Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi
Journal of Applied Sciences of Mehmet Akif Ersoy University

Effect of Parameter Selection on Fuzzy Clustering

Ozer Ozdemir[1*], Asli Kaya[2]

[1]Dr. Öğr. Üyesi, Anadolu University, Faculty of Science, Department of Statistics

[2]Anadolu University, Faculty of Science, Department of Statistics

Geliş Tarihi/Received: 02.11.2017
Kabul Tarihi/Accepted: 15.11.2017 Araştırma Makalesi/Research Article

ABSTRACT

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. Cluster analysis seeks to partition given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. Clustering can be performed in hard or fuzzy mode. One of the important conditions in order to reach accurate results in clustering analysis is to determine the initial parameters. In many studies, researchers do not have prior information about the number of clusters. Clustering algorithms in general need the number of clusters as a prior, which is mostly hard for domain expert to estimate. In this work, in order to overcome this problem, cluster validity indices in literature were reviewed and these indices were used in genetic data set. The result was simply analyzed and according to the analysis, validity indices do not always discover the optimal number of clusters.

Keywords: Clustering, Fuzzy clustering, Validity index.

Bulanık Kümeleme Analizinde Parametre Seçiminin Etkisi

ÖZET

Kümeleme, grupları keşfetmek ve veri setinin altında yatan ilginç dağılımları ve kalıpları saptamak için veri madenciliği işleminde en yararlı yöntemlerden biridir. Kümeleme analizi verilen bir veri kümesini belirlenmiş özelliklere göre gruplara parçalama çabasıdır. Böylece bir grup içindeki veri noktaları, farklı gruptaki noktalara göre birbirine daha çok benzerdir.

* Sorumlu yazar/Corresponding author
E-mail/e-ileti: ozerozdemir@anadolu.edu.tr

Kümeleme, sert veya bulanık modda gerçekleştirilebilir. Bulanık kümeleme analizinde sağlıklı ve anlamlı sonuçlara ulaşabilmek için önemli durum başlangıç parametrelerin belirlenmesidir. Kümeleme analizlerinde genel olarak başlangıç küme sayısına ihtiyaç vardır ancak bir veri kümesi için uygun küme sayısının önceden tahmin edilmesi alanın uzmanı için zor bir işlemdir. Bu çalışmada bu sorunun üstesinden gelebilmek için literatürdeki geçerlilik indeksleri araştırılmış ve genetik veri seti üzerinde uygulanmıştır. Sonuçlar basitçe analiz edilmiş olup bu indekslerin de her zaman en uygun sonuç vermediği görülmüştür.

Anahtar kelimeler: Kümeleme, Bulanık kümeleme, Geçerlilik indeksleri

1. INTRODUCTION

Fuzzy clustering algorithms require defining the number of clusters (c), but it is not always possible to know this number in beforehand. Selections of a different number of initial clusters result in different clustering partitions. Therefore, it is necessary to validate each of the fuzzy c- partitions. The process of selection of optimal cluster number is called “Validity index of clustering”. During the last years, many validity indexes have been proposed (Pal and Bezdek, 1995: 370).

There are many studies on validity indices for fuzzy clustering which exist in the literature. In Xie and Beni’s work (Xie and Beni, 1991), they have generalized a new validity function; separation index. Then they have applied this validity function to color image segmentation for IC ring defect detection. As a result of this work, they said separation index only measures compact and separate clusters, as defined (Xie and Beni, 1991: 841).

Pal and Bezdek (Pal and Bezdek, 1995), examined the role a subtle but important parameter- the fuzzifier exponent m- plays in determining the validity of fuzzy partitions. The functional considered are the partition coefficient and entropy indexes of Bezdek, the Xie Beni, and extended Xie-Beni indexes, and the Fukuyama-Sugeno index (Fukuyama and Sugeno, 1989). Analysis indicated that Fukuyama-Sugeno index is sensitive to both high and low values of m. On the other and, Xie-Beni index provided the best response over a wide range of choices for the number of clusters (Pal and Bezdek, 1995: 370).

Pakhira et al. (Pakhira et al., 2004), proposed a cluster validity index which can work for both crisp and fuzzy clustering in their work. They have provided a detailed mathematical analysis of the index in support of the work-ability of the proposed index (Pakhira et al., 2004: 481).

Melegy et al. (Melegy et al., 2007) have surveyed 16 well-known such indexes and made a comprehensive comparison between these indexes for the task of image segmentation. They also proposed a new index based on Akaike's information criterion (AIC). In addition, a new index for the same task based on cross-validation has been proposed. All 18 indexes have been assessed on 2D and 3D data corrupted with noise of varying levels (Melegy et al., 2007: 5).

