VALIDATION APPROACHES FOR FCM ALGORITHM

Elena Rihova – David Riha

Abstract

Clustering techniques can be used to organize into groups based on similarities among the individual data. In other words, clustering techniques are tools for discovering the previously hidden structure in a set, where the objects from one cluster are as similar as possible and objects from different clusters are dissimilar as possible. 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) are examined. Also, their strengths and weaknesses are studied. And finally, the new coefficient for evaluating the fuzzy C-means clustering results is presented. The proposed coefficient is compared with a number of popular validation indices on nine datasets. The experimental results show that the effectiveness and reliability of the proposal is superior to other indices. The main advantage of this new coefficient is that, it works correct on data sets with large and small number of clusters. This characteristic of the new coefficient is very significant, as this algorithm require the number of clusters as an input, and the analysis result can vary greatly depending on the value chosen for this variable.

Key words: fuzzy clustering, evaluating results, partition

JEL Code: C18, C38, C69

Introduction

Cluster analysis is one of the methods of multivariate statistics, which is used to detect hidden data structure, found the groups with the most similar groups. That's why cluster analysis is very useful in different areas as psychology, sociology, medicine, marketing, and etc. The precondition of cluster analysis is the following: objects in data sets are more or less different from each other, hence there are several groups (clusters) of those objects, which can be defined with cluster analysis. The aim of the cluster analysis is to partition a given set of data or objects into clusters (subsets, groups, classes). This partition should have the following optimal ties: homogeneity within the clusters, i.e. data that belong to the same cluster should be as similar as possible, and Heterogeneity between clusters, i.e. data that belong to different clusters should be as different as possible. Based on this reasoning it is clear that cluster analyses explore similarities between objects in input data matrix with the help of similarity measures. It is also important to realize what type of clusters we want whether we want to create a certain number of clusters or a hierarchy of clusters.

Most of well-known validity indices (Dunn index, Xie-Beni index, Modified Dunn index) have drawbacks pertaining to the evaluation of clustering results in a large number of clusters, and with increasing variability of data. There are many cluster validity indices offers conclusion that there is not generally the best validity index, and existing cluster validity indices are not very efficient in estimation of clusters of different sizes and densities. (Žalik and Žalik, 2011) They do not solve the problem of identifying the correct number of clusters. (Wang, Zhang, 2007) The problem of identifying the correct number of clusters on data set as with the small, as with the large number of clusters be discussed and solved in current research. Therefore, there are two objectives. First of all to find out advantages and disadvantages of existing the most successful coefficients (Dunn index, Xie-Beni index, Modified Dunn index). And the second one is to propose new validity index, which has no those disadvantages.

In this paper, we present a new coefficient for validating fuzzy C-means clustering results, which works correctly on both data sets with a small number of clusters and on data sets with a large number of clusters (more than 5), assuming that the input data have a normal distribution. This new index combines 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 centre and objects of this cluster should be minimal.

1 Validation approaches

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. The compactness is a measure, which indicates the degree of similarity of data objects in a cluster, is calculated from membership values of data objects that are strongly enough associated to one cluster. (Žalik, 2010) 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 centres. 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. In this work are presented the classification of indices by Wang. (Wang, Zhang, 2007). Start with validity indices involving only the membership values, and the first index is Dunn's index.

1.1 Dunn's index or Partition coefficient (PC)

Bezdek tried to define a performance measure based on minimizing the overall content of pair wise fuzzy intersection in the partition matrix. Those validity index is partition coefficient (*PC*). The index is defined as:

$$PC = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{2} \qquad (1)$$

The *PC* index shows 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. (Wang, Zhang, 2007). In general, we find an optimal cluster number *k** by solving $\max_{2\leq k\leq n\text{ - }1} PC$ to produce the best clustering performance for the data set *X*. (Dunn, 1974)

1.2 Modified Dunn's index (PCmod)

The next validity index is proposed by Dave as a modification of the previous one:

$$PC_{\mathrm{mod}}(k) = 1 - \frac{k}{k-1}(1 - PC(k)) \qquad (2)$$

This index can take values $\langle 0,1 \rangle$, where k* is the optimal number of clusters. This cluster number k* is defined by solving of (Dave, Bhaswan, 1992):

$$\max_{2\le k\le \acute{n}=1} PC_{\mathrm{mod}}(k) \qquad (3)$$

When the variability in clusters is small, this modified Dunn's coefficient PC$_{\mathrm{mod}}$ usually determined the number of clusters correctly. (Rezankova, Dusek, 2012) When the cluster variability is greater, the normalized Dunn's coefficient usually achieved its highest value for the highest possible number of clusters. (Rezankova, Dusek, 2012)

