THE FUZZY CLUSTERING PROBLEMS AND POSSIBLE SOLUTIONS

Elena Makhalová – Iva Pecáková

Abstract

With the widely used algorithms based on fuzzy logic, fuzzy clustering is attracting increasing attention. Nowadays there are many fuzzy clustering algorithms. One of their fundamental problems is to determine the optimal number of clusters, which has a deterministic effect on the clustering results. We can determine the optimal number of clusters with help of cluster validity indices. Cluster validity indices are used for estimating the quality of partitions produced by clustering algorithms and for determining the number of clusters in data.

In this paper, factors determining the number of clusters in the existing fuzzy clustering are researched and their advantages and disadvantages are examined. The important task is to estimate the proper number of clusters in actual dataset. This paper describes a new validity index for fuzzy clustering (modified index *E*) and modifications improving its performance as cluster number selection criterion for fuzzy k-means. The proposed indexes are tested and validated using several data sets. The paper also presents experimental results concerning them.

Keywords: fuzzy clustering, fuzzy sets, number of clusters.

JEL Code: C18, C38, C69

Introduction

Clustering has become a widely accepted synonym of a broad array of activities of exploratory data analysis and model development in science, engineering, life sciences, business and economics, defense, and biological and medical disciplines (J. Valente de Oliveira, 2007). Clustering techniques can be used to organize data (numerical or categorical or a mixture of both) into groups based on similarities among the individual data items. In other words, clustering techniques is a tool for discovering previously hidden structure in a set.

Like all clustering algorithms, also the fuzzy clustering algorithms are endowed with a distance function, which measures the dissimilarity in data, or with a special function, which is determined to measure their similarity. But there is a significant difference from the classical hard clustering algorithm. In the classical hard clustering, each data point x_i in the dataset of size n, $X \equiv \{x_1,...,x_n\}$ belongs to a single of k clusters. In the case of the fuzzy clustering, every data point x_i in the dataset X is assigned to all clusters, but with different membership degrees. This

membership degree expresses how ambiguously or definitely a data point should belong to a cluster (F. Höppner et al., 1999).

The concept of these membership degrees is substantiated by the definition of fuzzy sets by L. A. Zadeh (1965): Let X is a space of points, with a generic element of X denoted by x; thus $X = \{x\}$. A fuzzy set A in X is characterized by a membership function $f_A(x)$ which associates with each point in X a real number in the interval $\langle 0, 1\rangle$; the value of $f_A(x)$ at x represents the "grade of membership" of x in A. Thus, the closer the value of $f_A(x)$ is to one, the greater the degree of membership of x to A (Xie, N., 2011). A fundamental problem of this approach to clustering is to determine the best number of clusters, which has a deterministic effect on the clustering results (T. Löster, 2012).

1. Cluster validity indexes, their advantages and disadvantages

Clustering validity is a concept to evaluate how good clustering results are. There are many cluster validity indexes that have been proposed in the literature for evaluating fuzzy and other clustering techniques. In this paper we will research some of these indexes, their pros and cons.
The basic validity index associated with the fuzzy clustering is Dunn's coefficient defined by

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

Here $u_{ij} \in \langle 0, 1\rangle$ are the membership degrees of objects to clusters. This index assumes only the compactness measurement for each cluster and for the data structure. It is obviously a lack connection with the geometrical structure of data[1]. According to the theory of fuzzy sets, the sum of memberships of every object to all clusters is 1. Consequently, with the increasing number of clusters, the single degrees of cluster membership have a decreasing value. Squaring the membership's degrees we obtain even smaller values. So, with the increasing number of clusters the value of this coefficient is decreasing (S. Brodowski, 2011).

