Fuzzy Cluster Quality Index using Decision Theory

S.Revathy

Department of Information Technology,
Sathyabama University, Chennai,
Tamilnadu, India.
ramesh.revathy@gmail.com

B.Parvathavarthini,

Department of MCA
St.Joseph's College of Engineering, Chennai,
Tamilnadu, India.
parvathavarthini@gmail.com

Abstract

Clustering can be defined as the process of grouping physical or abstract objects into classes of similar objects. It's an unsupervised learning problem of organizing unlabeled objects into natural groups in such a way objects in the same group is more similar than objects in the different groups. Conventional clustering algorithms cannot handle uncertainty that exists in the real life experience. Fuzzy clustering handles incompleteness, vagueness in the data set efficiently. The goodness of clustering is measured in terms of cluster validity indices where the results of clustering are validated repeatedly for different cluster partitions to give the maximum efficiency i.e. to determine the optimal number of clusters. Especially, fuzzy clustering has been widely applied in a variety of areas and fuzzy cluster validation plays a very important role in fuzzy clustering. Since then Fuzzy clustering has been evaluated using various cluster validity indices. But primary indices have used geometric measures; this paper proposes decision theoretic measure for fuzzy clustering.

***Keywords*: Clustering, Fuzzy clustering, Fuzzy C Means Algorithm, Cluster Validity Index, Decision theory.**

1. Introduction

Clustering [1, 2, 3, 4, 5] is an unsupervised classification method when the only data available are unlabelled, and no structural information about it is available. In clustering, a set of patterns, usually vectors in a multidimensional space, are organized into coherent and contrasted groups in such a way that patterns in the same group are more similar than patterns in different groups . The purpose of any clustering technique is to obtain a partition matrix W(X) for the given data set X (consisting of, say, n patterns ,X = $\{x_1, x_2, \ldots, x_n\}$) so as to find a number, say c, of clusters (X_1,X_2, . . . , Xc). The partition matrix W(X) of size c × n may be represented as W = $[w_{ij}]_{c\times n}$, i = 1, . . . , c and j = 1, . . . , n, where $w_{ij}$is the membership of pattern xj to clusters Xi . In crisp partitioning, the following condition holds: w_{ij}= 1 if xj ∈ Xi, otherwise w_{ij} = 0.The purpose is to classify data set X such that

$Xi \neq \varphi$ for $i = 1, 2, \ldots, c$ (1)

$Xi \cap Xj = \varphi$ for $i = 1, 2, \ldots, c$, $j = 1, 2, \ldots, c$ and $i \neq j$, (2a)

$\bigcup_{i=1}^{c} Xi = X$ (2b)

In the case of fuzzy clustering, the purpose is to obtain an appropriate partition matrix W = $[w_{ij}]$c×n, where $w_{ij} \in [0, 1]$, such that w_{ij} denotes the grade of membership of the *j*th element to the *i*th cluster. In fuzzy partitioning of the data, the following conditions hold:

$0<\sum_{i=1}^{c} w_{ij}<1$ for j=1,2,……..,c (3)

$\sum_{t=1}^{c} w_{it}=1$ for i=1,2,……..,n (4)

$\sum_{j=1}^{c}\sum_{k=1}^{n} w_{kj} = n$ (5)

Fuzzy clustering has been applied in a variety of key areas. However, a main difficulty in fuzzy clustering is that the number of clusters c must be specified prior to clustering. Selections of a different number of initial clusters result in different clustering partitions. Thus, it is necessary to validate each of the fuzzy clusters once

they are found. Many validation criteria have been proposed. This paper put forwards decision theoretic measure for fuzzy clustering.

The remnant of this paper is organized as follows. In the next section we review the Fuzzy C-Means clustering algorithm. Section 2 describes Basic Fuzzy C Means Clustering Algorithm. Section3 provides cluster validity problem descriptions. A number of fuzzy cluster validity indices available in the literature are presented in Section 4. Section 5 gives the proposed cluster quality index. Conclusions are drawn in Section 6.

