ON EVALUATING OF FUZZY CLUSTERING RESULTS

Elena Říhová – Tatiana Makhalova

Abstract

Evaluating clustering results is a concept to estimate how good clustering results areh. A lot of validity indices and different validity's techniques have been proposed. However, the most of those indices have their weakness. For instance, some of indices have the adventure of being easy to compute, but are only useful for a small number of well-separated clusters. Other indices lack direct connection to the geometrical properties of the data set. Current study has shown that, of the several of the validity indices (*PC*, PC_{mod} and *E* index) have their advantages and disadvantages. Furthermore, all experiments were held on real and generated data sets with the small and large number of clusters, as with good separated clusters, as not. Based on results of the current analysis, it was discovered that the new *E* index is useful for evaluating fuzzy *C*-means clustering results with small and large numbers of clusters (from 2 to 8 clusters) on data sets with normal distribution. To sum up the results of current research, the new proposed index *E* has merit in cluster validity problems, and brings more reliable results than previously used indices.

Key words: clustering, indices, fuzzy, optimal number of clusters

JEL Code: C18, C38, C69

Introduction

There are many different coefficients for estimating the optimal number of clusters. Each of these coefficients has its strengths and weaknesses. In this research, several coefficients for estimating the optimal number of clusters (for fuzzy clustering techniques) will be examined. Those coefficients are: Dunn's coefficient (*PC*), modified Dunn's coefficient (PC_{mod}). Researchers have been studying fuzzy clustering problems for a long time. The current problem of evaluating clustering results and determining the correct number of clusters has been subject of several research projects. However, during all these years of research, the best coefficient has not yet been proposed. There are coefficients that work better than others, but

they are still not enough good. In recent years, significant discoveries have been made. The accuracy of the results has increased since using membership values and data sets together for index calculation was established.

1 Possibilistic Fuzzy *C*-Mean 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.

The range of applications of the cluster analysis is very wide: in medicine, in archeology, biology, chemistry, psychology, marketing and others. Nevertheless, the versatility of the cluster analysis led to large number of different clustering methods.

Regardless of the purpose and field of research cluster analysis involves the follow steps:

1. to select the data set sample for clustering,
2. to define the set of variables (feature space), in which the clustering will be done,
3. to calculate the value of measure similarity or dissimilarity between all objects in data set,
4. to applicate the method of cluster analysis to create groups of similar objects,
5. to validate the obtained results.

Of course, the main requirement for the data set is their uniformity and completeness. There are exist many methods of fuzzy clustering. The most-known methods are *k*-medoids and *C*-means clustering (with probabilistic and possibilistic algorithms).

As was mentoed above, there are exist several fuzzy *C*-means algorithms, for extance: possibilistic fuzzy C-mean clustering, probabilistic fuzzy *C*-means clustering and others. In this research will be used possibilistic fuzzy *C*-means clustering due for the following reasons:

1. While probabilistic memberships rather divide the data space, possibilistic membership degrees only depend on the typicality to the respective closest clusters. (Oliveira, 2007).

2. In the probabilistic fuzzy *C*-means algorithm the centers of every cluster are driven apart, it means includes a part of the object's membership values, hence thus leaves less that may attract other cluster centers. Consequently, to share out object between clusters disadvantageous. In the possibilistic fuzzy *C*-means algorithm has not this effect.

These algorithms are based on objective functions *J*, which are mathematical criteria that quantify the goodness of cluster models that comprise prototypes and data partition. Objective functions serve as cost functions that have to be minimized to obtain optimal cluster solutions (Řezanková, 2010). Thus, for each of the following cluster models the respective objective function expresses desired optional ties of what should be regarded as ''best'' results of the cluster algorithm.

The membership degrees for one datum now resemble the possibility 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:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{q} D_{ij}^{2} + \sum_{j=1}^{k} \eta_i \sum_{i=1}^{n} (1-u_{ij})^{q} , \tag{1.1}$$

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 to 0 is (Krishnapuram and Keller, 1992):

$$u_{ij} = \frac{1}{1+\left(\frac{D_{ij}^2}{\eta_i}\right)^{\frac{1}{q-1}}}. \tag{1.2}$$

Eq. 1.2 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 parameter η_i may have the different geometrical interpretation, this interpretation depends on the cluster shape. In case of the possibilistic *C*-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 η_i should be calculated.

$$\eta_i = \frac{\sum_{i=1}^{n} u_{ij}^q D_{ij}^2}{\sum_{i=1}^{n} u_{ij}^q}. \tag{1.3}$$

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).

2 Indices for evaluating fuzzy clustering results

Very often users do not have any information about the number of clusters in data set. Consequently, finding the optimal number of clusters is an important problem. The problem for finding an optimal number of clusters *k** 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.