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

$$PC_{\text{mod}}(k) = 1 - \frac{k}{k-1}(1 - PC(k)) . \tag{2}$$

This index can take values $\langle 0, 1\rangle$; k is the optimal number of clusters. This cluster number k is defined by solving of

$$\max_{2 \le k \le n} PC_{\text{mod}}(k) . \tag{3}$$

When the variability in clusters is small, this normalized Dunn's coefficient usually determined the number of clusters correctly (H. Řezanková a D. Húsek, 2012).When the cluster

[1]This is a characteristic for a cluster validity index for fuzzy clustering.

variability is greater, the normalized Dunn's coefficient usually achieved its highest value for the highest possible number of clusters (H. Řezanková a D. Húsek, 2012).

The total average silhouette coefficient *SC* (4) is the most complicated validity index for fuzzy clustering mentioned in this paper. This coefficient can determine the compactness and separation degree for the whole data structure, not only for each cluster. The silhouette coefficient for each point determines how that point is similar to points in its own cluster compared to points in other clusters; it ranges from –1 to +1. The *SC* coefficient combines ideas of both cohesion, which is the sum of the weight of all links within a cluster, and separation, which is the sum of the weights between points in the cluster and points outside the cluster, but for individual points, as well as clusters.

For an individual point *i*, the silhouette coefficient is based on two measures: on the average distance between *i* and every points in its cluster (A_{iC_i}) and on the minimal average distance between *i* and points in other clusters (B. Rezaee, 2010). The total average silhouette coefficient is than

$$SC = \frac{1}{n}\sum_{i=1}^{n}\frac{\min(A_{ij,}j \in C_{-i}) - A_{iC_i}}{\max(\min(A_{ij,}j \in C_{-i}), A_{iC_i})}. \quad (4)$$

Here C_{-i} denotes cluster labels which do not include the case *i* as a member, while C_i denotes the cluster label which includes the case *i*. If $\max(\min(A_{ij,}j \in C_{-i}), A_{iC_i})$ equals 0, the silhouette coefficient of case *i* is not used in the average operations. The average *SC* over all data in a cluster is a measure of how tightly grouped all the data in the cluster are. Thus the average *SC* over all data of the entire dataset is a measure of how appropriately the data has been clustered. If there are too many or too few clusters, as may occur when a poor choice of *k* is used in the k-algorithm, some of the clusters will typically display much narrower silhouettes than the rest. The silhouette coefficient is typically between 0 and 1. When the silhouette coefficient is closer to 1, it means the best clustering result. When the variability in clusters is small, the *SC* usually helps to determine the number of clusters correctly (H. Řezanková a D. Húsek, 2012).

2. The modified approach

As we can see, all indexes have drawbacks with evaluation of clustering results in a large number of clusters and with increasing variability of data. They do not solve the problem of identifying the correct number of clusters.

These drawbacks are evident also in results of the analysis of the well-known datasets Iris, Glass and Vowel (http://archive.ics.uci.edu/ml/datasets.html). For example, both of the

indexes (1) and (4) – Dunn's coefficient *PC* and *SC* – for the Iris dataset show, that the best number of clusters is two (instead of the correct number of three). The silhouette coefficient evaluates one-element clusters by zero. That is why the silhouette coefficient achieves its highest value for two clusters in analyzed data (K. Zalik, 2011).

This is due to the fact that the clusters are overlapping and Dunn's coefficient is not able to recognize the correct structure of the clusters. A similar situation is observed in the evaluation of data clustering in sets Vowel and Glass, where the coefficients do not determine the number of clusters correctly.

Our task is to propose an alternative coefficient that will work better with the increasing variability of data and with the different number of clusters. We introduce the next modified approach, which is to combine two components into one index; in doing so, we use the harmonic mean. One of these 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 centre and points of this cluster should be minimal.

Joining two elements based on different approaches into one index helps to reduce disadvantages of both. Let him be the first element Dunn's coefficient (1).

We can distinguish two extreme situations:

1) completely fuzzy clustering, where all $u_{ij} = 1/k$ and then $PC = 1/k$;

2) hard clustering, where one $u_{ij} = 1$, all others $u_{ij} = 0$ and then $PC = 1$.

