Volume : 4 | Issue : 11 | November 2015 • ISSN No 2277 - 8179

Research Paper

Internal Versus External Validity Indices for Fuzzy Clustring

Computer Science

KEYWORDS : Data Mining, Fuzzy Clustering, Cluster validity, Internal Validity Index, External Validity Index

Dr.S.Revathy. Department of Information technology Sathyabama University, Chennai, India

ABSTRACT

Cluster analysis aims at identifying groups of similar objects, and helps to discover distribution of patterns and interesting correlations in large data sets. A common approach for evaluation of clustering results is to use validity indices. Clustering validity approaches can use three criteria: External criteria (evaluate the result with respect to a pre-specified structure), internal criteria (evaluate the result with respect to information intrinsic to the data alone), Relative criteria (evaluate the result by comparing it with result obtained through other clustering algorithm). Different types of indices are used to solve different types of problems and index selection depends on the kind of available information. This paper shows a comparison between external and internal fuzzy cluster validity indexes. Results obtained in this study indicate that internal indexes are more accurate in group determining in a given clustering structure. Five internal indexes were used in this study: PC, PE, MPC, FS, XB and Three external indexes (F-measure, Entropy, and Purity). The clusters that were used were obtained through Fuzzy c means clustering algorithm.

I.INTRODUCTION

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters (Guha et al., 1998).For example, consider a super market database records containing items purchased by customers. A clustering procedure could group the customers in such a way that customers with similar buying patterns are in the same cluster. The conventional (hard) Clustering methods restrict each point of the data set to exactly one cluster. Since Zadeh [13] proposed fuzzy sets that produced the idea of partial membership of belonging described by a membership function. Hence In fuzzy clustering data point belongs to more than one cluster with different degree of membership values range in the interval [0,1].The most popular and commonly used fuzzy clustering algorithm is fuzzy-c-means clustering algorithm. Once the partition is obtained by a clustering method, the validity Index can help us to validate whether it accurately presents the data structure or not. This is called cluster validity. Indices of cluster validity can be divided into two categories: external and internal. External validity indices measure how well the clustering results match with the prior knowledge of the data. Internal validity indices measure the clustering results based on the information available in the dataset. In this paper we present the comparative study of these two approaches.

The rest of the paper is organized as follows: section 2 presents Fuzzy C Means Clustering Algorithm. Section 3 offers a various internal and external validity indices. Section 4 presents the study comparative; results obtained and discuss some findings from these results. Finally, we conclude by briefly showing our contributions and further works.

II. CLUSTERING ALGORITHM

A. Fuzzy C Means Clustering Algorithm

Main objective of fuzzy c-means algorithm is to minimize:

$$J_m = \sum_{i-1}^{N}\sum_{j-1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \qquad [1]$$

where,

||xi – vj||' is the Euclidean distance between ith data and jth cluster center. Algorithmic steps for Fuzzy c-means clustering:

Let X = {x1, x2, x3 ..., xn} be the set of data points and V = {v1, v2, v3 ..., vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the fuzzy membership 'μij' using:

3) Compute the fuzzy centers 'vj' using:

$$u_{ij} = \frac{1}{\sum_{k-1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \qquad [2]$$

$$c_j = \frac{\sum_{i-1}^{N} u_{ij}^m x_i}{\sum_{i-1}^{N} u_{ij}^m} \qquad [3]$$

4) Repeat step 2) and 3) until the minimum 'J' value is achieved or ||U (k+1) – U (k)|| < β.

where,

β is the termination criterion between [0, 1].

U = (μij)n*c' is the fuzzy membership matrix.

'J' is the objective function.

K is the iteration step

III.ANALYSIS OF INDICES

In this section, we offer an overview of internal and external validity indexes of fuzzy clustering that were used in our study.

A. Internal Validity Indices

Partition Coefficient(PC)

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

$$PC=\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{n_c} uij^2 \qquad [4]$$

The PC index indicates the average relative amount of membership sharing done between pairs of fuzzy subsets in U, by combining into a single number, the average contents of pairs of fuzzy algebraic products. The index values range in [1/c, 1], where c is the number of clusters.

