pairwise fuzzy intersection in U, the partition matrix with the hesitation degree (uncertainty).

3. Validity index involving Hesitation degree:

The second factor partition coefficient with hesitation degree (PCHD) is defined as

$$PCHD = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\left(\mu_{ij}^{m} \mp \pi_{ij}^{m}\right)}{n} \tag{16}$$

3.1. Fuzzy Sets:

In fuzzy set theory, the membership of an element to a fuzzy set is a single value between zero and one.

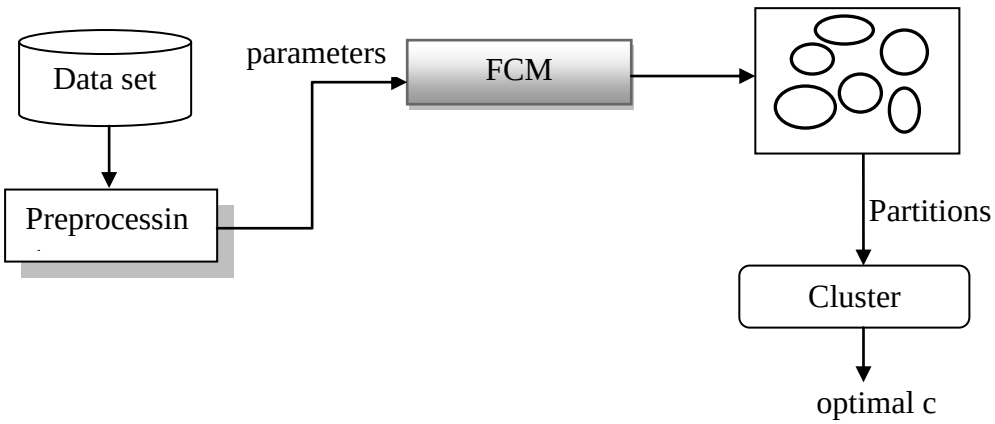

Fig. 1: Framework of the model.

But in reality, it may not always be certain that the degree of non-membership of an element in a fuzzy set is just equal to 1 minus the degree of membership. That is to say, there may be some hesitation degree. So, as a generalization of fuzzy sets, the concept of intuitionistic fuzzy sets was introduced by (Atanassov, 2003).

3.2 Intuitionistic Fuzzy Set:

The Intuitionistic Fuzzy Set (IFS) was defined as an extension of the ordinary Fuzzy Set(Atanassov, 1999). As opposed to a fuzzy set in *X*, given by:

$$A = \{(x, \boldsymbol{\mu_A}(x)) \mid x \in \boldsymbol{X} \} \tag{17}$$

where $\mu_A(x) \rightarrow [0,1]$ is the membership function of the fuzzy set *A*, an intuitionistic fuzzy set *B* is given by:

$$B = \{ x , \boldsymbol{\mu}_B (x)), \nu_B(x) \mid x \in X\} \quad (18)$$

where $\mu_B(x) \rightarrow [0,1]$ and $\nu_B(x) \rightarrow [0,1]$ are such that:

$$0 \leq \boldsymbol{\mu}_B (x)), \nu_{B(x)} \leq 1 \quad (19)$$

and $\mu_B(x)$,$\nu_B(x) \in [0,1]$ denote degrees of membership and non-membership of x ∈ B, respectively.

For each intuitionistic fuzzy set *B* in *X*, ''hesitation margin'' (or ''intuitionistic fuzzy index'') of $x \in B$ is given by:

$$\pi_B (x) = 1 - \boldsymbol{\mu}(x) - \nu_B(x) \quad (20)$$

which expresses a hesitation degree of whether *x* belongs to *B* or not. It is obvious that $0 \leq \pi_B(x) \leq 1$, for each x ∈X.

The CPCHD is defined for cluster i as

$$CPCHD_i = \sum_{l=1}^{N} \frac{e^{\min\left(||x_i - v_j||^2\right)/n} - \sum_{i=1}^{n}\sum_{j=1}^{c}\frac{\left(\mu_{ij}^m + \pi_{ij}^m\right)}{n}}{n} \quad (21)$$

The exponential function sets the compactness measure in the interval (0,1] and have the same degree(range) of measure. The total average of C-PCHD detects the data structure with a compact partition and well-separated clusters. Thus, an optimal c* can be found by solving $min_{2 \leq c \leq n-1} V_{CPCHD}$ to produce the best clustering performance for the dataset. Table 1 shows the membership value and the hesitation degree for the microarray data sets.