The second element is based on the hard clustering theory: we calculate the ratio of the distance minimum in case of k clusters to the distance minimum in case of a single cluster,

$$C = \frac{\sum d_{i,\min}}{d_{1,\min}}. \tag{5}$$

Here $\sum d_{i,\min}$ is the minimal sum on the Euclidean distances between points in the case of k clusters; and $d_{1,\min}$ is the minimal sum on the Euclidean distances between points in the case of a single cluster (when the dataset is one cluster, it means before clustering). The sum of $d_{i,\min}$ should be minimal for the best clustering, it means the minimal value of C should achieve the minimum for the best clustering.

We want to combine two parts in the aggregate function: one of them is Dunn's coefficient (the maximum value for the best clustering), and the second one should also strive to maximum; that's why we use $1 - C = N$ (which achieves the maximum value for the best clustering).

And now the optimization problem is to solve. It can be represented in the following way: it consists of minimizing a real function $f(PC, N)$ by systematically choosing input values from within an allowed set and computing the value of the function. This set we can describe as:

$$\begin{cases} \frac{1}{PC}+\frac{1}{N}\neq 0 \\ PC\neq 0 \\ N\neq 0 \\ N\neq 1 \\ PC\in\langle 0,1\rangle, N\in(0,1) \end{cases}. \tag{6}$$

In the case that $N = 0$ we have a dataset as a single cluster.

These conditions are met by function E,

$$E=1-\frac{2}{\left(\frac{1}{PC}+\frac{1}{N}\right)}\rightarrow \min. \tag{7}$$

It tends to its minimum for the best clustering, because the inverse values of the indexes of PC and N receive its maximum for the best clustering.

3. The experiments

Now, we introduce how our approach works in three known datasets: *Iris*, *Glass* and *Vowel* (http://archive.ics.uci.edu/ml/datasets.html). The datasets are described as follow:

The dataset *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 dataset *Glass*:

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

Dataset *Vowel*:

The number of cases is 528, 11 classes, the number of attributes is 10 (numeric, predictive attributes). The task of our experiments is to estimate the proper number of clusters in the actual datasets with help of the modified approach.

We applied the k-means clustering algorithm with Euclidean distance on these three different datasets. And we applied fuzzy k-means clustering to calculate the degree of belonging to clusters for every case. The reason to choose these approaches is because they are simple and well known by the scientific community.

To determine the correct number of clusters, we now calculate the *PC* (1) and PC_{mod} (2) coefficients, which is based on the membership degrees, the *SC* coefficient (4) based on the hard clustering technique and the modified coefficient *E* (7), which includes hard clustering (distances between points) and fuzzy clustering (memberships degrees) basis.

Iris dataset:

When the clusters are well-separated and the number of clusters is clearly defined, there is no problem with evaluating the right number of clusters. However, when data are overlapped, the concept of what is a cluster can be distinct for different methods, and consequently, the number of clusters too. The Iris data are overlapped and as we can see in Table 1, the *SC, PC* and PC_{mod} coefficients estimate the number of cluster as two. The modified approach shows the right number of clusters three. The behavior of all the coefficients is also described by Fig. 1.

Glass dataset:

The right number of clusters for these data is six. The *SC* and *PC* coefficients show that the best number of clusters is two (Table 2); the modified method shows six clusters (the minimum value of the coefficient for all possible solutions). We can see the behavior of the coefficients on Fig. 2, gradually declining curve tells us about the declining value of the coefficient with increasing number of clusters. The same case with the Silhouette coefficient declining curve means worse clustering results. And rising curve of the modified coefficient means worse clustering results.