Partition Entropy(PE):

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

$$PE=\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{n_c} uij \log_a uij \qquad [5]$$

The PE index is a scalar measure of the amount of fuzziness in a given U. According to [10], the limitation of the PE can be attributed to its apparent monotonicity and to an extent, to the heuristic nature of the rationale underlying its formulation.

Modified Partition Coefficient(MPC)

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

$$MPC=1-\frac{c}{c-1}(1-PC) \qquad [6]$$

Fukuyama and Sugeno (FS) Index

A validity function proposed by Fukuyama and Sugeno (FS) [9] was defined by

$$F=\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\| x_j - v_i \right\|^2 - \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\| v_i - \bar{v} \right\|^2 \qquad [7]$$

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

The first term is Jm (u, v), which combines the fuzziness in U with the geometrical compactness of the representation of X via the c prototypes V. The second term, Km (u, v), combines the fuzziness in each row of U with the distance from the ith prototype to the grand mean of the data.

.Xie-Beni Index (XB)

Xie and Beni proposed a validity function [10] and it is defined as

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \left\| x_j - v_i \right\|^2}{n \times \min i \neq j \left(\left\| v_i - v_j \right\| \right)} \qquad [8]$$

This index XB focused on two properties: compactness and separation. In this equation, the numerator indicates the compactness of the fuzzy partition, while the denominator indicates the strength of the separation between clusters.

B.External Validity Indices

F-measure

It combines the precision and recall concepts from information retrieval. We then calculate the recall and precision of that cluster for each cluster as:

$$\text{Recall (i,j)} = \frac{n_{ij}}{n_i} \qquad [9]$$

$$\text{Precision (i,j)} = \frac{n_{ij}}{n_j} \qquad [10]$$

$$\text{F-measure} = \frac{2 * \text{Recall(i,j)} * \text{Precision(i,j)}}{\text{Recall(i,j)} + \text{Precision(i,j)}} \qquad [11]$$

Purity

Purity[14] is very similar to entropy. We calculate the purity of a set of clusters. First, we cancel the purity in each cluster. For each cluster, we have the

purity $p_{j=\frac{1}{n_j}Max_i\left(n_j^i\right)}$ is the number of objects in j with class label i. In other words, Pj is a fraction of the overall cluster size that the largest class of objects assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and given as:

$$\text{Purity}=\sum_{j=1}^{m}\frac{n_j}{n}\text{P}_\text{j} \qquad [12]$$

Where nj is the size of cluster j, m is the number of clusters, and n is the total number of objects.

Entropy

Entropy[10] measures the purity of the clusters class labels. Thus, if all clusters consist of objects with only a single class label, the entropy is 0. However, as the class labels of objects in a cluster become more varied, the entropy increases. To compute the entropy of a dataset, we need to calculate the class distribution of the objects in each cluster as follows:

$$\text{Entropy}=\sum_{j=1}^{m}\frac{n_j}{n}\text{E}_\text{j} \qquad [13]$$

Where nj is the size of cluster j, m is the number of clusters, and n is the total number of objects.

IV.COMPARITIVE STUDY

In this section, we show experimentally tested results using the Fuzzy C Means clustering algorithm. We used 8 synthetic data sets. These data sets were used by WeinaWanga, Yunjie Zhanga[12]. To find the best partition, we have used the Fuzzy C Means algorithm with its input parameters (K) ranging between 2 and 8. Table 1 presents a summary of the tests carried out in the datasets that were clustered with the FCM algorithm, using internal clustering validity indices; from there we can see that PC,XB indexes identified the correct number of groups in 7 of the trials, and FS index gave wrong results in datasets 2, 4 and 7. Additionally, none of the indexes identified the correct number of clusters of the datasets 2 and 8. Table 2 presents a summary of the tests carried out in the datasets that were clustered with the FCM algorithm, using external cluster validity indexes; from where we can see that the F-measure index correctly identified the number of groups in all trials and Entropy only gave correct results in Dataset1,3. In conclusion, from these tests it can be said that internal validity indexes are more accurate in the identification of correct group numbers in clusters formed with the FCM algorithm.