Saad and Alımı (Saad and Alımı, 2012), reviewed several validity indexes and then proposed a new validity index, called Modified Partition Coefficient and Exponential Separation, which is developed to obtain optimal partition. Moreover, they have conducted extensive comparisons of the mentioned indices in conjunction with the FCM algorithm on a number of widely used data sets. These results prove that our new index (MPCAES) provides the majority of cases the value of the desired classes (Saad and Alımı, 2012).

Zanaty and Afifi (Zanaty and Afifi, 2013), in their study, an alternative reliable validity index algorithm has proposed in order to improve the image clustering. The proposed method had been tested with discrete image example to show the applicability of this method. Also, they had compared it with the results obtained from cluster validity indexes such as PC, CE, and XB (Zanaty and Afifi, 2013: 38).

The proposed method is applied to two simulations and one real life data. In the results obtained for the simulation data, the criteria which are PC, CE, XB and the proposed method is appointed the appropriate number of clusters correctly. As a result of the applications, it can be seen that the most appropriate number of clusters can be appointed in fuzzy clustering with the proposed method (Zanaty and Afifi, 2013: 38).

This paper presents of fuzzy cluster validity indices available in the literature, classified in two important types for c-means: one is based on the fuzzy partition of the dataset and the other is based on the geometric structure and membership values. These indices were used in well-known two data sets with fuzzy c-means algorithms and changeable fuzzifier parameter.

2. THE FUZZY C-MEANS CLUSTERING ALGORITHM

Fuzzy C-Means (FCM) unsupervised classification algorithm dates back to through 1973 (Bezdek, 1973).

Fuzzy c-means allows data points to be assigned into more than one cluster each data point has a degree of membership of belonging to each cluster (Hartigan, 1975). FCM attempts to find the most characteristic point in each cluster, which can be considered as the "centroid" of the cluster and, then, the grade of membership for each object in the clusters. Such aim is achieved by minimizing the following objective function:

$$J(u,v) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m} \left\| x_j - v_i \right\|^2 \quad (1)$$

where *n* is the total number of patterns in a given data set and *c* is the number of cluster $X = \{x_1, x_2, \ldots, x_n\} \subset R^s$ and $V = \{v_1, \ldots, v_c\} \subset R^s$ are the feature data and cluster centroids; and *U* = [*uij*]*c*×*n* is a fuzzy partition matrix composed of the membership grade of pattern x_j to each cluster *i*. $\left\| x_j - v_i \right\|^2$ is the Euclidean norm between x_j and v_i. The weighting exponent *m* is called the being effective on the clustering performance of FCM.

The cluster centroids and the respective membership functions that solve the constrained optimization problem in Equation (1) are,

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}, \quad 1 \le i \le c \quad (2)$$

$$u_{ij} = \frac{1}{\left[\sum_{k=1}^{c} \left(\frac{\left\| x_j - v_i \right\|}{\left\| x_j - v_k \right\|} \right)^{1/(m-1)} \right]}, \quad 1 \le i \le c, 1 \le j \le n \quad (3)$$

These equations are obtained from iterative optimization process. The FCM algorithm is executed in the following steps:

Step 1: Given a pre-selected number of cluster *c*, a chosen value of *m*, initialize memberships u_{ij} of x_j belonging to cluster *i* such that

$$\sum_{i=1}^{c} u_{ij} = 1 \quad (4)$$

Step 2: Calculate the fuzzy cluster centroid v_i *for* $i = 1, 2, \ldots, c$ using Equation (2).

Step 3: Update the membership u_{ij} using Equation (3).

Step 4: If the improvement in *J(U, V)* is less than a certain threshold *(ε)*, then halt; otherwise go to step 2

3. SOME VALIDITY INDICES

The problem for finding an optimal c is usually called cluster validity. Once the partition is obtained by a clustering method, the validity function can help us to validate whether it accurately presents the structure of the data set or not. The problems of deciding the number of clusters better fitting a data set as well as the evaluation of the clustering results has been subject of several research efforts.

When the data are in the two dimensional space, the number of clusters can be decided upon by commenting on the cluster results visually. However, as the dimension of the problem increases in space, visually gets harder and there becomes a need for validity indexes.

As a result, two criteria can be mentioned for value clusters and the most suitable cluster planning. In the sequel, two criteria proposed for clustering evaluation and selection of an optimal clustering scheme are presented (Zanaty and Afifi, 2013: 38).