Table 1: Results of membership value and hesitation value for yeast, colon cancer, splice and leukemia data set.

Data set / No. of clusters	Yeast		Colon cancer		Splice		Leukemia	
$\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}$	μ_{ij}	π_{ij}	μ_{ij}	π_{ij}	μ_{ij}	π_{ij}	μ_{ij}	π_{ij}
2	0.7542	0.2001	0.3787	0.1325	0.7601	0.1209	0.7924	0.1098
3	0.5442	0.3962	0.6167	0.1783	0.4829	0.3482	0.2909	0.5829
4	0.3297	0.5108	0.1488	0.5108	0.5349	0.4323	0.4892	0.4423
5	0.5323	0.2534	0.5644	0.1519	0.6599	0.2369	0.6892	0.2428
6	0.1964	0.7332	0.2034	0.7316	0.3892	0.5822	0.2091	0.4782
7	0.3244	0.4792	0.8621	0.0012	0.2439	0.6982	0.5921	0.3821
8	0.9152	0.0022	0.4579	0.3162	0.4523	0.3369	0.1879	0.7528
9	0.4233	0.5631	0.7391	0.2213	0.9237	0.0013	0.6232	0.2781
10	0.9201	0.0281	0.5429	0.3901	0.4236	0.2244	0.5291	0.3462

The procedural steps for the validation of the FCM using the proposed validity index V_{CPCHD}, where $V_{CPCHD}^{(\min)}$ denotes the minimum value of index is given as follows:

Step 1: Initialize the parameters related to the FCM and the validity index:

c=2, c_{max} =10, $V_{CPCHD}^{(\min)}$=0, m=2, ε =0.001.

Step 2: With the initial assignment of m, weighting exponent, the membership values are initialized such that $\sum_{i=1}^{c} \mu_{ij}$ =1.0, where i=1, 2,…c, j=1,2…n.

Step 3: Update the fuzzy cluster centroid v_i and fuzzy membership using the equations

$$Cj = \frac{\sum_i [\mu_j(x_i)]^m x_i}{\sum_i [\mu_j(x_i)]^m} \quad (22)$$

$$\mu_j(x_i) = \frac{[\frac{1}{d_{ji}}]^{1/m-1}}{\sum_{k=1}^{c}[\frac{1}{d_{ki}}]^{1/m-1}} \quad (23)$$

Step 4: If the improvement in objective function is less than a certain threshold ε, then go to step 5: otherwise go to step 3.

Step 5: Compute the non-membership value and hesitation degree for the fuzzy partition obtained in step 4.

Step 6: Find the minimum$V_{CPCHD}^{(\min)}$, and report the value of c that minimizes V_{CPCHD} as the optimal number of clusters.

$$V_{CPCHD} \leftarrow \min V_{CPCHD} \tag{24}$$

The validation algorithm runs the FCM algorithm and computes the proposed validity index with respect to c=2,3,….,c_{max}.

RESULTS AND DISCUSSION

Comparisons were made with various data sets to demonstrate the proposed validity index performance. The proposed index was compared with seven fuzzy cluster validity indices mentioned in section 3: Bezdek's partition coefficient(PC) and classification entropy(CE), partition index(SC), separation index(S), Xie-Beni's index(XB), dunn's index(DI) and alternative dunn index(ADI).

In the experiments presented here, the cluster validation is determined to obtain the optimal cluster number c.

a. Validation performance:

The cluster validity index was tested for four data sets (available at http://kzi.polsl.pl/~jbiesiada/Infosel/files/datasets.html,http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html). The validity indexes discussed for the study have been implemented using MATLAB. The fuzzy cluster validity index performance varies with the fuzzy clustering algorithm. The FCM algorithm can easily able to discriminate the cluster validity with cluster number c varying from 2 to c_{max}.