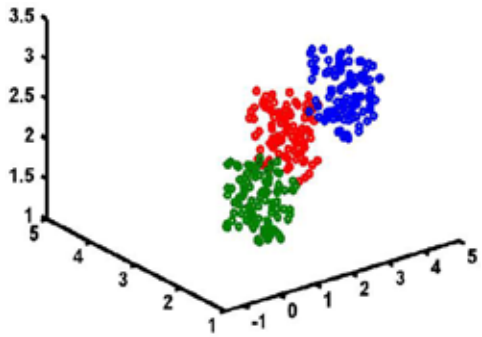

Fig.1.Data Set-1

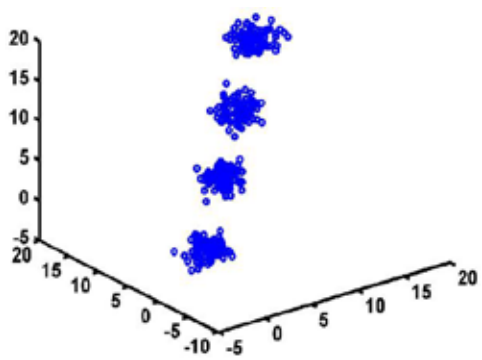

Fig.2.Data Set-2

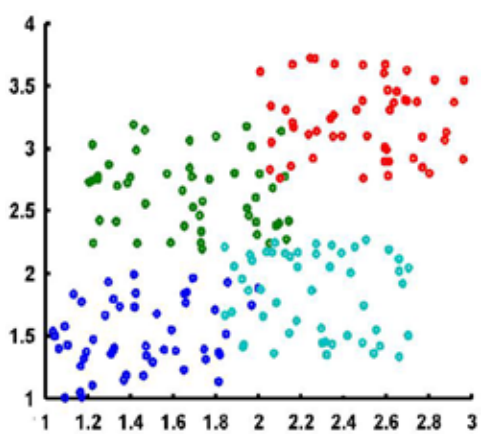

Fig.3.Data Set-3

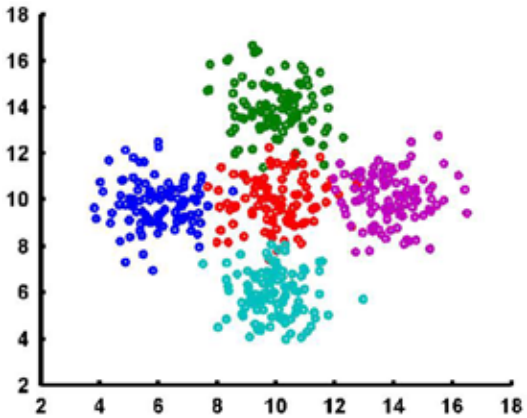

Fig.4.Data Set-4

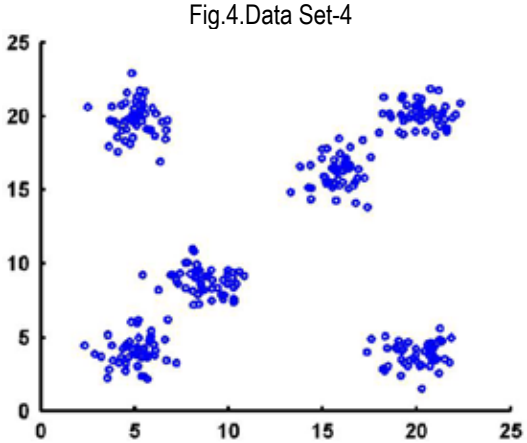

Fig.5.Data Set-5

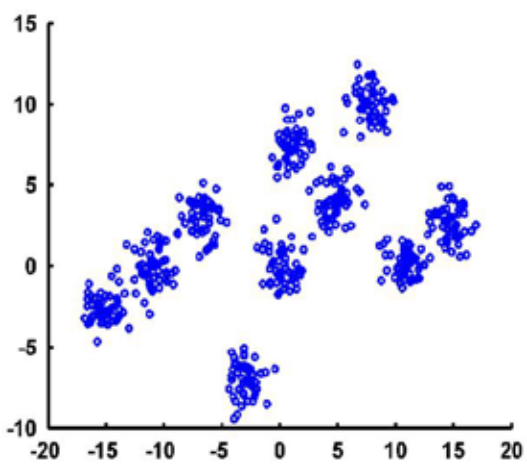

Fig.6.Data Set-6

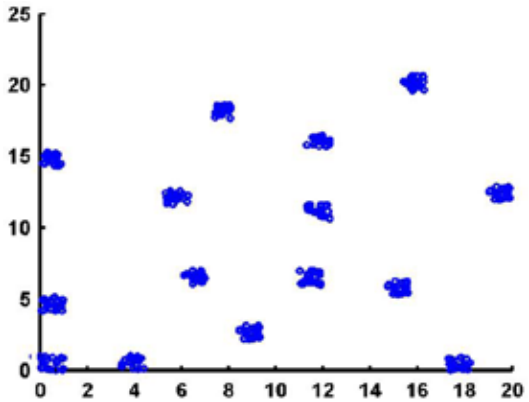

Fig.7.Data Set-7

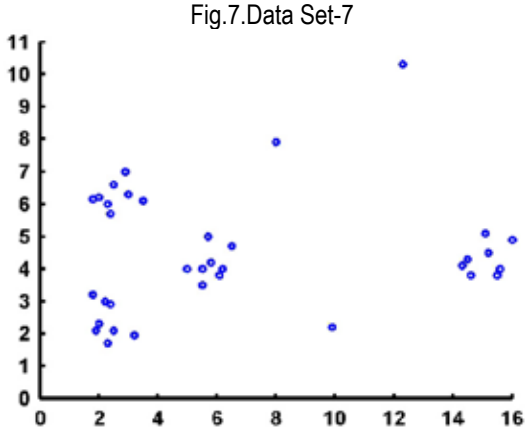

Fig.5.Data Set-8

Table 1. Overview of the results obtained with internal validity indexes applied to Fuzzy C Means Clustering algorithm.

DataSet	PC	PE	MPC	FS	XB
Dataset-1	√	√	√	√	√
Dataset-2	√		√		√
Dataset-3	√	√		√	√
Dataset-4	√	√	√		
Dataset-5	√	√	√	√	√
Dataset-6	√			√	√
Dataset-7	√	√			√
Dataset-8		√	√	√	√
Total	7	6	5	5	7

Table 2. Overview of the results obtained with external validity indexes applied to Fuzzy C Means clustering algorithm.

DataSet	F-Measure	Purity	Entropy
Dataset-1	√	√	√
Dataset-2	√	√	
Dataset-3	√		√
Dataset-4	√		
Dataset-5	√		
Dataset-6	√		
Dataset-7	√	√	
Dataset-8	√	√	
Total	7	4	2

VI. CONCLUSION

This paper presents a comparison between two clustering validity index approaches, internal and external; carrying out analysis of three external indices and five internal indices. Here 8 datasets were used, which were clustered using FCM algorithm. Each dataset was clustered with different K values (K= 1,….,8 groups). Out of 40 (8*5) cases where the results of the FCM algorithm using internal indices, correct group numbers were obtained 86% of the time, and in 51.9% when external indices(8*4) were used. When clusters of the FCM algorithm were clustered using internal indices, 76.9% of accuracy was obtained; and61.5% with external indices. From which we can infer that, internal indices are more precise in real group number determination than external indices

