A New Validation Index for Determining the Number of Clusters in a Data Set

Haojun Sun[1], Shengrui Wang[1], Qingshan Jiang[2]

[1]Dept. of Math. and Computer Sci.
Univeristy of Sherbrooke
Sherbrooke, Qc, Canada, J1K 2R1
{sun,wang}@dmi.usherb.ca

[2]Generation 5
243 Consumers Road, 8th Floor
Toronto, Ont. Canada, M2J 4W8
qingshan@generation5.net

Abstract

Clustering analysis plays an important role in solving practical problems in such domains as data mining in large databases. In this paper, we are interested in Fuzzy C-Means (FCM) based algorithms. The main purpose is to design an effective validity function to measure the result of clustering and detecting the best number of clusters for a given data set in practical applications. After a review of the relevent literature, we present the new validity function. Experimental results and comparisons will be given to illustrate the performance of the new validity function.

1 Introduction

Clustering analysis is a fundamental process of data analysis. It is based on partitioning a set of data points into a number of clusters, where the data points inside each cluster exhibits similarity. It is a very active subject of research because of the important role it plays in solving practical problems in such domains as data mining in large databases, financing, pattern recognition and image processing. Similarity is often defined by a distance measure and an objective function is optimized in order to find a good partition of data. A common class of such algorithms for clustering are partitioning methods in which a set of centers (also called seeds or representative objects) are computed, each of which represents a cluster, and the membership of a data point to a cluster is defined based usually based on its distances with each center. Examples of these algorithms are K-Means, FCM and PCM [1][2] [3]. In the work reported in this paper, the target algorithm is Fuzzy C-Means (FCM). The fuzzy clustering approach was chosen for its robust performance in dealing with real data. In fact, according to many studies such as that of *Baraldi* and *Blonda* [4], fuzzy clustering has the potential to decrease dependency on initialization and reduce the presence of dead units which are two serious problems with hard competitive clustering algorithms such as the K-means algorithm. The FCM algorithm is also one of the most widely used fuzzy clustering algorithms.

Cluster validation is an important issue in clustering analysis because clustering algorithms are unsupervised in nature and the result of clustering needs to be validated in most applications. For instance, most algorithms assume that the number of clusters in a data set is a user parameter. However, it is hardly the case that the user is able to answer is how many clusters are contended in the data set. The cluster validity problem here relates to the measurement of how well the structure that is present in the data set has been identified. It is clear that a "good" clustering of a data set should be based on an accurate number of clusters. Therefore, the practical approach to clustering a data set tests a range of possible numbers of cluster, and discerns a "score" for each possible number of clusters based on the clustering results. The best number of cluster would have the best "score". In the literature, this score is named as validity function, validity measure, validity index, etc.

Several validity functions for fuzzy clustering algorithms, such as partition coefficient, classification entropy and so on, have been used for measuring the validity mathematically. A better method to define a validity function for measuring the clustering results is to consider two conflicting factors: compactness within each cluster and separation between clusters. *Xie et al* (1991) defined a well-known validity index [5] using the ratio between the compactness and the separation, *Fakuyama* and *Sugeno* (1989) defined another validity index [6] using the discrepancy

0-7803-7044-9/01/$10.00 ©2001 IEEE

of the compactness and the separation. *Rezaee* and *Letlieveldt* (1998) gave another validity index [7] using a liner combination of the compactness and the separation. All these validity functions have proved to be effective in detecting the number of the clusters when they do not overlap each other.

In this paper, we define a new validity index based on combining the average within-cluster scatter and the relative between-clusters distance. The new validity index, coupled with the FCM algorithm, has been compared with a number of major indices found in the literature. Experiments have been made on synthetic and real data sets. The results show that the new validity index is particularly efficient when there are overlapping clusters. In the follows, we will first introduce the general FCM-based algorithm for determining the number of clusters. Then, we will introduce the new validity index as well as a number of existing indices. Experiments and comparison results will be presented thereafter before the concluding remarks.