The parameters of the FCM were set to a termination criterion ε=0.001, and weighting exponent m=2.0, and $\|*\|^2$ was the Euclidean norm. Random selection made for the assignment of initial centroids. Four data sets were used to evaluate the validation performance of each index: the yeast, colon cancer, splice and leukemia data sets.

Table 2: Description of the five data sets

Data set	No. of samples	No. of genes
Yeast	79	2467
Colon cancer	62	2000
Leukemia	38	7129
Splice	3190	60

Table 2 describes the number of samples and genes of the microarray data sets. Table 3-6 display the results of the evaluation of each index for the four data sets; the optimal value of c for each index is marked in bold face.

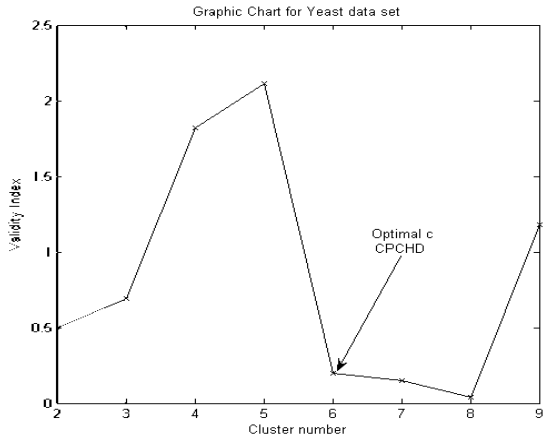

Fig. 2: Plot of Validity indices of Yeast data set.

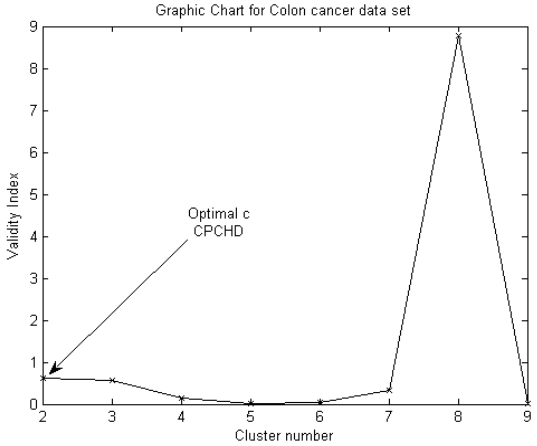

Fig.3: Plot of Validity indices of colon cancer data set.

V_{PC} and V_{DI} take their maxima as optimal values, whereas the other indices take their minima as optimal values. Table 3 lists the results of validity indexes for yeast data set which contains 79 samples where, c=2,3,…,10. For each c≥2, index values were computed for each of the 8 validity indexes considered. The optimal c's of V_{PC} and V_{CE} were at c=2, whereas for V_S and V_{XB} were at c=10 and for the proposed index were at c=6.

Table 4 shows the validity indices values for colon cancer data set obtained from various validity indices with c=2,3,…,10. The optimal number of clusters c=2 is correctly identified by V_{PC}, V_{CE} and V_{CPCHD} whereas V_{XB} yielded the optimal partitions at c=4.The optimal values are identified at c=10 by V_S and V_{SC}.