2 FUZZY C MEANS CLUSTERING ALGORITHM

This is a fuzzification of the *c*-means algorithm, proposed by Bezdek [6]. It partitions a set of *N* patterns $\{X_k\}$ into *c* clusters by minimizing the objective function $J = \sum_{k=1}^{N}\sum_{i=1}^{c}(w_{ik})^m\|X_k - V_i\|^2$, where $1 \le m < \infty$ is the fuzzifier, V_i is the *i*th cluster center, $w_{ik} \in [0, 1]$ is the membership of the *k*th pattern to it, and $\|.\|$ is the distance, such that,

$$v_i = \frac{\sum_{k=1}^{N}(w_{ik})^m X_k}{\sum_{k=1}^{N}(w_{ik})^m} \quad (6)$$

$$w_{ik} = \frac{1}{\sum_{i=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \ \forall i, \quad (7)$$

With $d_{ik} = \|X_k - V_i\|^2$, subject to $\sum_{i=1}^{c} w_{ik} = 1$, $\forall k$, and $0 < \sum_{k=1}^{N} w_{ik} < N$, $\forall i$.. The algorithm proceeds as in *c* - means, along with the incorporation of membership.

The FCM algorithm is executed in the following steps:

Step 1: Given a cluster number c, a value of m, initialize memberships w_{ij} of xj belonging to cluster i such that

$\sum_{j=1}^{c} w_{ij}=1$

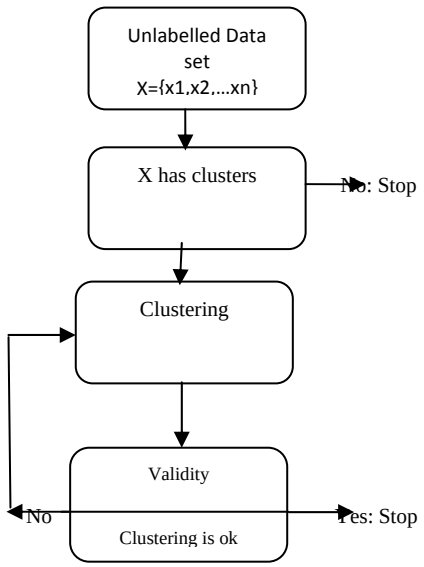
Step 2: Calculate the fuzzy cluster prototype vi for i = 1, 2, . . .c using Eq. (6).

Step 3: Use Eq. (7) to update the fuzzy membership w_{ij}.

Step 4: If the improvement in value of objective function is less than a certain threshold (Ξ), then halt; otherwise go to step 2.

2. CLUSTER VALIDITY DESCRIPTIONS

Clustering is considered as an unsupervised process since there are no predefined classes. This would show what kind of desirable relations should be valid among the data. The final cluster partitions of a data set require some sort of evaluation in most applications [7]. For example questions like “how many cluster partitions can happen in a data set?”, “does the resulting clustering structure is rightly applicable to our data set?”, “Is there a more desirable clustering for our data set?” To overcome the above issues, validation should be done at the end of clustering process. Some of the popular fuzzy cluster validity indices have been discussed in the successive literature survey.



4. OVERVIEW OF FUZZY CLUSTER VALIDITY INDICES

A validity index is a function which assigns output of the clustering algorithm to a value which is intended to measure the quality of the clustering provided by the clustering algorithm. Since most of the fuzzy clustering methods, such as FCM, need to pre-assume the number c of clusters, a validity index for finding an optimal c, denoted c* , which can completely describe the data structure, becomes the most research oriented topic in cluster validity.

Two categories of fuzzy validity indices are discussed. The first category uses only the membership values, w_{ij}, of a fuzzy partition of data. The second involves both the W matrix and the dataset.

