e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2019. № 4
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2019. № 4

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

UDC 519.237.8

A COMPARATIVE STUDY OF CLUSTER VALIDITY INDICES

Kondruk N. E. – PhD, Associate professor, Associate Professor of Department of Cybernetics and Applied Mathematics, Uzhgorod National University, Uzhgorod, Ukraine.

ABSTRACT

Context. Cluster analysis is a method of classification without a teacher, that is, under conditions where preliminary information on the number of clusters is previously unknown. Therefore, defining the optimal number of clusters and test results of partitioning data sets is a complex task and requires further research.

Objective. The aim of paper is to study the efficiency of finding the natural data structure by crisp and fuzzy clustering validity indices, when the partition is realized by the clustering method based on fuzzy binary relations and conducting their comparative analysis.

Method. For partition of data sets the method based on fuzzy binary relation was used that provides an opportunity to simultaneously conduct crisp and fuzzy grouping of objects by different types of similarity measures. The distance similarity measure, which divides data into ellipsoid clusters, is used in the research. Two synthetic 2-dimensional data sets of a special type are generated, natural clustering of which is possible in two ways. Both sets are Gaussian. The most effective and frequently used groups of crisp and fuzzy cluster validity indices, which allow to find the optimal data set structure are described.

Results. The study of estimating the quality of clustering was conducted by means of method of fuzzy binary relations with six indices in two data sets. A comparative analysis of the effectiveness of determining the cluster and sub-cluster data structures by validity indices is made.

Conclusions. In practice, for some cluster validity indexes it is important to find not only the global extreme, but also local ones. They can fix the optimal sub-cluster data structure with less separation. To ensure the effectiveness of estimating the quality of clustering and to obtain objective results it is appropriate to take into account not only one index, but several of them. In perspective studies, creating a combined criterion that would join the most effective cluster validity indices by means of method based on fuzzy binary relations by a distance similarity measure is anticipated as well as implementing generalized cluster validity index for any similarity measures of fuzzy binary relations method; developing a software system that would ensure the automatic grouping of objects into clusters by concentric spheres, cones, ellipses without the preliminary determination of the clustering threshold.

KEYWORDS: cluster validity indices, cluster, clustering.

ABBREVIATIONS

B is a Bensaid partition Index;
BIC is a Bayesian information criterion;
CVI is a cluster validity index;
D is a Dunn's cluster validity index;
R is a Ren's cluster validity index;
SSWC is a Simplified silhouette width criterion;
XB is a Xie and Beni's Index.

NOMENCLATURE

C is a set of vector features;

$\overline{c_i}\left(c_1^i, c_2^i, \ldots, c_n^i\right)$ is a feature vector of clustering objects;

$d(\ )$ is a distance between clusters;

$diam(\ )$ is a cluster diameter;

m is a number of clustering objects;

m_i is a number of element in the *i*-th cluster;

n is a number of feature (dimension of the data sets);

O_i is an *i*-th clustering object;

r is a coefficient of fuzziness;

R is a fuzzy binary relation;

R^V is a fuzzy binary relation, which characterizes the distance between the feature vectors;

$\overline{v^*}$ is a center of the centroid;

$\overline{v_i}$ is a centroid of the *i*-th cluster;

z is a number of clusters;

μ_{ij} is a membership function of the *j*-th element to the *i*-th cluster;

$\mu_R(\ )$ is a membership function of the fuzzy binary relation R;

DOI 10.15588/1607-3274-2019-4-6

$\mu^*_{R^V}$ is a clustering threshold using a distance similarity measure;

ρ is a some distance metric;

$\| \|$ is the Euclidean metric.

INTRODUCTION

Cluster analysis is a method of classification without a teacher, that is, under conditions where preliminary information on the number of clusters is previously unknown. It consists in grouping data so that their similarities by some criteria within the same group would be strong and within different – weak. The resulting data partitioning can be crisp where each object belongs to only one cluster and fuzzy if it can simultaneously belong to different clusters with some measure. Obviously, there is a problem of determining the "natural" ("real") partition, which corresponds to the input data. The solution of this problem makes it possible to estimate the obtained clustering results and choose the "true" structure of the basic data.