Compactness is a measure of the proximity of object's vectors comprising the same class of its center (Saad, 2012). Separation – a measure of how similar that object is to objects in its own cluster compared to objects in other clusters, shows the isolation of clusters. The basic measure of separation is the deviation between two fuzzy cluster centers.

This two values are the basic values of validity, as for hard, as for fuzzy clustering. The small local value of compactness shows, that each cluster is compact and the great local value of separation shows, that clusters are good separated.

Noise – noisy objects are objects that do not belong to any clusters of data set. According by Saad, if the data set contains some noise objects, then we can see that the validity indices take the noisy object in a compact and separated class from the rest of the classes. Thus, the noise aspect is crucial in the classification of data (Saad, 2012).

Overlap – is a measure, that indicating the degree of overlapping two clusters, the measure with which two clusters overlap and have similar future vectors. (Rezankova, 2010)

Large number of validity indices for fuzzy clustering exist in the literature. Early indices such as the partition coefficient and classification entropy make use only of membership values and have the advantage of being 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 work will be presented the classification of indices by Wang (Wanga, 2007). In this section we review some cluster validity indices available in the literature.

2.1 Dunn's index (*PC*)

Bezdek (Bezdek, 1974) attempted to define a performance measure based on minimizing the overall content of pair wise fuzzy intersection in *U*, the partition matrix. He proposed cluster validity index for fuzzy clustering: partition coefficient (*PC*). The index was defined as

$$PC = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{2} \,. \tag{2.1}$$

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 $\langle 1/C,\ 1\rangle$, where *C* is the

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

2.2 Modified Dunn's index (PC_{mod})

The next validity index was proposed by Dave (Dave, 1992) as a modification of the previous one:

$$PC_{\mathrm{mod}} = 1 - \frac{C}{C-1}(1 - PC(C)). \qquad (2.2)$$

This index can take values $\langle 0,1\rangle$, where C^* is the optimal number of clusters. This cluster number C^* is defined by solving of $\max_{2\le C\le \acute{n}=1} PC_{\mathrm{mod}}$

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)

2.3 *E* index (*E*)

The last one index, which is as follows: to combine into one index two components using the harmonic mean. One of the components is based on fuzzy clustering theory and the other one is based on hard clustering theory. The theory of fuzzy clustering is based on the assumption that each object belongs to each cluster with a membership degree u_{ij}. The hard clustering theory is based on the assumption that each object belongs to one cluster, the average distance from the cluster center and objects of this cluster should be minimal.

Joining two elements based on different approaches into one index helps us to reduce disadvantages of both. The first element here is Dunn's coefficient.

The second element is based on the hard clustering theory: to sum the ratio of the distance minimum in case n clusters to the distance minimum in case 1 cluster (k). And now we have to solve the optimization problem. It can be represented in the following way:

$$f(x) = \frac{2}{\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^2} + \frac{1}{1 - \frac{\sum_{h=1}^{k} D_{h,\min}}{D_{1,\min}}}} \to \max . \tag{2.3}$$

This function tends to its maximum for the best clustering because the inverse values of the indexes of *PC* and *N* receive its minimum for the best clustering.

An optimization problem consists of maximazing a real function by systematically choosing input values from within an allowed set and computing the value of the function. The *E* coefficient can be defined as:

$$E = \frac{2}{\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^2} + \frac{1}{1 - \frac{\sum_{h=1}^{k} D_{h,\min}}{D_{1,\min}}}} . \tag{2.4}$$

The optimal number of clusters *C** for the data set *X* can be found by solving $\max_{2 \le C \le n-1} E$.

3 Case studie

The main objective of this subsection is to compare the performance of some of the abovementioned indices in determining the true number of clusters. In the following experiments presented here, were tested the cluster validity indices for some well-known data sets from *UCI Machine Learning Repository* and generated data sets with the different number of clusters and different overlapped degree. All data sets are illustrated on Figs 1-12.

The data set Iris

It contains three classes of 50 cases each (the total number of cases is 150), where each class refers to a type of iris plant. One class is good separable from the other two; the latter are not linearly separable from each other

The data set Glass

The number of cases is 106, six classes, the number of attributes is 9 (numeric, predictive attributes).

The data set Wine

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. A data frame with 178 observations.

The data set Statlog (Shuttle)

Current data set obtains 58000 objects, divined into 5 clusters by 7 variables. Clusters are very high-overlapped and in 2-dimensional space are impossible to show five cluster's centers. Approximately 80% of the data belongs to class 1, others 20% are distributed by 4 clusters.

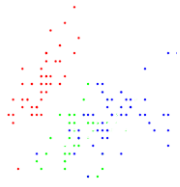
Generated Data Sets 1-4

Current data set obtains 160 objects, divined into 4 clusters by 4 variables. Clusters have the different high-overlapped and in 2-dimensional space are impossible to show five cluster's centers.