REFERENCE

[1] Halkidi M., Vazirgiannis, M. Quality scheme assessment in the clustering process. In Proc. PKDD (Principles and Practice of Knowledge in databases). Lyon, France. Lecture Notes in Artificial Intelligence. Spring –Verlag Gmbh, vol.1910, 2000, pp. 265-279. [2] Chow C.H, Su M.C and Lai Eugene. Symmetry as A new measure for Cluster Validity. 2 th. WSEAS Int.Conf. scientific Computation and Soft Computing, Crete, Greece, 2002, pp. 209-213 [3] Su M.C, Chow C.H. A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry. IEEETrans. Pattern Anal. And Machine Intelligence, vol. 23. No.6, 2001,pp.674-680. [4] Chow C.H, Su M.C and Lai Eugene. A new Validity Measure for Clusters with Different Densities. Pattern Anal. Applications, 7,2004, pp.2005-2020. [5] Theodoridis, S., Koutroubas, K. Pattern Recognition, Academic Press, USA, 1999. [6] Volker Roth, Tilman Lange, Mikio Braun, and Joachin Buhmann.A Resampling Approach to Cluster Validation, Proceeding in Computational Statistics COMPSTAT. Physika Verlag, 2002, pp.123-128 [7] Athena Vakali, Jaroslav Pokorný and Theodore Dalamagas. An Overview of Web Data Clustering Practices, Lecture Notes Computer Science, Vol. 3268, 2005, pp.597-606. [8] M.J.L. Hoon, S. Imoto, J. Nolan and S. Miyano. Open source clustering software. Bioinformatics, Vol. 20 No. 9, 2004, pp. 1453-1454. [9] Guha Sudipto, Rastogi Rajeev, Shim Kyuseok. CURE: An Efficient Clustering Algorithm for Large DataBases. In Proceedings of the CAM SIGMOD Conference on Magnagement of Data, Seatle,Washington, U.S.,01-04 Jun., 1998, pp. 73-83. [10] B. MacQueen. Some Methods for classification and Analysis of Multivariable Observations, Proceeding of 5th Berkeley on Mathematical Statistics and Probability, University of CaliforniaPress, 1967, pp. 281-297. [11] Bernd Drewes. Some Industrial Applications of Text Mining,Knowledge Mining, Springer Berlin, Vol. 185, 2005, pp. 233-238. [12] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis.Cluster Validity methods: Part I, SIGMOD Record, Vol.31(2),2002, pp. 40-45. [13] Larsen and C. Aone. Fast and effective text mining using lineartime document clustering. In Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 1999, pp. 16-22. [14] Strehl A and Ghosh J.Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research (JMLR) 3, 2002, pp.583–617. [15] Shannon C. E. A Mathematical Theory of Communication, The Bell System Technical ournal, 1948, pp. 379 – 423. [16] Raftery A. A note on Bayes factors for log-linear contingency table models with vague prior information. Journal of the Royal Statistical Society. 48(2), 1986, pp. 249-250. [17] Pacual D., Pla F., Sánchez J.S. Cluster validation using information stability measures, Pattern Recognition Letters 31, 2010, pp.454-461. [18] Rendon L. Eréndira, Garcia Rene, Abundez Itzel, GutierrezCitlallih, et. al. Niva:A Robust Cluster Validity. 2 th. WSEASInt.Conf. Scientific Computation and Soft Computing, Crete,Greece, 2002, pp. 209-213. [19] Legány C., Juhász S. and A. Babos. Cluster Validity Measurement Techniques. Proceeding of the 5th. WSEAS Int.Conf. on Artificial,Knowledge Engineering and Data bases, Madrid, Spain, February15-17,2006, pp. 388-393. [20] Kovács F. and R. Ivancsy. Cluster Validity Measurement forarbitrary Shaped clustering. Proceeding of the 5th. WSEAS Int.Conf. on Artificial, Knowledge Engineering and Data bases,Madrid, Spain, February 15-17,2006, pp. 372-377.