Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

FUZZY C-MEANS CLUSTERING. TECHNIQUE AND EVALUATING RESULTS.

RIHOVA Elena (CZ), MALEC Miloslav (CZ)

Abstract. The present study aims to provide a better understanding of the fuzzy C-means clustering, their technique and evaluating their results. In order to do so, the fuzzy C-means clustering technique was described in detail, then those technique was applicated on the data sets, as generated, as real. In doing so, novel insights into the key drivers of fuzzy C-means clustering and evaluating of fuzzy clustering results. For determining the right number of clusters, were used indices based not only on membership function, but indices based on membership function and data sets. In most cases, those type of indices were able to recognize the right number of clusters. Indices involving the membership values and data sets were detected as more successful indices, although not always successful.

Keywords: fuzzy clustering, C-means, number of clusters

Mathematics Subject Classification: 62H30, 62A86

1 Possibilistic Fuzzy C-means Clustering

Clustering is an unsupervised process and can be classified into two categories: hard and fuzzy clustering. Although those two different clustering categories, they have the common goal. The task of clustering is to divide the set in to the optimal number of groups. The objects in the same group (this group is called a cluster) must be more similar to each other than to those objects in other clusters. In this article will deal with the second category of clustering - fuzzy clustering.

Fuzzy clustering is a form of clustering in which each object can belong to more than one cluster. The main concept in fuzzy clustering is based on membership degrees. These membership grades indicate the degree to which data points belong to each cluster. A better reading of the memberships, avoiding misinterpretations, would be (Höppner, Klawonn, Kruse and Runkler, 1999): If the object x_i has to be assigned to a cluster, then with the probability u_{ij} to the cluster j. However, the normalization of u_{ij} can lead to unexpected bad result in finding and discovering outliners. The membership values affects the clustering results.

By dropping the normalization constraint (1) in the following definition one tries to achieve a more intuitive assignment of degrees of membership and to avoid undesirable normalization effects (Oliveira, 2007).

Let $X = \{x_1; \ldots ; x_n\}$ is a data set with k clusters, where number of clusters is $1 < k < n$ and represented by the fuzzy sets μ_{Ch}. Hence, $U_p = (u_{ij}) = \mu_{Ch}(x_i)$ is a possibilistic partition of X if:

$$\sum_{i=1}^{n} u_{ij} > 0 \, , \; \forall j \in \{1,\ldots,k\} \, , \tag{1}$$

holds. The degree of the object x_i to cluster C_h is u_{ij} and achieves value [0, 1]. The membership degrees for one datum now resemble the possibility (in the sense of possibility theory (Dubois and Prade, 1988) of its being a member of the corresponding cluster (Daver and Krishnapuram,1997; Krishnapuram and Keller, 1992).Consequently, J would not be appropriate for this type of fuzzy clustering. The normalization term leads to following problem: J would reach its minimum for $u_{ij} = 0$ for all objects in data set, it means no one object is not assigned to cluster. Consequently, clusters are empty. According Krishnapuram and Keller (Krishnapuram and Keller, 1992) to avoid this problem the objective function must be modified to:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{q} D_{ij}^{2} + \sum_{j=1}^{k} \eta_i \sum_{i=1}^{n} (1 - u_{ij})^{q} \, , \tag{2}$$

where $\eta_i > 0$ ($i = 1; \ldots ; k$).

The first part of this objective function leads to a minimization of the weighted distances. The second part puts down the first part of this function: when the first part leads to 1, the second part suppresses it: $(1 - u_{ij})^q$. In tandem with the first term the high membership can be expected especially for data that are close to their clusters, since with a high degree of belonging the weighted distance to a closer cluster is smaller than to clusters further away (Oliveira, 2007). The updating the membership degrees that is derived from J by setting its derivative to0 is (Krishnapuram and Keller, 1992):

$$u_{ij} = \frac{1}{1+\left(\frac{D_{ij}^2}{\eta_i}\right)^{\frac{1}{q-1}}} \tag{3}$$

Eq. 3.8 shows that the membership u_{ij} (belonging the object $x_i$to cluster C_h) depends on the distance from this object to cluster. Small value of the distance (strong similarity) leads to high membership degree, and the large value of distance means to low membership value. And the other one parameter is η_i - the distance from object x_j to the cluster C_h ,which membership degree should be 0,5.