Vowel dataset:

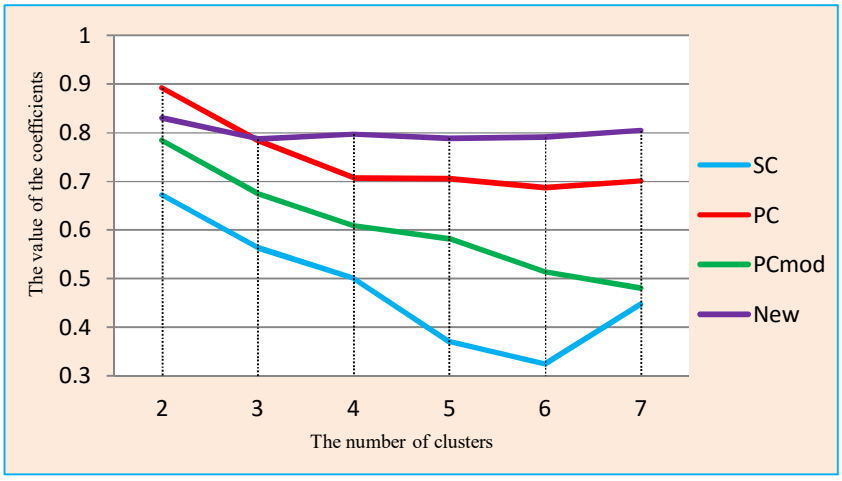
Now the result of the modified approach was compared with *SC,* PC_{mod} and *PC* coefficient on Vowel dataset (Table 3). The right number of clusters is eleven. The *SC* coefficient and the modified method estimate the right number of clusters eleven, and the *PC* coefficient shows the right number of clusters as three. The behavior of the coefficients is also described by Fig. 3.

Table 1. Dataset Iris; the results

The name of the coefficient / **The number of clusters**	***SC***	***PC***	PC_{mod}	***E* (New)**
2	**0,6723**	**0,8922**	**0,7844**	0,8302
3	0,5629	0,7833	0,6751	**0,7869**
4	0,5006	0,7067	0,6091	0,7978
5	0,3701	0,7057	0,5821	0,7892
6	0,3244	0,6878	0,5143	0,7914
7	0,4466	0,7009	0,4801	0,8046

Source: own calculation

Fig 1. Dataset Iris; the results



Source: own calculation

Table 2. Dataset Glass; the results

The name of the coefficient / The number of clusters	*SC*	*PC*	PC_{mod}	*E* (New)
2	**0,9421**	**0,9553**	**0,9107**	0,4631
3	0,5947	0,4233	0,1350	0,4701
4	0,5559	0,4259	0,2345	0,4999
5	0,5468	0,4103	0,2629	0,4971
6	0,4933	0,3318	0,1982	**0,4358**
7	0,2614	0,8023	0,7693	0,7300
8	0,2415	0,3861	0,2984	0,4917
9	0,3350	0,7832	0,7561	0,7151
10	0,2159	0,7723	0,7471	0,7230