4.1Validity Indices involving only membership values:

4.1.1. Partition Coefficient:

Bezdek proposed in [8] the partition coefficient which was defined as

$$PC=\frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{n_c} {w_{kl}}^2 \quad (8)$$

4.1.2..Partition Entropy Coefficient:

Bezdek suggested the partition entropy (PE) [8–10] that was defined as,

$$PE=\frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{n_c} w_{kl}\log_a w_{kl} \quad (9)$$

The best partition is obtained when the value for PC has a maximum or the value PE has a minimum, for a certain number of clusters.

4.1.3. Modified Partition Coefficient:

Both PC and PE possess monotonic evolution tendency with c. Modified form of PC index proposed by Dave [11] can reduce the monotonic tendency and was defined as

$$MPC=1-\frac{c}{c-1}(1-PC) \quad (10)$$

The above indexes used only fuzzy memberships. These indices have some downsides.

(1) Their monotonous dependency on the number of clusters.

(2) Their sensitivity to the fuzzifier, m.

(3) The lack of direct connection to the geometry of the data, since they do not use the data itself.

4.2. Validity Indices involve both membership values and the data set

4.2.1. Fukuyama and Sugeno (FS) Index

A quality Index proposed by Fukuyama and Sugeno (FS) [9] was defined by

$$FS=\sum_{i=1}^{c}\sum_{j=1}^{n} w_{ij}^m \left\|x_j - v_i\right\|^2 - \sum_{i=1}^{c}\sum_{j=1}^{n} w_{ij}^m \left\|v_i - \bar{v}\right\|^2 \quad (11)$$

Where $\bar{v}=\sum_{i=1}^{c} v_i/c$

4.2.2. Xie-Beni Index

Xie and Beni proposed a validity measure [10] and it was defined as

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} w_{ij}^2\left\|x_j-v_i\right\|2}{\text{n x mini}\neq j(\left\|v_i-v_j\right\|)} \quad (12)$$

The above indices have the common objective of finding a good estimate of cluster number c so that each one of the c clusters is compact or/and separated from the other clusters. Since these measures are geometric measures, next section outlines the proposed decision theoretic measure.

5.PROPOSED FUZZY CLUSTER QUALITY INDEX

5.1 DECISION THEORY

Decision-theoretic framework [14] has been helpful in providing a better understanding of the classification process. The decision-theoretic rough set model considers various classes of loss functions. It is

possible to construct a cluster validity measure by considering various loss functions based on decision theory. The following section derives risk measure for fuzzy clustering structure.

5.2 Risk for assigning objects under Clusters

According to Bayesian decision theory risk for an object can be calculated by multiplying loss and probability.

$$R_{\overrightarrow{x_l}}(FC)=\sum_{i=1..k}\lambda_{\overrightarrow{x_l}}(c_i) \ * \ P(\overrightarrow{c_l}|\overrightarrow{x_l}) \quad (13)$$

Loss for assigning an object x_l to c_i is given by

$$\lambda_{\overrightarrow{x_l}}(c_i)=1\text{-}w_{li} \quad (14)$$

Probability of cluster c_i with given object x_l is

$$P(\overrightarrow{c_l}|\overrightarrow{x_l})=\frac{sim(\overrightarrow{x_l},\overrightarrow{c_l})}{\sum_{1\le j\le k} sim(\overrightarrow{x_l},\overrightarrow{c_l})} \quad (15)$$

In Fuzzy Clustering membership values are used to specify similarity. Hence Risk for fuzzy clustering structure can be defined as,

$$R(FC)=\sum_{i=1..n} R_{\overrightarrow{x_l}}(FC) \quad (16)$$

5.3. Proposed Algorithm

Input: Fuzzy Clusters derived using FCM

Output: Optimal value for number of clusters

Method used: Bayesian Decision Theory

Step1: Calculate Loss value for each object.

Step 2: Calculate conditional probability for each object

Step 3: Calculate Risk for assigning objects under multiple clusters.

Step 4: Calculate Total Risk.

Step 5: Repeat step1 to step 4 for different number of clusters.

Step 6: Select the cluster number with minimum risk.