2 Algorithm for Determining the Number of Clusters

2.1 Basic FCM Algorithm

The FCM algorithm dates back to 1973. Many derivatives have been proposed with modified definitions for the norm and prototypes for cluster centers [8]. FCM-based algorithms are the most widely used fuzzy clustering algorithms in practice. The basic FCM algorithm can be formulated as follows :

$$Minimize\ J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}^{m}\|x_k - v_i\|^2, \quad (1)$$

where n is the total number of data in a given data set and c is the number of clusters; $X = \{x_1, x_2, \cdots, x_n\} \subset R^s$ and $V = \{v_1, v_2, \cdots, v_c\} \subset R^s$ are the given set of feature data and the set of cluster centers; $U = (u_{ki})_{n\times c}$ is a fuzzy partition matrix composed of the membership of each feature vector x_k in each cluster i. u_{ki} should satisfy $\sum_{i=1}^{c} u_{ki} = 1$ for $k = 1, 2, \ldots, n$ and $u_{ki} \geq 0$ for all $i = 1, 2, \ldots, c$ and $k = 1, 2, \ldots, n$. The exponent $m > 1$ in $J_m(U, V)$ (Equation (1)) is a parameter which modifies the weighting effect of the membership value. Large m tends to result in approximately equal membership values u_{ki}, thus increasing the fussiness of clusters. $d(x,y) = \|x-y\|$, $x, y \in R^s$ is a distance function (for example, Euclidean distance). To minimize $J_m(U,V)$, the cluster centers (prototypes) v_i and the membership matrix U need to be computed according to the following iterative formula:

$$u_{ki} = \left\{ \begin{array}{ll} \left(\sum\limits_{j=1}^{c} \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{\frac{2}{m-1}} \right)^{-1} & \text{if } x_k \neq v_i \\ 1 & \text{if } x_k = v_i \end{array} \right\} \quad (2)$$

$$v_i = \frac{\sum\limits_{k=1}^{n} u_{ki}^{m}\, x_k}{\sum\limits_{k=1}^{n} u_{ki}^{m}} \quad (3)$$

Where $k = 1, 2, ..., n$, $i = 1, 2, ..., c$. The cluster centers v_i, $i = 1, 2, \ldots, c$, are initialized by some method (for example, Random initialization) and the initial elements of the membership matrix, u_{ki} ($k = 1, 2, \ldots, n$, $i = 1, 2, \ldots, c$), are computed using Equation (2). To refine V and U, Equations (3) and (2) are used iteratively until the changes in V or U are sufficiently small. For final classification, the largest value of $u_{ki}(i = 1, 2, \ldots, c)$ is selected for any x_k, and the corresponding i_0 identifies the cluster to which the x_k belongs.

In practical applications of the FCM algorithm, one has to solve several problems including determination of the number of clusters and initialization of prototypes. The problem related to determination of the number of clusters is particularly important because the user does not generally know the exact number of clusters in the data set. The performance of clustering algorithms (FCM for example) in terms of the clustering results can be affected significantly if the number of clusters given is not accurate.

2.2 Determination of the Number of Clusters

Next, we give a general algorithm to detect the number of clusters in a given data set. The following algorithm applies the FCM clustering algorithm to the data set for $c = C_{\max}, ..., C_{\min}$ and chooses the best number based on a (cluster) validity criterion. Here the $C_{\max}$ and $C_{\min}$ are, respectively, the maximal and minimal numbers of clusters, and need to be provided by the user.

Algo: FCM based algorithm

```
1. Choose $C_{\max}$ and $C_{\min}$.
2. For $c = C_{\max}$ to $C_{\min}$:
    2.1. Initialize cluster centers ($V$).
    2.2. Apply the basic FCM algorithm to update the membership matrix ($U$) and the cluster centers ($V$).
    2.3. Test for convergence, if no, go to 2.2.
    2.4. Compute a validity value $V_d(c)$.
3. Compute $c_f$ such that the cluster validity function $V_d(c_f)$ is optimal.
```

There exist several techniques for initializing cluster centers (for the Step 2.1). For the sake of simplicity, random initialization has been used in our experiments.