Source: own calculation

Fig 2. Dataset Glass; the results

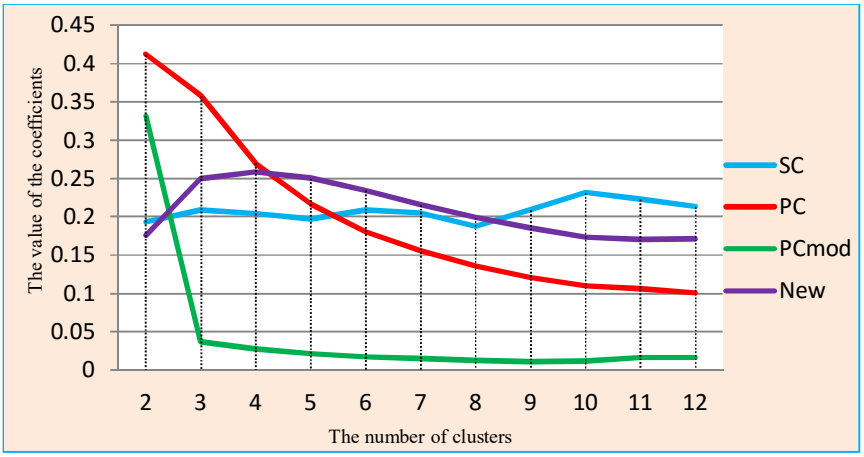
Source: own calculation

Table 3. Dataset Vowel; the results

The name of the coefficient / The number of clusters	*SC*	*PC*	PC_{mod}	*E* (New)
2	0,1930	**0,4123**	**0,3318**	0,1757
3	0,2085	0,3578	0,0368	0,2494
4	0,2044	0,2705	0,0273	0,2580
5	0,1973	0,2170	0,0213	0,2505
6	0,2085	0,1809	0,0172	0,2342
7	0,2047	0,1552	0,0145	0,2157
8	0,1880	0,1357	0,0123	0,1993
9	0,2086	0,1209	0,0111	0,1853
10	0,2315	0,1100	0,0112	0,1737
11	**0,2230**	0,1059	0,0165	**0,1701**
12	0,2134	0,1004	0,0161	0,1713

Source: own calculation

Fig 3. Dataset Vowel; the results



Source: own calculation

Conclusions

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

From experimental results can be drawn that modification method determine the number of clusters correctly. We plan to study this approach in other data sets.

Acknowledgment

This paper was created with the help of the Internal Grant Agency of University of Economics in Prague No. 6/2013 (Evaluation of results of cluster analysis in Economic problems.)

References

BRODOWSKI, S. (2011): A Validity Criterion for Fuzzy Clustering. *Computational collective intelligence: technologies and applications*, Berlin: Springer-Verlag, 113-122 p.

GUOJUN, G., CHAOGUN, M., JIANHONG, W. (2007): *Data Clustering theory, algorithms, and applications*. Philadelphia: ASA-SIAM Series on Statistics and Applied Probability

HÖPPNER, F., KLAWONN, F., RUNKLER, T. (1999): *Fuzzy Cluster Analysis*. London: Wiley

KRUSE, R., DÖRING, C., LESOT, M. J. (2007): *Fundamentals of fuzzy clustering, advances in fuzzy clustering and its applications*. London: Wiley.

LÖSTER, T., PAVELKA, T. (2013) Evaluating of the Results of Clustering in Practical Economic Tasks. *International Days of Statistics and Economics*. Slaný: Melandrium, 804-818 p.

LÖSTER, T. (2012): Kritéria pro hodnocení výsledků shlukování se známým zařazením do skupin založená na konfuzní matici. *Forum Statisticum Slovacum*, 8/7, 85-89 p.

REZAEE, B. (2010): A cluster validity index for fuzzy clustering. *Fuzzy sets and systems.* Amsterdam: Elsevier Science BV, 237-246 p.

ŘEZANKOVÁ, H., HÚSEK, D. (2012): Fuzzy Clustering: determining the number of clusters. *Computational Aspects of Social Networks* 2012 (CASoN)

XIE, N. (2011): A Classification of Cluster Validity Indexes Based on Membership Degree and Applications. *Web information systems and mining.* Berlin: Springer-Verlag Berlin, 43-50 p.

VALENTE DE OLIVEIRA, J. (2007): *Advances in Fuzzy Clustering and its Applications.* London: Wiley.

ZADEH, L. A. (1965): Fuzzy Sets. *Information and Control* 8/3. Elsevier, 338-353 p.

ZALIK, K. (2011): Validity index for clusters of different sizes and densities.*Computer science, artificial intelligence.* Amsterdam: Elsevier Science BV

The database of real datasets:

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

Contact

Mgr. Elena Makhalova
University of Economics, Prague
W. Churchill Sq. 4
130 67 Prague 3
elena.makhalova@vse.cz

doc. Ing. Iva Pecáková, CSc.
University of Economics, Prague
W. Churchill Sq. 4
130 67 Prague 3
iva.pecakova@vse.cz