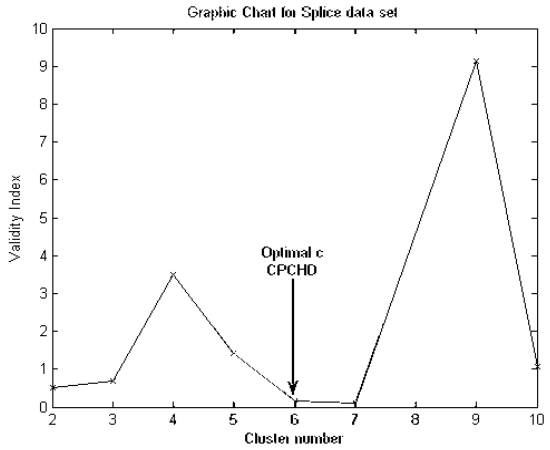

Fig. 4: Plot of Validity indices of Splice data set

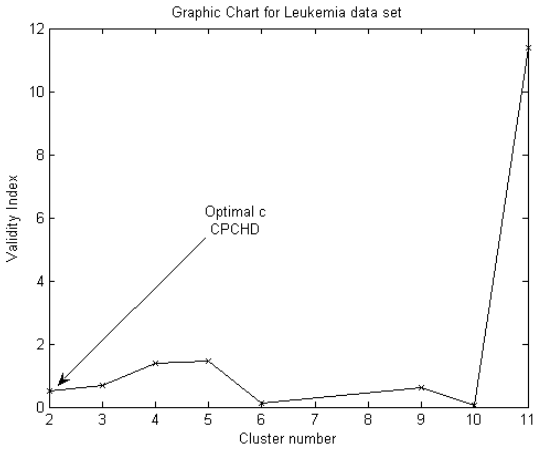

Fig. 5: Plot of Validity indices of Leukemia data set.

Table 3: Values of Validity indices for yeast dataset.

c	PC	CE	SC	S	XB	DI	ADI	CPCHD
2	**0.5000**	**0.6931**	4.3617	2.6190	1.0146	**0.1566**	**0.0485**	3.9483
3	0.3333	1.0986	4.2134	5.1674	0.6764	0.1542	0.0394	5.9225
4	0.2500	1.3863	3.1654	5.7446	0.5073	0.1478	0.0110	7.8966
5	0.2000	1.6094	3.7544	6.7196	0.4058	0.1542	0.0067	9.8708
6	0.1667	1.7918	2.5185	4.6297	0.3382	0.1542	0.0078	**1.1845**
7	0.1429	1.9459	**1.8211**	3.5533	0.2898	0.1494	0.0066	1.3819
8	0.1250	2.0794	3.5534	3.2562	0.2536	0.1542	0.0092	1.5793
9	0.1111	2.1972	4.2700	3.2861	0.2254	0.1542	0.0042	1.7767
10	0.1000	2.3026	5.2058	**2.1197**	**0.2029**	0.1542	0.0091	1.9742

Table 4: Values of Validity indices for Colon cancer dataset.

c	PC	CE	SC	S	XB	DI	ADI	CPCHD
2	**0.6233**	**0.5603**	0.2184	0.0035	1.0279	0.3163	0.0931	**0.0187**
3	0.4167	0.9610	0.1895	0.0048	0.6546	0.2792	0.0738	0.0368
4	0.3144	1.2434	0.1805	0.0047	**0.0523**	0.3016	0.0132	0.0523
5	0.2532	1.4647	0.1795	0.0044	0.3997	0.2784	0.0029	0.0685
6	0.2117	1.6448	0.1767	0.0041	0.3374	0.3081	0.0042	0.0703
7	0.1824	1.7972	0.1712	0.0040	0.2936	0.2785	0.0023	0.0668
8	0.1612	1.9258	0.1654	0.0040	0.2597	0.3081	0.0034	0.0715
9	0.1502	2.0250	0.1469	0.0036	0.2472	0.2939	5.3408	0.0748
10	0.1377	2.1257	**0.1446**	**0.0033**	0.2314	**0.3258**	**8.7724**	0.0817

Table 5 shows the performance of the validation methods for the splice data set of the various validity indices with c=2, 3,…,10. The optimal number of clusters c=6 is correctly identified by $V_{CPCHD,}$ where as V_{PC}, V_{CE} and V_S yielded the optimal partitions at c=2.The optimal values are identified at c=10 by V_{XB} and V_{SC}.

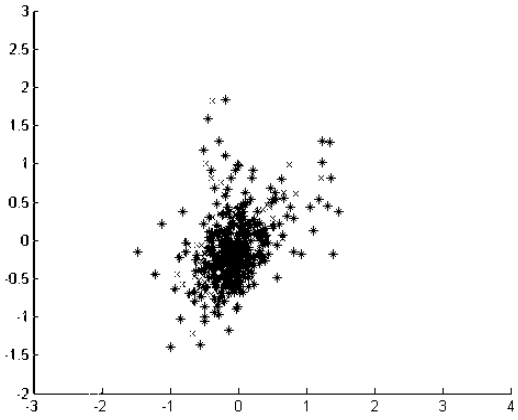

Fig. 6: Clustered yeast data after application of FCM for c=6

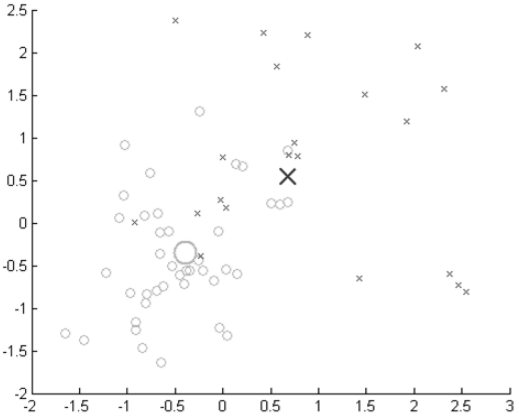

Fig. 7: Clustered colon cancer data after application of FCM for c-2.