3 New Validity Index

The function $V_d(c_f)$ in the **Algo** measures the *goodness* of the results of a clustering algorithm. A partition is considered good if it optimizes at least two conflicting criteria. One of these is related to within-class scattering, which needs to be minimized; the other to between-class scattering, which needs to be maximized. Several major validity functions are reviewed in this section and their performance is compared to the new one proposed here.

3.1 Validity Indices For Fuzzy Clustering

There are a number of cluster validity indices available. Some of them use only the membership value of a fuzzy partition of the data (membership matrix), others use original data and computed cluster centers as well as the membership matrix. Here are some of the indices most frequently referred to in the literature.

- Partition coefficient V_{PC}:

$$V_{PC}(U,C) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}^2 \qquad (4)$$

- Partition entropy V_{PE}:

$$V_{PE}(U,C) = -\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\log_a(u_{ki}) \qquad (5)$$

V_{PC} and V_{PE} are two simple indices that are computed using only the elements of the membership matrix. Both indices have a monotonic increasing (decreasing) tendency when c increases and they do not handle the data well when there is overlap between (true) clusters.

- *Xie*'s validity V_{Xie}:

$$V_{Xie}(U,V,c) = \frac{\sum\limits_{k=1}^{n}\sum\limits_{i=1}^{c} u_{ki}^m \|v_i - x_k\|^m}{n \star \min\limits_{i \neq j} \|v_i - v_j\|} \qquad (6)$$

- FS validity V_{FS}:

$$V_{FS}(U,V,c) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}^m (\|x_k - v_i\|^2 - \|v_i - \overline{v}\|^2) \qquad (7)$$

where $\overline{v} = \frac{1}{n}\sum\limits_{i=1}^{c} v_i$. $V_{Xie}(U,V,c)$ is a trade-off between compactness and separation. To obtain good clustering results, $V_{Xie}(U,V,c)$ needs to be minimized. $V_{FS}(U,V,c)$ measures the discrepancy between compactness and separation. Minimum $V_{FS}(U,V,c)$ is believed to correspond to the best clustering result.

- *Rezaee*'s validity V_{Rez}:

$$V_{Rez}(U,V,c) = \alpha * Scat(c) + Dis(c) \qquad (8)$$

where $\alpha = Dis(C_{max})$. $Scat(c)$, the average scattering, is defined as $\frac{\frac{1}{c}\sum\limits_{i=1}^{c}\|\sigma(v_i)\|}{\|\sigma(X)\|}$, where $\sigma(X) = \{\sigma(X)^1, \sigma(X)^2, \cdots, \sigma(X)^s\}^T$, $\sigma(X)^p = \frac{1}{n}\sum\limits_{k=1}^{n}(x_k^p - \overline{x}^p)^2$, $\sigma(v_i) = \{\sigma(v_i)^1, \sigma(v_i)^2, \cdots, \sigma(v_i)^s\}^T$, $\sigma(v_i)^p = \frac{1}{n}\sum\limits_{k=1}^{n} u_{ki}(x_k^p - v_i^p)^2$, $\overline{x} = \frac{1}{n}\sum\limits_{k=1}^{n} x_k$, for $p = 1, 2, \cdots, s$. The distance function is defined as $Dis(c) = \frac{D_{max}}{D_{min}}\sum\limits_{i=1}^{c}\left(\sum\limits_{j=1}^{c}\|v_i - v_j\|\right)^{-1}$, where $D_{min} = \min\limits_{i \neq j}\|v_i - v_j\|(i,j \in [1,c])$, $D_{max} = \max\limits_{i,j}\|v_i - v_j\|(i,j \in [1,c])$. $Scat(c)$ indicates the compactness of the partition. A small value of $Scat(c)$ means that, in average, the clusters are compact compared to the variance of the data set. $Dis(c)$ indicates the total scattering (separation) between the clusters. The weighting factor $\alpha = Dis(C_{max})$ is introduced to compensate for differences in the scales of $Dis(c)$ and $Scat(c)$. The minimum value of V_{Rez} is believed to correspond to the best clustering.