Since that value of membership can be seen as definite assignment to a cluster, the permitted extension of the cluster can be controlled with this parameter (Oliveira, 2007), but the parameterη_i may have the different geometrical interpretation, this interpretation depends on the cluster shape. In case of the possibilistic *k*-means, the clusters diameter is $\sqrt{\eta_i}$ (Höppner, Klawonn, Kruse and Runkler 1999). If a kind of information about clusters is known a prior, η_i can be set to any value. In case the same optionalities of all clusters this parameter can be the same for all clusters. But in the real world this information about cluster optionalities is unknown in advance. Hence, parameter $\eta_i$should be calculated. To calculate the optional value of η_i can be used a probabilistic clustering model. The parameters η_i are then estimated by the fuzzy intra-cluster distance using the fuzzy memberships matrix U_f as it has been determined by the probabilistic counterpart of the chosen possibilistic algorithm (Krishnapuram and Keller, 1992):

$$\eta_i = \frac{\sum_{i=1}^{n} u_{ij}^q D_{ij}^2}{\sum_{i=1}^{n} u_{ij}^q} \tag{4}$$

2 Validation indices

The problem for finding an optimal number of clusters *C** is usually called cluster validity problem. In order to solve the cluster validity problem, validity indices must enclose, take into account, some specific are as which enable to solve this problem successfully. Those areas are: compactness, separation, noise and overlap. A lot of validity indices for fuzzy clustering are. Early indices such as the partition coefficient and classification entropy make use only of membership values. The main advantage of those indices they are easy to compute. Now, it is widely accepted that a better definition of a validity index always consider both partition matrix *U* and the data set itself. In this paper we will go with the classification of indices by Wang (Wang, 2007). The first group of those indices is indices involving only the membership values.

Partition coefficient (*PC*). *PC* index is based on minimizing the overall content of pair wise fuzzy intersection in *U*, the partition matrix**.** The index was defined as

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

Those 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. In general, we find an optimal cluster number *C** by solving $\max_{2 \le C \le n\text{-}1} PC$ to produce the best clustering performance for the data set *X*.¨

Modified *PC* (*PC_{mod}*) index (Dave, 1992) as a modification of the previous one and can take values ⟨0,1⟩:

$$PC_{\mathrm{mod}}(C) = 1 - \frac{C}{C-1}(1 - PC(C))\,. \tag{6}$$

The optimal number of cluster *C** is defined by solving of:

$$\max_{2 \le C \le \acute{n} = 1} PC_{\mathrm{mod}}(C)\,. \tag{7}$$

When the variability in clusters is small, this modified Dunn's coefficient PC_{mod} usually determined the number of clusters correctly (Řezanková, Húsek, 2012).When the cluster variability is greater, the normalized Dunn's coefficient usually achieved its highest value for the highest possible number of clusters. (Řezanková, Húsek, 2012) Validity indices involving the membership values and the data set are more successful compared with validity indices involving the only membership values. (Wang, 2007). The most succesfull index from those group is **Xie and Beni (*XB*) validity index.** *XB* index was proposed in by Xie and Beni (Xie, Beni, 1991) with *q* = 2 and modified by Pal and Bezdek (Pal, Bezdek, 1995) was defined as:

$$XB = \frac{J_m(u,v)/n}{Sep(v)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{2}\left\|x_j - x_i\right\|^2}{n.\min_{i,j}\left\|v_i - v_j\right\|^2} \tag{8}$$

XB includes two components: compactness in the numerator and separation, which is represented in denominator. The optimal number of clusters C^* can be find by solving $\min_{2 \leq C \leq n-1} XB$ to produce the best clustering performance for the data set X. XB index has tendency to monotonically decrease with increasing number of clusters.

3 Case Study

The main objective of third section is to compare the performance of those indices in evaluating the optimal number of clusters. In the following experiments presented here, were tested the cluster validity indices for 5 well-known data set from *UCI Machine Learning Repository* and for 5 generated data sets.