This problem occurs for several reasons. Firstly, there is no optimal clustering algorithm. Thus, partition of the same data by different algorithms may give different results, which are not effective in all situations where clustering is needed. Consequently, the process of clustering should perform different partitions and identify the most optimal for each set and each individual task. Secondly most effective methods involve pre-setting the number of clusters in a situation where no such prior information is available. Therefore, in this situation, it's necessary to run an algorithm with different input parameters and then to estimate the resulting partition. The process of assessing how the resulting partition corresponds to the input data set determines the quality of the clustering or cluster validation. This problem is complex and requires further research.

So **the object of study** is the process of cluster validation and **the subject of study** is the cluster validity indices.

In addition, most developed clustering methods provide grouping of objects only by one similarity criterion, usually determined by some metric of distance. In this case, clusters of only ellipsoidal shape are formed. But there are many applied tasks, for example [1] where this kind of grouping objects is inadequate to the determined purpose and ineffective. The method based on fuzzy binary relations [2], makes it possible to group data according to different similarity measures, thus forming clusters in the shape of ellipses, cones and concentric spheres. Additionally, this method allows to conduct crisp and fuzzy clustering simultaneously, so data clustering.

Therefore, **the purpose of the work** is to study the efficiency of finding the natural data structure by crisp and fuzzy clustering validity indices, when the partition is realized by the clustering method based on fuzzy binary relations and conducting their comparative analysis.

To achieve the purpose of the work, the following problems shall be solved:

– to analyze existing cluster validity indices;

– to generate special research data sets;

– to implement the procedure for determining the optimal number of clusters for each data set;

– to make a comparative analysis of the obtained results.

1 PROBLEM STATEMENT

Let us consider the general problem of cluster analysis in the following statement.

Let there be given some objects $O_1,\ldots,O_m$, characterized by n quantitative features. Each object O_i, $i=\overline{1,m}$ definitely corresponds to the feature vector $\overline{c_i}\left(c_1^i,c_2^i,\ldots,c_n^i\right)$, $i=\overline{1,m}$.

It is necessary to find the optimal ("natural") number of clusters in partition of the given objects O_i, $i=\overline{1,m}$ into homogeneous "similarity" groups (clusters) by means of a method based on fuzzy binary relations and a distance similarity measure. Moreover, a comparative analysis of different types of crisp and fuzzy indices should be done to find the optimal number of clusters on a special data sets with a sub-cluster structure.

2 REVIEW OF THE LITERATURE

Estimating the quality of clustering is quite a complicated task. In the general statement, it is difficult to be formulated semantically and it is also difficult to find the appropriate mathematical model for it. But despite that, determining the quality of clustering is a very important task.

Nowadays, there are many cluster validity indices, which are very useful in practice as quantitative measures of determining the optimal structure of the partition. Thus, in paper [3] about 20 different most used CVIs are described. In paper [4] a comparative study was conducted investigating 40 validity indices (basic and modified ones) by their computational complexity on the basis of asymptotic analysis.

The set of papers [5–13] displays a comparison of crisp indices by the possibility of correct determination of the true (natural) number of clusters of different kinds of data sets and according to certain clustering methods. In [5] 8 crisp CVIs were compared on the basis of the *k*-means algorithm, in [6] we evaluated the internal and external indices applied to partitioning and density-based clustering algorithms, and the effectiveness of the silhouette index [7] has been stressed out. In [8, 9], a number of indexes was compared on Big Data and it was indicated that the silhouette and Dunn's index [10] give the best results for finding the optimal number of clusters. The works [11, 12] show the effectiveness of the BIC index [13].

In [14, 15] the optimal number of clusters for fuzzy clustering is researched. In [14] five fuzzy indexes were

DOI 10.15588/1607-3274-2019-4-6

used while solving the problems of image segmentation, in [15] a comparative analysis of ten fuzzy indexes was made. In these studies and that of [16], the effectiveness of fuzzy indices XB [17] and B [18] is shown.

Thus, it can be noted that different validity indices have some characteristics that may outweigh the others in certain classes of problems. In addition, it is difficult for the user to choose one of CVIs when there is such a variety of them. For this reason, the relevant problem of cluster analysis is to compare the characteristics and effectiveness of existing validation criteria.