Generated Data Sets 5-8

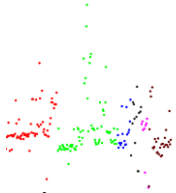
Current data set obtains 320 objects, divined into 4 clusters by 4 variables. Clusters have the different high-overlapped and in 2-dimensional space are impossible to show five cluster's centers.

Fig. 1: Data Set Iris



Source: Autor

Fig. 2: Data Set Glass



Source: Autor

Fig. 3: Data Set Statlog (Shuttle)

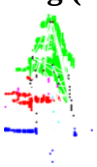
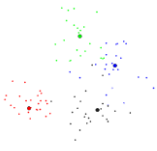


Fig. 5: Generated Data Set 1

Source: Autor

Fig. 6: Generated Data Set 2



Source: Autor

Fig. 7: Generated Data Set 3

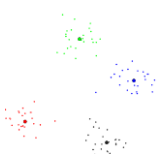
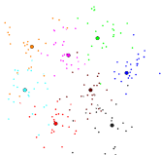


Fig. 9: Generated Data Set 5

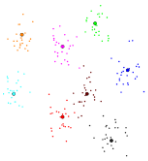
Source: Autor

Fig. 10: Generated Data Set 6



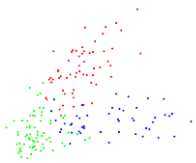
Source: Autor

Fig. 11: Generated Data Set 7



Source: Autor

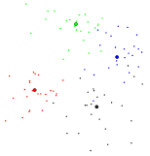
Fig. 4: Data Set Wine



Source: Autor

Source: Autor

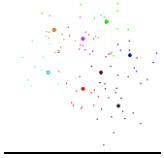
Fig. 8: Generated Data Set 4



Source: Autor

Source: Autor

Fig. 12: Generated Data Set 8



Source: Autor

Results and Discussion

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. Obtained results of evaluating fuzzy *C*-means clustering are presented in Tab.1.

Tab. 1: The Results of Evaluating

The Data Set	Right Number of Clusters	*PC*	PC_{mod}	*E*
Iris	3	2	2	3
Glass	6	2	2	6
Statlog (Shuttle)	5	2	3	2
Wine	3	2	2	2
Generated Data Set 1	4	4	4	4
Generated Data Set 2	4	2	10	4
Generated Data Set 3	4	4	10	4
Generated Data Set 4	4	2	9	4
Generated Data Set 5	8	2	8	8
Generated Data Set 6	8	2	10	8
Generated Data Set 7	8	2	9	8
Generated Data Set 8	8	2	8	8
Successfulness, %	-	16.67	25.00	83.33

A better works *E* index, in 10 from 12 of the cases shows correct results. However, PC_{mod} is not able to recognize the optimal number of clusters for data sets with more than 4 clusters. Incorrectly identifies the optimum for the data sets: Iris, Glass, Statlog (Shuttle), and Wine and for the most part of generated data sets. The most successful coefficient in those experiments was the *E* coefficient. Its successfulness is 83,33%. *E* incorrectly identifies the optimum for the data sets: Statlog (Shuttle), and Wine.

Conclusion

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.

Based on results of the current analysis, it was discovered that the new *E* index is useful for evaluating fuzzy *C*-means clustering results with small and large numbers of clusters (from 2 to 8 clusters) on data sets with normal distribution. As Saad stated (Saad, 2012): Moreover, the main idea of the functions of validity is based on the geometry of objects, within the same class must be compact and in different classes should be separated.

To sum up the results of current research, the new proposed index *E* has merit in cluster validity problems, and brings more reliable results than previously used indices.

Acknowledgment

The section Case Studie was made by T. Makhalova supported by the Russian Science Foundation under grant 17-11-01294 and performed at National Research University Higher School of Economics, Russia.

References

BEZDEK, J.C. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, vol.3, no.3, p.: 58–73.

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

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.

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

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.

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

REZANKOVA, H., HUSEK, D., LÖSTER, R.(2010): Clustering with Mixed Type Variables and Determination of Cluster Numbers, In: *CNAM and INRIA*, Paris, p.: 1525-1532.

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.

SAAD, M., F., (2012). Validity Index and number of clusters. *International Journal of Computer Science Issues*, vol. 9, no. 1, no 3, p.: 52 – 57

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

http://www.ics.uci.edu/~mlearn/databases.html

Contact

Mgr. Elena Říhova, Ph.D.

University of Business, Prague

Spalena 51, Prague, Czech Republic

elena.rihova@gmail.com

Mgr. Tatiana Makhalova

National Research University Higher School of Economics

Kochnovsky Proezd 3, Moskva, Rusko, 125319

tpmakhalova@hse.ru