(1) Compactness: the members of each cluster should be as close to each other as possible. A common measure of compactness is the variance, which should be minimized.

(2) Separation: This indicates how distinct two clusters are. It computes the “distance” between two different clusters. The distance between representative objects of two clusters is a good example. This measure has been widely used due to its computational efficiency and effectiveness for hyper sphere-shaped clusters.

A good clustering result should have the properties of being both small intra-cluster compactness and large inter cluster separation at the same time. The two approaches are based on statistical tests and their major drawback is their high computational cost.

The performance of a fuzzy cluster validity index depends on the outcome of a fuzzy clustering algorithm, and a validity index is not able to provide desirable evaluation when the used clustering algorithm is not appropriate to the partitioning of a given data set (Kim et al., 2004: 2009).

3.1. Validity Indices Involving Only Membership Value

3.1.1. Partition Coefficient (PC)

Bezdek attempted to define a performance measure based on minimizing the overall content of pair wise fuzzy intersection in *U*, the partition matrix (Bezdek, 1981). The index was defined

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2} \tag{5}$$

Empirical studies think that the maximum V_{PC} lead to a correct interpretation of the samples considered. The best performance is achieved when the V_{PC} gets it maximum value.

3.1.2. Partition Coefficient (PC)

Bezdek proposed the classification entropy defined as below (Bezdek, 1974);

$$V_{CE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log_{a} u_{ij} \tag{6}$$

where the α is the base of the logarithm. The classification entropy index is a scalar measure of the amount of fuzziness in a given *U*. The index is computed for values of *c* greater than 1 and its values range in *[0, loga c]*. The best performance is achieved when the V_{CE} gets it minimum value.

3.1.3. Classification Entropy (CE)

Index developed by Dave (Dave, 1996) aimed to reduce the monotonous of V_{PC} index and defined as;

$$V_{MPC} = 1 - \frac{c}{c-1}\left(1 - V_{PC}\right) \tag{7}$$

The index values range in [0, 1]. The best performance is achieved when the V_{MPC} gets it maximum value. These mentioned indices use only the membership values of the fuzzy partition and they are used to measure the fuzziness of the fuzzy partition matrix. For this reason, the data are not directly related to the geometric shape and tend to decrease with cluster numbers *(c)*, which may be disadvantage of these scores.

3.2. Validity Indices Involving the Membership Values and the Data Set

3.2.1. Fukuyama and Sugeno Index (FS)

Validity function proposed by Fukuyama and Sugeno is defined by (Fukuyama and Sugeno, 1989)

$$V_{FS} = J_m(u,v) - K_m(u,v)$$
$$= \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m \left\|x_j - v_i\right\|^2 - \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m \left\|v_i - \overline{v}\right\|^2 \quad (8)$$

where $\overline{v} = \sum_{i=1}^{c} \frac{v_i}{c}$

In Equation (8) the first term shows that cluster density, second term shows that distances between cluster centers. The index should be minimum value for well partitioning.

3.2.2. Xie-Beni Index (XB)

This index developed by Xie and Beni, it is also kwon as density and separating validity function, is defined by (Xie and Beni, 1991: 841)

$$V_{XB} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m \left\|x_j - v_i\right\|^2}{n \min_{i,j} \left\|v_i - v_j\right\|^2} \quad (9)$$

The proposed validity index V_{XB} focused on two properties: compactness (closeness of elements) and separation (distinction of two different clusters). In their equation for V_{XB} (Equation (9)), the numerator indicates the compactness of the fuzzy partition, while the denominator indicates the strength of the separation between clusters. Xie Beni index should be minimum value.

3.2.3. Kwon Index (K)

Kwon (Kwon, 1998) tries to decrease monotonous increase tendency. To achieve this punishing function was introduced to the numerators of Xie and Beni's original index. In situations that cluster number closes to data number by developing Xie-Beni index. Index is defined;

$$V_K = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^2 \left\| x_j - v_i \right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\| v_i - \overline{v}_j \right\|^2}{\min_{i \neq j} \left\| v_i - v_j \right\|^2} \tag{10}$$

3.2.4. Partition Index (SC)

Index has defined as;

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} (u_{ik})^m \left\| x_k - v_i \right\|^2}{\sum_{k=1}^{n} (u_{ik}) \sum_{j=1}^{c} \left\| v_j - v_i \right\|^2} \tag{11}$$

This is the ratio of the sum of compactness and separation of the cluster (Zahid et al., 1999: 1089). It is a sum of individual cluster validity measures normalized by dividing it by the fuzzy cardinality of each cluster. *SC* is useful when comparing different partitions having an equal number of clusters.