3.2 A New Validity Index

The validity index $V_{WSJ}(U, V, c)$ we propose has the following form:

$$V_{WSJ}(U, V, c) = Scat(c) + \frac{Sep(c)}{Sep(C_{\max})} \qquad (9)$$

Here $Scat(c)$ is defined in the same way as in Rasae's index. It represents the compactness of the obtained clusters. The value of $Scat(c)$ generally decreases when c increases because the clusters become more compact. The range of $Scat(c)$ is between 0 and 1. The term representing the separation between clusters is defined as $Sep(c) = \frac{D_{max}^2}{D_{min}^2} \sum_{i=1}^{c} \left(\sum_{j=1}^{c} \|v_i - v_j\|^2 \right)^{-1}$, where $D_{min} = \min_{i \neq j} \|v_i - v_j\|$ and $D_{max} = \max_{i,j} \|v_i - v_j\|$. We can also write $Sep(c)$ as $Sep(c) \doteq \frac{D_{max}^2}{D_{min}^2} E[\frac{1}{d_c^2}]$, where d_c is the average distance between two cluster centers. Both $\frac{D_{max}^2}{D_{min}^2}$ and $E[\frac{1}{d_c^2}]$ in $Sep(c)$ are influenced by the geometry of the cluster centers. $Sep(c)$ tends to increase with the number of clusters, c. $Sep(c)$ is more sensitive to the distance between clusters than $Scat(c)$. Consequently, the value of $V_{WSJ}(U, V, c)$ changes more signifficantly when we merge two overlapping clusters to one cluster or split one cluster into two separated clusters. The expression $\frac{Sep(c)}{Sep(C_{\max})}$ is utilized in order to scale the value of $Sep(c)$ into the same range as $Scat(c)$. A coefficient could be used to modulate the contribution of each of the two terms in $V_{WSJ}(U, V, c)$. For all the data sets used in this paper, the expression of $V_{WSJ}(U, V, c)$ given in Equation (9) has yielded more accurate results than any other index tested (see the next section). A cluster number which minimizes $V_{WSJ}(U, V, c)$ is considered to be the optimal value for the number of clusters present in the data.

4 The Comparison of the New Validity and the Others

To show the performance of the new validity index $V_{WSJ}(U, V, c)$, we report here the experimental results for three data sets, of which two come from the public domain, one is generated using Gaussian mixture distribution. In all the experiments, the fuzzier m in the algorithms was set at $m = 2$, the test for convergence in the basic FCM algorithm is performed using $\varepsilon = 0.001$, and the distance function $\| \cdot \|$ is defined as Euclidean distance. The Random method for cluster initialization is used in **Algo**. For determination of the number of clusters, the validity indices V_{PC},V_{PE},V_{Xie},V_{FS} and V_{Rez} were compared with V_{WSJ}.

c	V_{PC}	V_{PE}	V_{Xie}	V_{FS}	V_{Rez}	V_{WSJ}
2	0.90	0.18	0.04	-694	1.26	0.15
3	0.97*	0.08*	0.02*	-894	0.58*	0.03*
4	0.93	0.15	0.08	-976	1.23	0.05
5	0.92	0.18	0.05	-1190	1.24	0.04
6	0.92	0.19	0.09	-1328	2.03	0.10
7	0.92	0.19	0.16	-1426	3.23	0.24
8	0.93	0.18	0.35	-1496	5.54	0.65
9	0.94	0.16	0.32	-1620	5.80	0.71
10	0.96	0.12	0.40	-1721*	7.05	1.00

Table 1: Validity values on data set X30

4.1 Test on Data Set 1 (X_{30})

The first test here uses the data set X_{30} from [9]. It has 30 2-dimensional data vectors divided into 3 well separated clusters, each of which contains 10 data vectors. The results for $c = 2$ to 10 are shown in Table 1. For this simple data set, only *Fakuyama*'s index V_{FS} are not able to give the correct number of clusters (which is 3).