1.3 Xie-Beni index (XB)

The second group of the validity indices is indices involving membership values and the data set. The most successful coefficient from this group is Xie Beni index. Those index is proposed by Xie and Beni with q = 2 (Xie, Beni, 1991) and modified by Pal and Bezdek (Bezdek, Pal. 1995) is defined as:

$$XB = \frac{J_m(u,v)/n}{Sep(C)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^2 \left\|x_j - x_i\right\|^2}{n.\min_{i,j}\left\|C_i - C_j\right\|^2} \qquad (4)$$

Like we can observe, this index includes two components: compactness in the numerator and separation, which is represented in denominator. Small value of compactness is evidence of a good partition, and high value of separation is evidence of a good partition. The optimal number of clusters k* can be find by solving $\min_{2\le k\le n-1} XB$ to produce the best clustering performance for the data set X. Unfortunately, this index has tendency to monotonically decrease with increasing number of clusters.

1.4 E index (E)

The theory of fuzzy clustering is based on the assumption the 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 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. We can distinguish two extreme situations: completely fuzzy clustering: all $u_{ij} = 1/k \Rightarrow PC = 1/k$ hard clustering: for one u_{ij} : u_{ij} = 1 and for all others: $u_{ij} = 0 \Rightarrow PC = 1$. 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). Traditional (hard) clustering considers the geometrical optionality of the data structure. (Žalik, 2010) We combine in the aggregate function two parts: 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 introduce the complement into one N (which achieves the maximum value for the best clustering). And now we have to solve the optimization problem. It can be represented in the following way:

$$f(x)=\frac{2}{\dfrac{1}{\dfrac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^{2}}+\dfrac{1}{1-\dfrac{\sum_{j=1}^{k}D_{k,\min}}{D_{1,\min}}}}\rightarrow\max \tag{5}$$

This function tends to its maximum for the best clustering because the inverse values of those two parts 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.

$$E=\frac{1}{k}\frac{2}{\left[\dfrac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^{2}\right]^{-1}+\left[1-\dfrac{\sum_{j=1}^{k}D_{k,\min}}{D_{1,\min}}\right]^{-1}} \tag{6}$$

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

2 Case study

To find validity indices, an extensive comparison of some of the abovementioned indices are conducted with the possible fuzzy C-means algorithm on an artificial number and a number of well-known data sets. In all experiments the distance function used is Euclidian distance. Choosing the best range in the number of clusters is quite a difficult problem. In this work, for all mentioned data sets, Bezdek's suggestion is adopted: $k_{\min}=2$ and $k_{\max}=\sqrt{n}$. (Bezdek, 1998)

For every data set 4 validity indices are calculated: Dunn's index, modified Dunn's index, the Xie-Beni index and finally the E coefficient. Afterwards, the success of each index is calculated. The main objective of this section is to compare the performance of the aforementioned indices in determining the optimal (actual) number of clusters. The real data sets are from the open internet-database UCI Machine Learning Repository.

All results are shown in Table 1. Summing up, we can determine in which of the abovementioned cases the indices worked incorrectly, in other words, to find out what affected it. As Saad stated: The disadvantages of the coefficients PC are the lack of direct connection to the geometrical structure of data, and its tendency to decrease with the number k. (Saad, 2012) When the variability in clusters is small, this normalized Dunn's coefficient usually determined the number of clusters correctly. (Rezankova, Dusek, 2012) When the cluster variability is greater, the normalized Dunn's coefficient usually achieved its highest value for the highest possible number of clusters. (Rezankova, Dusek, 2012) The Xie-Beni index is focused on two aspects of optionality: compactness and separation. As we know, compactness is a measure of the proximity of objects' vectors that share the same clusters as their centre. A small value of compactness indicates the of each cluster. Separation is a distance between two different clusters; hence separation indicates how two clusters are distinct and isolated from one another. A high value of separation shows that the clusters are well separated.

Tab. 1: Validity Indices

Data Sets	The Results of Optimal Number of Clusters				
	Optimal Number	PC	PC_{mod}	XB	E
Breast Tissue	6	2	2	2	6
Banknote Authentication	2	2	2	2	2
Climate Model Simulation Crashes	2	2	2	2	2
Fertility	2	2	2	2	2
Ionosphere	2	2	2	2	2
Parkinson Train Data	2	2	2	2	2
User Knowledge	4	2	3	4	4
Vowel	11	3	2	8	11
Wine	3	2	2	2	2
Sucessfulness,%	-	56	56	67	89

Source: author

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.