The results of the validity indices of leukemia data set are presented in table 6. It shows that $V_{CPCHD,}$ V_{PC} and V_{CE} have yielded the optimal partitions at c=2, whereas V_{XB} and V_S gives optimal c at 10. Figures 6,7,8 and 9 show respectively, the partitions on yeast, colon cancer, splice and leukemia data sets acquired by applying FCM with the number of clusters identified by the proposed cluster validity index, CPCHD index.

In Figure 2, validity function for the yeast data set is graphed to provide the optimal cluster estimate at c=6. It is shown that c=2 as the optimal cluster number for the colon cancer data set plotted in Figure 3. Figure 4 shows the optimal cluster value c=6 for splice data set. Graph shown in Figure 5 describes c=2 as the optimal cluster estimate.

Conclusion:

The quality of the partition can be determined by the cluster validity index. One of the important initial parameter for the FCM fuzzy clustering algorithm is the number of clusters to be generated, which highly reflects the quality of the resulting partition. The validity indices proposed in the literature are dependent upon the membership and the data itself for validity calculation.

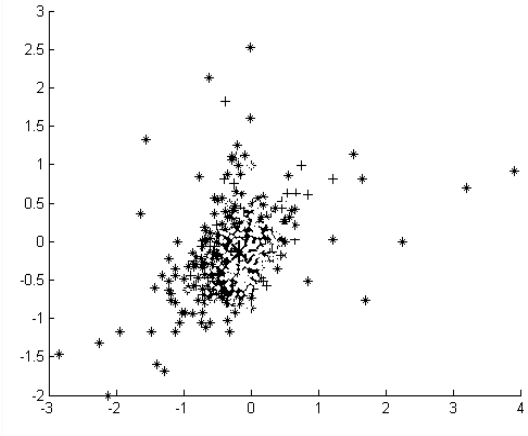

Fig. 8: Clustered splice data after application of FCM for c=6

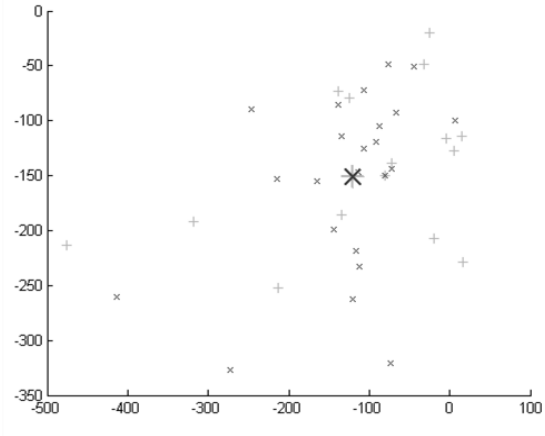

Fig. 9: Clustered Leukemia data after application of FCM for c-2.

Table 5: Values of Validity indices for splice dataset

c	PC	CE	SC	S	XB	DI	ADI	CPCHD
2	**0.5000**	**0.6931**	4.5703	**1.4327**	0.7788	**0.1107**	0.0060	3.6060
3	0.3333	2.0986	6.3082	2.9374	0.5192	0.0842	0.0063	5.4090
4	0.2500	1.3863	8.2003	4.1635	0.3894	0.0842	0.0063	7.2121
5	0.2000	1.6094	**3.5018**	1.5952	0.3115	0.0852	0.0064	9.0151
6	0.1667	1.7918	5.9621	2.8283	0.2596	0.0595	0.0063	**1.0818**
7	0.1429	1.9459	7.7484	3.2955	0.2225	0.0825	0.0015	1.2621
8	0.1250	2.0794	8.6251	4.0114	0.1947	0.0484	0.0015	1.4424
9	0.1111	2.1972	7.1447	3.8956	0.1731	0.0595	9.0040	1.6227
10	0.1000	2.3026	4.4623	2.2648	**0.1558**	0.0593	**9.1491**	1.8030