4.2 Test on Data Set 2 (IRIS)

The second data set is IRIS Data [10][11]. This is a biometric data set consisting of 150 measurements belonging to three flower varieties, generally known as the IRIS data set. Each class contains 50 observations, in which two variables (length and width of the petal and sepal) are measured. So the data are represented as a point in 4-dimensional measurement space. IRIS Data is one of the most commonly used benchmark data sets in data analysis. Figures 1 show 2D projections of these data. Of the three classes, two are overlapped.

Halgamuge and *Glesner* [12] have shown that a very good classification can be obtained by using only two features. In [7], *Rezaee* indicated that for the cluster validity index V_{Rez}, it is necessary to use one feature (petal length) to obtain the best number of classes, which is 3. In fact, as shown in Table 2, when using all four features, only the validity index V_{WSJ} was able to yield the correct number of classes. The optimal number of classes is given as 2 using other validity indices. The ability of the new validity index to compute the correct number of classes without using a feature selection procedure is a great advantage

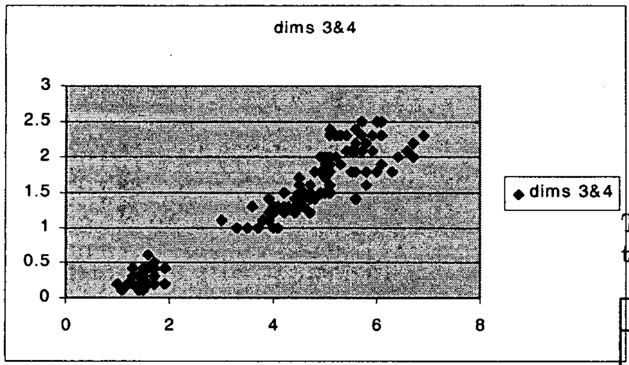

Figure 1: IRIS Data: dimensions 3 and 4

c	V_{PC}	V_{PE}	V_{Xie}	V_{FS}	V_{Rez}	V_{WSJ}
2	0.89*	0.20*	0.05*	-399	1.86*	0.18
3	0.79	0.39	0.14	-450	1.94	0.11*
4	0.70	0.56	0.20	-474	3.44	0.20
5	0.66	0.68	0.23	-542	3.39	0.20
6	0.61	0.61	0.81	-547*	4.67	0.36
7	0.55	0.92	0:42	-396	8.77	1.18
8	0.53	0.99	0.25	-393	6.61	0.73
9	0.51	1.07	0.38	-406	8.91	1.28
10	0.50	1.13	0.35	-400	8.12	1.03

Table 2: Validity values on IRIS data set

over the other indices. Our results for $c \in[2,10]$ are summarized in Table 2.

4.3 Test on Data Set 3 (5 clusters)

Data Set 3 was generated using a Gaussian mixture distribution. This data set is 3-dimensional and contains 5 clusters. Their means, variances and mixing coefficient are listed in Table 3. There are 50 data in each of the five clusters. In the data set, Cluster1 and Cluster3 strongly overlap with each other. Table 4 summarizes the results obtained using different validity indices. As can be seen, our validity index V_{WSJ} as well as the index V_{PC} were able to result in the correct number of clusters, which is 5. The indices V_{PE},V_{Xie} and V_{Rez}, which correctly predict the number of clusters in example 1, yield the numbers 2, 3 and 4 respectively in this case.

To conclude, the new validity index proposed here significantly improves the performance of the FCM-based algorithm in determining the number of clusters. In fact, based on results for the sets of experimental results reported above, only our validity index was able to predict the correct number of clusters in all cases. V_{PE} performs well when the clusters are well separated and the data set does not contain too many outliers (noise points). V_{PC}, V_{Xie},V_{FS} and V_{Rez} sometimes yield unpredictable results although they may perform well in difficult cases in which clusters overlap or there are a lot of noise points.