VI. EXPERIMENTAL RESULTS

6.1. Synthetic Data.

Table 1:Data Set

Object (2 Dimension)	
1.7	1.7
2.1	1.8
1.6	2.1
3.5	2.7
3.5	5.1
3.1	5.2
3.3	4.7
7.7	4.6
7.8	5.2
8.2	4.7

Table 2 shows loss values for each object which has been calculated using equation 14.Table 3 shows probability of assigning each object to specific cluster.

Table 2: Loss values When c=3(for six objects)

Cluster1	0.0168	0.0020	0.0258	0.3148	0.9950	0.9910
Cluster2	0.9964	0.9996	0.9951	0.9301	0.9969	0.9954
Cluster3	0.9868	0.9984	0.9791	0.7551	0.0082	0.0136

Table 3: Probability values when c=3 (for six objects)

Cluster1	0.9832	0.9980	0.9742	0.6852	0.0050	0.0064
Cluster2	0.0036	0.0004	0.0049	0.0699	0.0031	0.0028
Cluster3	0,0132	0.0016	0.0209	0.2449	0.9918	0.9908

Table 4 shows risk values .The results show the risk to be minimum when the objects are grouped into 3clusters. Hence we can prove that this measure provides exact number of clusters for the synthetic data set.

Table 4: Risk Values

Number of clusters	Risk Values
2	1.2198
3	0.3851
4	0.7181
5	0.8878
6	1.3484

6.2. Iris Data

Iris data is a typical benchmark consisting of 150 samples of three classes of the iris flower. There are 50 samples in each class which are expressed in terms of the four features viz., sepal length, sepal width, petal length, petal width. This data has been clustered using FCM. The results have been evaluated using the proposed measure. The following table shows the risk values. Here also we proved that the proposed measure provided optimal number of clusters.

Table 5: Risk Values for Iris Data

Number of clusters	Risk Values
2	32.4846
3	16.1794
4	44.0385
5	56.3832
6	60.7622

VI. **CONCLUSION**

In fuzzy clustering, the role of a validity index is very important. It helps to evaluate the correct number of clusters present in a data set. Many cluster validity indices have been used to evaluate quality of fuzzy clustering. Since those measures are geometric measures, this paper put forwards decision theoretic measure for evaluating fuzzy clustering results. The proposed measure has been tested using syntactic data as well as the real

time iris data set. The results show that the proposed measure shows appropriate number of clusters for both data sets. This measure is most useful in business oriented data mining.

REFERENCES

[1] M.R. Anderberg, Cluster Analysis for Application, Academic Press, NewYork, 1973.
[2] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, London, 1982.
[3] J.A. Hartigan, Clustering Algorithms,Wiley, NewYork, 1975
[4] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[5] Y.G. Tang, F.C. Sun, Z.Q. Sun, Improved validation index for fuzzy clustering, in: American Control Conf., June 8–10, 2005, Portland, OR,USA.
[6] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.
[7] Rezaee, R, Lelieveldt, B.P.F., Reiber, J.H.C. "A new cluster validity index for the fuzzy c-mean", Pattern Recognition Letters, 19, pp. 237-246, 1998.
[8] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NewYork, 1981
[9] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, in: Proc. Fifth Fuzzy Systems, Symp., 1989, pp. 247–250.
[10] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 841–847.
[11] R.N. Dave, Validating fuzzy partition obtained through c-shells clustering, Pattern Recognition Lett. 17 (1996) 613–623.
[12] Horng-Lin Shieh1, Po-Lun Chang, A New Robust Validity Index for Fuzzy Clustering Algorithm, 978-1-4244-8503-1/10/$26.00 ©2010 IEEE
[13] WeinaWanga,Yunjie Zhang, On fuzzy cluster validity indices, 0165-0114/$ - see front matter © 2007 Elsevier
[14] P. Lingras, M. Chen, and D.Q. Miao, “Rough Multi-Category Decision Theoretic Framework,” Rough Sets and Knowledge Technology, pp. 676-683, Springer, 2008