The data set Abalone:

This data set comes from an original, non-machine study. Those data was made in 1994 by Warwick J Nash and collective. Data set Abalone contains 8 variables:

1. length (continuous in mm) Longest shell measurement,
2. diameter (continuous in mm) perpendicular to length,
3. height (continuous in mm) with meat in shell,
4. whole weight (continuous in grams) whole abalone,
5. shucked weight (continuous in grams) weight of meat,
6. viscera weight (continuous in grams) gut weight (after bleeding),
7. shell weight (continuous in grams) after being dried,
8. rings (integer) +1.5 gives the age in years.

Total number of samples is 4177. Abalone data set contains two clusters. Data set samples are highly overlapped and none correlated.

The data set Breast Tissue:

This database includes 106 instances. Six classes of freshly excised tissue were studied using electrical impedance measurements:

1. carcinoma,
2. fibro-adenoma,
3. mastopathy,
4. glandular,
5. connective,
6. adipose.

The data set can be used for predicting the classification of either the original 6 classes or of 4 classes by merging together the fibro-adenoma, mastopathy and glandular classes whose discrimination is not important (they cannot be accurately discriminated anyway). All those data were divided into 6 groups with help of 9 variables:

1. impedivity (ohm) at zero frequency,
2. phase angle at 500 khz,
3. high-frequency slope of phase angle,

4. impedance distance between spectral ends,
5. area under spectrum,
6. area normalized,
7. maximum of the spectrum,
8. distance between impedivity and real part of the maximum frequency object,
9. length of the spectral curve.

The data set Cardiography:

The represented data set obtains 2126 samples and 21 variables, which are divided on 3 clusters by fetal state class code: normal, suspect and pathologic. Data set samples are middling overlapped and none correlated.

The data set Connectionist Bench:

For this data set 15 people were gradually microphones into 6 series, where each series has 11 vowels pronounced. The data file contains of 990 subjects (voiced vowels), where for each vowel recorded 10 characteristics. Because the multicollinearity is not present in the data file, all of the variables will be enter into the analysis.

The data set Wholesale customers

Current data set contains 440 objects, divined into 2 clusters by following attributes:

1. fresh: annual spending (m.u.) on fresh products,
2. milk: annual spending (m.u.) on milk products,
3. grocery: annual spending (m.u.)on grocery products,
4. frozen: annual spending (m.u.)on frozen products,
5. detergents paper: annual spending (m.u.) on detergents and paper products,
6. delicatessen: annual spending (m.u.)on and delicatessen products (continuous).

3.2 Generated Data Sets

The data set with 8 clusters (Example_2):

Data set contains 320 objects, which were put in 8 clusters. The number of variables is 8.All data sets have the normal distribution.
The data set Example_2, obtains overlapped clusters.

The data set with 9 clusters (Example_7):

Current data set contains 360 objects, which were put in 9 clusters. The number of variables is 8. Data set have the normal distribution. The data set Example_7 obtains middling - overlapped clusters.

The data set with 10 clusters (Example_11):

The data set contains 400 objects, which were put in 10 clusters. The number of variables is 8 with normal distribution. The data set Example_11obtain middling - overlapped clusters.

The data set with 11 clusters (Example_19):

The data set contains 440 objects, which were put in 9 clusters. The number of variables is 8 with normal distribution. The data set Example_19 obtain middling - overlapped clusters.

The data set with 12 clusters (Example_25):

The data set contains 480 objects, which were put in 9 clusters. The number of variables is 8 with normal distribution. The data set Example_25 obtain middling - overlapped clusters.

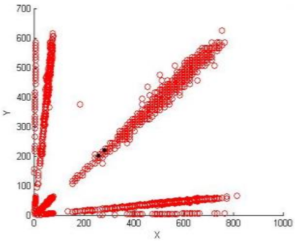

Fig. 1. The data set Abalone.

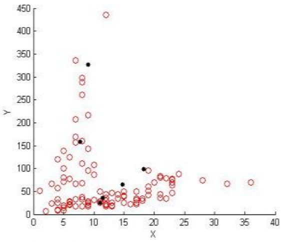

Fig. 2. The data set Breast Tissue.