	Mean	*Varince*	*Coeff.*
C1	(-1 , -2 , -1)	(0.2, 0.3, 0.3)	0.2
C2	(1.5 , 1 , -1)	(0.3 ,0.2, 0.2)	0.2
C3	(-1 , -0.5 , -1)	(0.3, 0.4, 0.2)	0.2
C4	(2 , -1 , -1)	(0.2, 0.3, 0.3)	0.2
C5	(-1.5, -1.5, 1.7)	(0.3, 0.3, 0.4)	0.2

Table 3: Data Set 4 generated with a Gaussian mixture distribution

c	V_{PC}	V_{PE}	V_{Xie}	V_{FS}	V_{Rez}	V_{WSJ}
2	0.78	0.38*	0.16	-78	2.71	0.38
3	0.79	0.42	0.09	-605	1.62	0.18
4	0.818	0.40	0.08*	-835	1.39*	0.11
5	0.82*	0.41	0.09	-893	1.58	0.10*
6	0.77	0.51	0.35	-880	3.17	0.32
7	0.73	0.60	0.64	-946*	4.44	0.59
8	0.68	0.71	0.55	-890	4.61	0.66
9	0.63	0.81	0.79	-860	5.97	1.11
10	0.58	0.90	0.67	-710	5.66	1.02

Table 4: Validity values on Data Set 4

5 Conclusion and Discussion

The major contribution of this paper is a new efficient measure for validating clustering results and its application in determining the number of clusters. The new validity index has been tested on many public domain data, on generated test data and on data sets from real applications. The results have shown significant advantage of the new index over other indices, especially in the cases with overlapping clusters. Since the new validity V_{WSJ} is function of original data, cluster centers and cluster memberships, it can be used to measure the quality of the clustering results obtained by various clustering methods other than FCM. On the other hand, V_{WSJ} has allowed us to improve the FCM-based cluster number determination algorithm used in this paper by introducing a cluster splitting strategy into the algorithm. We are currently studying the effect of the distance function $d(x, y)$ and the fuzzifier (exponent parameter) m on the new validity index.

References

[1] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L.M. Le Cam & J. Neyman (Eds.)*, vol. 1, (Berkeley, CA), pp. 281–297, University of California Press, 1967.

[2] J. Bezdek, "Fuzzy mathematics in pattern classification." Ph.D. Dissertation, Cornell University, 1973.

[3] J. M. K. R. Krishnapuram, "A possibilistic approach to clustering," *Fuzzy System*, vol. 1, pp. 98–109, May 1993.

[4] A. Baraldi and P. Blonda, "A Survey of Fuzzy Clustering Algorithms for Pattern Recognition - Part I," IEEE *Transactions. Systems, Man, and Cybernetics*, vol. SMC-29, pp. 778–785, Dec 1999.

[5] X. Xie and G. Beni, "A Validity Measure for Fuzzy Clustering," IEEE *Transactions on Pattern Analysis and Machine Intelligence (*PAMI*)*, vol. 13, no. 8, pp. 841–847, 1991.

[6] Y. Fukuyama and M. Sugeno, "A New Method of Choosing the Number of Clusters for the Fuzzy C- means Method," in *Proceedings of 5th Fuzzy System Symposium*, pp. 247–250, 1989.

[7] B. L. M.R. Rezae and J. Reiber, "A new cluster validity index for the fuzzy c-means," *Pattern Recognition Letters*, vol. 19, pp. 237–246, 1998.

[8] Y. Man and I. Gath, "Detection and Separation of Ring-shaped Clusters Using Fuzzy Clustering," IEEE *Transactions on Pattern Analysis and Machine Intelligence (*PAMI*)*, vol. 16, pp. 855–861, Aug 1994.

[9] J. C. Bezdek, *Chapter F6: Pattern Recognition in Handbook of Fuzzy Computation*. IOP Publishing Ltd, 1998.

[10] E. Anderson, "The Iris of the Gaspé Peninsula," *Bulletin of American Iris Society*, vol. 59, pp. 2–5, 1935.

[11] N.R.Pal and J.C.Bezdek, "On Cluster Validity for the Fuzzy C-Means Model," *IEEE Trans. on Fuzzy Systems*, vol. 3, no. 3, pp. 370–390, 1995.

[12] S. Halgamuge and M. Glesner, "Neural networks in designing fuzzy systems for real world applications," *Fuzzy Sets and Systems*, vol. 65, no. 1, pp. 1–12, 1994.