By analysing each data set, we can observe the different behaviour of each index: *PC*, *PC_{mod}*, *XB* and *E*. In the evaluation of fuzzy clustering results, it is necessary in the case of quantitative variables to compute multiple indices because there is no universal index to determine the correct number of clusters. In some cases, the local extreme exists. However, it is not very dependable for the determining the correct number of clusters, and does not show the correct number of clusters (as in many cases with *PC* index). Summary of evaluating of fuzzy clustering results are shown in Table 2.

Tab. 2: Summary of Evaluating Fuzzy Clustering Results

Characteristic	Affect	Does not Affect
The number of clusters		Does not affect the results (Bezdek, 1995)
The number of variable		Does not affect the results (Bezdek, 1998)
The Distance Measure		Does not affect the results (Bezdek, 1998; Oliviera, 2007)
Chosen algorithm	Affect the results (Oliviera, 2007)	
Overlapping[a]	Affect the results (Krishnapuram, Joshi, Nasraoui, and Yi,1993)	

Source: author

a)Krishnapuram, Joshi, Nasraoui, and Yi [16] recommended a value of overlapping(q) between 1 and 1, 5.

In most cases, the proposed index *E* works correctly: both for real and generated data sets (88% successfulness on real data sets and the same success rate on generated data sets), the worst results are shown by the *PC* index (63% successfulness on real data sets and a 32% success rate on generated data sets). Even better is the PC_{mod} index with success rates of 50% and 40%, respectively. The *XB* index showed better results than the *PC* and PC_{mod} indices. Its success rates are 75% and 84%, which is worse than the index results of *E*.

A significant merit in the evaluation of fuzzy clustering results is index *E*, which even determines the correct number of clusters in cases with large cluster overlap.

To sum up, based on above mentioned analysis, it can be stated that the newly proposed *E* index has significant merit in the problem of evaluating fuzzy clustering results. Using the *E* coefficient brings more reliable results than using the previously proposed indices (*PC*, PC_{mod}, and *XB*). The *XB* index is less successful because the best clustering results are achieved with an overlapping value of 1.5 (Krishnapuram,1993) and *XB* index has the best results with an overlapping value of 2 (Wang, Zhang, 2007). If the value of overlapping fundamentally affects the result of certain data sets, in that case *XB* index fails.

Testing many well-known previously formulated and proposed index on well-known data sets showed the superior reliability and effectiveness of the proposed index in comparison to other indices especially when evaluating partitions with clusters that widely differ in size or overlapping.

References

Bezdek, J., & Pal, N. (1995). Cluster validation with generalized Dunns indices. *Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*. doi:10.1109/annes.1995.499469

Bezdek, J. C. (1998) Fuzzy clustering. *Handbook of Fuzzy Computation*. doi:10.1887/0750304278/b438c73

Dave, R., & Bhaswan, K. (1992). Adaptive fuzzy c-shells clustering and detection of ellipses. *IEEE Transactions on Neural Networks, 3*(5), 643-662. doi:10.1109/72.159055

Dunn, J., C., (1974). A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*. Vol.3, no.3

Krishnapuram, R., Joshi, A., Nasraoui, O., & Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for Web mining. *IEEE Transactions on Fuzzy Systems, 9*(4), 595-607. doi:10.1109/91.940971

Oliveira, J. V., & Pedrycz, W. (2007). *Advances in fuzzy clustering and its applications*. Chichester: Wiley.

Rezankova, H., & Husek, D. (2012). Fuzzy clustering: Determining the number of clusters. *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)*. doi:10.1109/cason.2012.6412415

Saad, M., F., (2012). Validity Index and number of clusters. *2012 International Journal of Computer Science.* Issues, 9(1/3)

Wang, W., & Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy Sets and Systems, 158*(19), 2095-2117. doi:10.1016/j.fss.2007.03.004

Xie, X., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(8), 841-847. doi:10.1109/34.85677

Žalik, K. R. (2010). Cluster validity index for estimation of fuzzy clusters of different sizes and densities. *Pattern Recognition, 43*(10), 3374-3390. doi:10.1016/j.patcog.2010.04.025

Žalik, K. R., & Žalik, B. (2011). Validity index for clusters of different sizes and densities. *Pattern Recognition Letters, 32*(2), 221-234. doi:10.1016/j.patrec.2010.08.007

UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/databases.html

Contact

Elena Rihova
Skoda Auto University, o.p.s.
Na Karmeli 1457 , 293 01 Mladá Boleslav
Czech Republic
elena.rihova@savs.cz

David Riha
University of Economics, Prague
W. Churchilla 1938/4, 130 67 Praha 3 – Žižkov
Czech Republic
david.riha@vse.cz