3.2.5. Separation Index (S)

The Separation index uses minimum distance separation for validity. It is on a ratio scale in the metric of the root mean square measurement error of the test for the sample postulated. It quantifies "reliability" in a simple and direct way and has a clear interpretation.

4. ANALYSIS AND RESULTS

We tested the cluster validity indices for two well-known data sets (https://archive.ics.uci.edu/ml/datasets.html) in R studio. To test validity indices, we conducted extensive comparisons of some of the mentioned indices in conjunction with the FCM algorithm on a number of widely used data sets.

Table 1. Values of c preferred by validity indices for two data sets

Dataset	Abalone	Bupa Live Disorder
D	8	6
N	4147	345
c*	2	2
PC	**5**	**2**
MPC	**5**	**2**
CE	**5**	**2**
XB	**2**	**2**
S	**2**	**4**
SC	**2**	**8**
KWON	**2**	**2**
FS	**2**	**4**

Table 1 summarizes the results obtained when different validity indices were applied to two well-known data sets. The column c * in Table 1 gives the actual number of clusters for each data set, and the other columns show the optimal cluster numbers obtained using each index. *PC, CE* and *MPC* incorrectly identify the optimum for the abalone data set, but for bupa data set these indexes correctly identify the optimum cluster number. Also, *S, FS* and *SC* indexes fails to recognize c * in bupa dataset.

Table 2. Values of fuzzy validity indices in the range of c=2,…,9 (when m=1.15) for abalone data set using FCM algorithm

c	**PC**	**MPC**	**CE**	**XB**	**SC**	**S**	**Kwon**
2	.979	.958	.038	**.167**	**.262**	**.163**	**698.4**
3	.994	.991	.015	.184	.389	.182	774.3
4	.973	.964	.047	.261	.446	.248	1110.9
5	**.995**	**.994**	**.011**	.331	.674	.319	1339.2
6	.992	.991	.016	.473	.464	.467	2085.2
7	.991	.990	.018	.319	.532	.313	1467.2
8	.990	.989	.020	.590	.703	.577	2749.3
9	.985	.983	.027	.619	1.04	.606	2987.1

Table 2 shows values of fuzzy validity indices in the range between c equals to 2 and c equals to 9 while m equals to 1.15 for abalone data set using FCM algorithm. The optimum number of cluster equals to 2 for XB, SC, S and Kwon fuzzy validity indices when m equals to 1.15 for abalone data set. On the other hand, that the number of cluster equals to 5 is the best for PC, MPC and CE fuzzy validity indices according to Table 2.

Table 3. Values of fuzzy validity indices in the range of c=2… 9 (when m=1.15) for Bupa live disorders data set using FCM algorithm

c	**PC**	**MPC**	**CE**	**XB**	**SC**	**S**	**Kwon**
2	**.991**	**.982**	**.016**	**.104**	.624	.501	**36.81**
3	.956	.933	.074	.465	.720	.423	172.5
4	.936	.914	.107	.623	.553	**.351**	246.2
5	.938	.922	.108	.595	.543	.525	249.9
6	.943	.966	.100	.518	.562	.451	227.4
7	.913	.899	.153	1.267	.471	1.09	589.7
8	.915	.903	.146	.997	**.379**	.903	548.6
9	.909	.898	.159	1.027	.438	.907	584.1

The difference between Table 2 and Table 3 is the data set. Table 3 summarizes values of fuzzy validity indices for the same range and m value but for the Bupa live disorders data set. The optimum number of cluster is 2 for PC, MPC, CE, XB and Kwon. The optimum number of cluster equals to 8 for SC fuzzy validity index and 4 for S fuzzy validity index.

Fuzzifier is important parameter to determining the optimal number of cluster. For this reason, we change fuzzifier parameter for abalone data set. Optimal obtained cluster numbers are given in the Table 4 when the fuzzifier parameter *m* changes. When fuzzifier parameter is equal 1.5 *MPC* index fails to recognize c * in abalone dataset. At the same time, when *m*= 2, *MPC*, *XB*, *S* indices identify the optimum numbers incorrectly.