Table 6: Values of Validity indices for leukemia dataset

c	PC	CE	SC	S	XB	DI	ADI	CPCHD
2	**0.5000**	**0.6931**	2.0552	5.4084	0.6378	0.5586	0.0320	**11.3831**
3	0.3333	1.0986	6.3721	2.3904	0.4252	**0.6128**	**0.0351**	17.0746
4	0.2500	1.3863	5.2264	2.0068	0.3189	0.5360	0.0284	22.7662
5	0.2000	1.6094	5.8293	2.5136	0.2551	0.5882	0.0038	28.4577
6	0.1667	1.7918	3.5899	1.4639	0.2126	0.5586	0.0019	34.1493
7	0.1429	1.9459	2.0216	7.5091	0.1822	0.5998	0.0039	39.8409
8	0.1250	2.0794	4.8351	1.9503	0.1595	0.5882	0.0020	45.5323
9	0.1111	2.1972	**1.3936**	5.1390	0.1417	0.5126	0.0243	51.2241
10	0.1000	2.3026	3.6922	**1.4582**	**0.1276**	0.5129	0.0243	56.9155

After reviewing several validity indices a new validity index is proposed, CPCHD index. The proposed CPCHD index uses hesitation degree which copes with the uncertainty issue associated with the current real data sets. Along with the membership degree, another measure that helps in proper evaluation result is the

hesitation degree. The compactness and partition coefficient with hesitation degree provides the necessary component in identifying the well-separated and compact cluster. It has given valid results when applied for the microarray data sets: yeast, colon cancer, splice and leukemia. The data sets were compared with the existing validity indices: PC, CE, SC, S, XB, DI, ADI and CPCHD. The potential associated with the proposed index CPCHD is the hesitation degree which assess the validness of the partitions generated from the FCM clustering algorithm. The optimal fuzzy c-partition is obtained by minimizing V_{CPCHD} with respect to c. The results of the experimental tests in which various indices were used to determine the optimal number of clusters for microarray data sets showed that the proposed index delivers a reliable result.

REFERENCES

Atanassov, K., 1986. Intuitionistic fuzzy sets, Fuzzy Sets and Systems, 20: 87-96.

Atanassov, K., 1999. Intuitionistic Fuzzy Sets Theory and Applications, Studies in Fuzziness and Soft Computing, Phisica-Verlag.

Atanassov, K., 2003. Intuitionistic fuzzy sets: past, present and future. In: Proceedings of the 3rd Conference of the European Society for Fuzzy Logic and Technology, 12-19.

Babuska, R., 2009. Fuzzy Clustering in Fuzzy and neural control disc : Course lecture notes. Delft, the Netherlands: Delft University of Technology.

Bensaid, A.M., L.O. Hall, J.C. Bezdek, L.P. Clarke, M.L. Silbiger, J.A. Arrington and R.F. Murtagh, 1996. Validity-guided (Re)Clustering with applications to image segmentation.*IEEE Transactions on Fuzzy Systems*, 4: 112-123.

Bezdek, J.C., 1974a. Numerical taxonomy with fuzzy sets. J. Math. Biology, 1: 57-71.

Bezdek, J.C., 1974b. Cluster validity with fuzzy sets. J.Cybernit., 3: 58-72.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

Dunn, J.C., 1974. Well separated clusters and optimal fuzzy partitions, *J. Cybern.*, 4: 95-104.

Halkidi, M., Y. Batistakis, M. Vazirgiannis, 2001. On Clustering Validation Techniques. Journal of Intelligent Information Systems, 17(2/3): 107-145.

Rokach, L., O. Maimon, 2005. Clustering Methods in The Data Mining and Knowledge Discovery Handbook, (Editors: O. Maimon and L. Rokach), pp: 321-352.

Xie, X.L. and G.A. Beni, 1991. Validity measure for fuzzy clustering. *IEEE Trans. PAMI*, 3(8): 841-846.