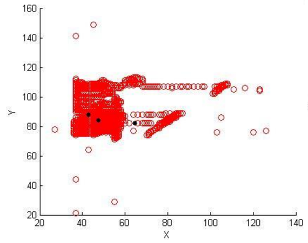

Fig. 3. The data set Cardiography.

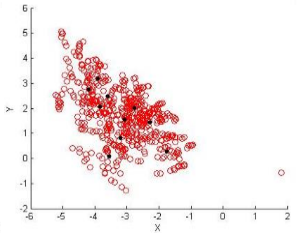

Fig. 4. The data set Connectionist Bench (Vowel Recognition – Deterring Data).

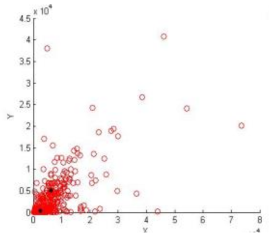

Fig. 5. The data set Wholesale customers.

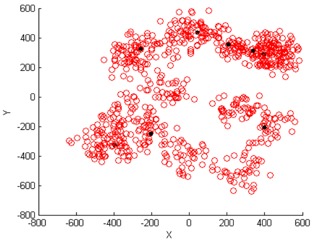

Fig. 6. Example_2: The data set with 8 clusters (obtaining 320 objects); overlapped clusters with normal distribution.

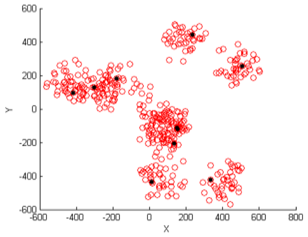

Fig. 7. Example_7: The data set with 9 clusters (obtaining 360 objects);middling - overlapped clusters with normal distribution.

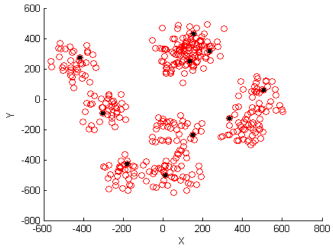

Fig. 8. Example_11: The data set with 10 clusters (obtaining 400 objects) middling - overlapped clusters with normal distribution.

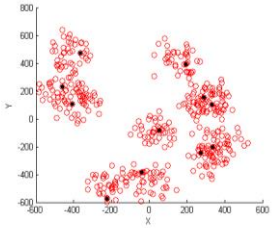

Fig. 9. Example_19:
The data set with 11 clusters (obtaining 440 objects) middling - overlapped clusters with normal distribution.

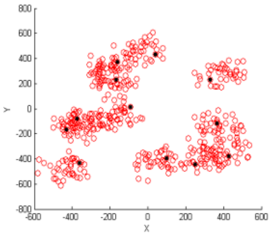

Fig. 10.: Example_25:
The data set with 12 clusters ; (obtaining 480 objects); middling - overlapped clusters with normal distribution.

4 Results and Discussion

The result of evaluating fuzzy C-means clustering results are shown in Tab. 1. and 2. In Tab. 1. are shown the number of clusters, which are calculated for *PC*, PC_{mod}, and *XB* index for real data sets.

Data set	*C**	*PC*	PC_{mod}	*XB*
Abalone	2	2	2	2
Breast Tissue	6	2	2	2
Cardiotocography	3	2	2	3
Connectionist Bench (Vowel Recognition - Deterring Data)	11	3	2	8
Wholesale customers	2	2	3	2
Successfulness,%	-	40%	20%	60%

Tab. 1. Values of *C* by validity indices for real data sets.

Summing up, we can determine in which of the abovementioned cases the indices worked incorrectly, in other words, to find out what affected it.

To sum up, the most successful indexes are *XB* with success rates of 60%.The worst-performing index is PC_{mod} with a success rate of 20%. However the *PC* index was better with a success rate of 40%. In Table 2 are shown the number of clusters, which are calculated for *PC*, PC_{mod}, and *XB* index for generated data sets.