Table 4. Values of some fuzzy validity indexes in the range of c=2,…,9 using different fuzzifier parameters in abalone data set

m=1.5 **c**	**PC**	**MPC**	**CE**	**XB**	**S**	**Kwon**
2	**.917**	.834	**.139**	**.132**	**.120**	**550.9**
3	.895	.842	.185	.178	.157	749.4
4	.901	**.868**	.196	.170	.151	727.8
5	.873	.842	.232	.247	.209	1073.2
6	.847	.817	.282	.355	.299	1573.7
7	.884	.865	.236	.287	.246	1327.5
8	.869	.850	.264	.505	.420	2378.3
9	.859	.841	.294	.743	.625	3608.4
m=2 **c**	**PC**	**MPC**	**CE**	**XB**	**S**	**Kwon**
2	**.812**	.624	**.307**	.154	.154	**517.9**
3	.753	**.630**	.442	**.148**	**.148**	621.4
4	.703	.605	.561	.151	.151	648.2
5	.662	.578	.660	.201	.201	879.6
6	.632	.558	.708	.271	.271	1226.9
7	.631	.569	.778	.262	.262	1247.2
8	.627	.574	.809	.334	.334	1639.7
9	.575	.521	.934	.711	.711	3584.2

As a result of analyzes, we find that some of the mentioned indices incorrectly recognizes optimal cluster numbers c * for all mentioned data sets and that the use of the weighting parameter m = 2 in the general fuzzy clustering algorithm is not suitable for some data sets. Also, the analysis shows that as the fuzzifier parameter goes to 1, the results are closer to the desired conditions for the indexes.

5. CONCLUSION

Clustering is one of the multivariate statistical techniques that help to divide data groups according to similarities. Clustering can be performed either in crisp or fuzzy mode. In fuzzy clustering, the role of a validity index is very important.

For real data, it is clearly more difficult to estimate the number of clusters. In the literature of clustering, a large number of cluster validity indices of fuzzy clustering are there.

In this paper, we reviewed several validity indexes. Moreover, we conducted comparisons of the mentioned indices in conjunction with the FCM algorithm on widely used data sets and make a simple analysis of the experimental results.

We find some of the mentioned indices incorrectly recognize optimal cluster numbers c for all mentioned data sets. They have their own drawbacks. Therefore, we must to select the suitable index for different data sets.

Acknowledgment

This study was supported by Anadolu University Graduate School of Sciences as BAP project (1703F081).

REFERENCES / KAYNAKLAR

Bezdek J.C., Fuzzy mathematics in pattern classification, Ph.D. Dissertation, Cornell University, Ithaca, NY, 1973.

Bezdek J.C., “Cluster validity with fuzzy sets”, J. Cybernet., 3, 58–73, 1974.

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

Dave R.N., “Validating fuzzy partition obtained through c-shells clustering”, Pattern Recognition Lett., 17, 613–623, 1996.

El-Melegy, M.T., Zanaty, E.A., Abd-Elhafiez, W.M. and Farag, A., "On cluster validity indexes in fuzzy and hard clustering algorithms for image segmentation”, IEEE international conference on computer vision, vol. 6, VI 5-8, 2007.

Fukuyama Y. and Sugeno M., “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.

Hartigan J.A, Clustering Algorithms, Wiley, NewYork, 1975.

https://archive.ics.uci.edu/ml/datasets.html.

Kim, D. -W., Lee, K. H. and Lee, D., “On Cluster Validity Index for Estimation of the Optimal Number of Fuzzy Clusters”, Pattern Recognition, 37, pp.2009–2025, 2004.

Kwon S.H., “Cluster validity index for fuzzy clustering”, Electron. Lett. 34 (22), pp. 2176–2177, 1998.

Pakhira, M.K., Bandyopadhyay, S. and Maulik, U., “Validity index for crisp and fuzzy clusters”, Pattern Recognition, 37, 481–501, 2004.

Pal N.R. and Bezdek J.C., "On cluster validity for fuzzy c-means model", IEEE Trans. Fuzzy Systems, 3 (3), 370–379, 1995.

Saad, M. F. and Alimi, A. M., "Validity index and number of clusters", IJCSI International Journal of Computer Science, Vol. 9, Issue 1, No 3, 2012.

Xie X.L. and Beni G., "A validity measure for fuzzy clustering", IEEE Trans. Pattern Anal. Mach. Intell., 13, 841–847, 1991.

Zahid N., Limouri M. and Essaid A., "A new cluster-validity for fuzzy clustering", Pattern Recognition, 32, pp. 1089–1097, 1999.

Zanaty, E. A. and Afifi, A., "A new approach for automatic fuzzy clustering applied to magnetic resonance image clustering", American Journal of Remote Sensing, 1(2), 38-46, 2013.