Data set	*C**	*PC*	PC_{mod}	*XB*
Example_2	8	3	8	8
Example_7	9	2	2	9
Example_11	10	2	2	2
Example_19	11	4	4	11
Example_25	12	2	6	12
Successfulness, %	-	0%	20%	60%

Tab. 2. Values of *C* by validity indices for generated data sets.

The number of clusters ranged from 8 to 12. As Bezdek (Bezdek, 1987) pointed out: the number of variables does not affect the fuzzy *C*-means clustering results. For this reason, the same number of variables was chosen for all generated data sets – eight. However, the degree of overlapping is different for each data set as shown in Fig 1 – 10. Fuzzy *C*-means clustering with Euclidean distance was applied to each data set. To sum up, the most successful indexes are *XB* with success rates of 60%.The worst-performing index is *PC* with a success rate of 0%. However the modified *PC* index was better with a success rate of 20%.

The validation of clustering structures is the most difficult and frustrating part of cluster analysis. That’s why the issue of the definition of the indexes, which would be good for data with large variability and a large number of clusters, has not yet been resolved. As shown by the results of the approach, which we suggest, this modification can increase the efficiency of the correct determination of the number of clusters.

From the experiment's results, it can be drawn that the modification method determines the number of clusters correctly. The next part of current study is represented in following subchapter.
Study the data can help to define what behavior we can expect from the clusters with different overlap but with normal distribution. Let’s observe how the behavior of those indices changes with an increasing number of clusters.

References

[1] URL, http://www.ics.uci.edu/∼mlearn/databases.html

[2] BEZDEK, J.C., HATHAWAY, R.J., SABIN, M.J. (1987). Convergence theory for fuzzy *c*-means: counter-examples and repairs. *IEEE Transaction on Systems*, vol.17, no.5, p.: 873–877.

[3] DAVE, R.N., BHASWAN, K. (1992) Adaptive fuzzy *c*-shells clustering and detection of ellipses. *IEEE Transactions of Neural Networks*, vol.3, no.5, p: 643–662.

[4] DAVEҐ, R., KRISHNAPURAM, R. (1997) Robust clustering methods: a unified view. *IEEE Transactions on Fuzzy Systems*, vol. 5, no.2, p: 270–293.

[5] DUBOIS, D., PRADE, H., (1988). *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. New York: Plenum Press.

[6] HÖPPNER, F., KLAWONN, F., KRUSE, R., AND RUNKLER, T. (1999). *Fuzzy Cluster Analysis*. John Wiley & Sons, Inc., New York

[7] KRISHNAPURAM, R.,NASRAOUI, O., KELLER, J. (1992). The fuzzy *c* spherical shells algorithm: a new approach. *IEEE Transactions of Neural Networks*, vol.3, no. 5, p.: 663–671.

[8] OLIVEIRA, J.W. (2007) *Advances in Fuzzy Clustering and its Applications*. London: John Wiley & Sons, Ltd.

[9] PAL, N.R., BEZDEK, J.C. (1995). On cluster validity for fuzzy *C*-means model. *IEEE Transactions of Fuzzy Systems*, vol.3, no.3, p.: 370–379.

[10] REZANKOVA, H., HUSEK, D.(2012): Fuzzy Clustering: Determining the Number of Clusters. In: *Computational Aspects of Social Networks*. Sao Carlos: Research Publishing Services, p.: 277--282.

[11] WANG, W., ZHANG,Y. (2007). On fuzzy cluster validity indices. *Fuzzy Sets and Systems*. vol. 158, no. 19. P.: 2095 – 2110.

[12] XIE, X.L., BENI, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions of Pattern Analalysis and Machine Intelligence*, vol.13, p.: 841–847.

Current address

Rihova Elena, Mgr., Ph.D.
University of Business in Prague
Department of Information Technologies and Analytical Methods,
Spalená 76/14 110 00, Praha, Czech Republic
E-mail: rihova.elena@gmail.com

Malec Miloslav, doc. RNDr., CSc.
University of Business in Prague
Department of Information Technologies and Analytical Methods,
Spalená 76/14 110 00, Praha, Czech Republic
E-mail: malecm@vso-praha.eu