In this paper, it is proposed to compare the efficiency of finding the natural number of clusters by CVIs for crisp and fuzzy clustering. A method that provides a crisp and fuzzy partition is the clustering method based on fuzzy binary relations [2, 19, 20]. It makes it possible to conduct partition the clustering of objects by clusters of different geometric shapes (ellipses [2], cones [19] and concentric spheres [20]), according to the chosen similarity measure.

3 MATERIALS AND METHODS

Clustering method.

As a clustering method, the crisp single-level method is chosen (subsection 6 in [2]) based on fuzzy binary relations. This method is centroidous and allows realize crisp and fuzzy grouping for the same data. These advantages of the method allow to conduct comparative analysis of CVIs for estimation of the crisp and fuzzy clustering results. At the same time, the similarity of objects according to some criterion is characterized by the fuzzy binary relation R on the set of vector features $C = \left\{\overline{c_i} \middle| i = \overline{1,m}\right\}$ with the membership function $\mu_R\left(\overline{c_i}, \overline{c_j}\right)$, where $\mu_R : C^2 \to [0,1]$. In particular, a qualitative change in the type of similarity measure of objects leads to changes in the geometry of clusters. So, in the studies [2, 19, 20], examples are provided showing possible types of similarity measures, which lead to the formation of ellipsoid, conical clusters and clusters in the form of concentric spheres. But the existing criteria for estimating the quality of the partition are developed mostly for ellipsoidal clusters, therefore, for the researching, the similarity measure of "distance" is chosen, which is described by the fuzzy binary relation R^V [1] and membership function

$$\mu_{R^V}\left(\overline{c_i}, \overline{c_j}\right) = e^{-\left(\frac{\rho\left(\overline{c_i}, \overline{c_j}\right)}{\Delta}\right)},$$

$$\Delta = \max_{i,j=\overline{1,m}} \rho\left(\overline{c_i}, \overline{c_j}\right).$$

Cluster validity indices.

The concept of cluster validity is based on compactness and separation.

Compactness means that the elements of the cluster should be closest to each other. This property is expressed, as a rule, by the distance between the internal elements of the cluster, the density in the middle of the cluster, or the volume occupied by the cluster in a multidimensional space.

Separation displays that the distance between different clusters should be as large as possible. It can be defined as the distance between the closest, most distant neighbours of different clusters, or centroids.

Let us consider the relative metrics that evaluate the quality of clustering by comparing several cluster structures.

A group of crisp cluster validity indices.

1. Dunn's index [10]:

$$D = \min_{\substack{i,j\in\overline{1,z},\\ i\neq j}} \left\{ \frac{d(i,j)}{\max\limits_{k\in\overline{1,z}} diam(k)} \right\}.$$

Where d is the distance between the clusters i and j, which can be defined as the distance between the two closest elements or between the centroids of clusters, $diam(k)$ is the cluster diameter that can be calculated as the maximum distance between the elements of the same cluster. Thus, the Dunn's index compares the intercluster distance with the diameter of the cluster. It is generally accepted that, if the cluster's diameter is smaller then the intercluster distance, then the clusters of the resulting structure are compact and separated. Hence, the higher the index value is, the better is the clustering. The Dunn's index is sensitive to noise and data emissions, in addition, it can not be used when each element forms an isolated cluster.

2. Simplified silhouette width criterion [7].

The silhouette of each cluster is determined. First, for each element cj of the cluster p its "silhouette" is determined:

$$S_{cj} = \frac{\min\limits_{q\in\overline{1,z},q\neq p} \left\|\overline{c_j}, \overline{v_q}\right\| - \left\|\overline{c_j}, \overline{v_p}\right\|}{\max\left(\left\|\overline{c_j}, \overline{v_p}\right\|, \min\limits_{q\in\overline{1,z},q\neq p} \left\|\overline{c_j}, \overline{v_q}\right\|\right)}.$$

Denominator is introduced for normalization. The high value of the index S_{cj} characterizes the better belonging of the element j to the cluster p. The evaluation of the whole cluster structure is defined as the average for all elements:

$$SSWC = \frac{1}{m}\sum_{j=1}^{m} S_{cj}.$$

Clearly, the best partitioning is achieved when the SSWC is maximal, it means minimizing the internal cluster distance by maximizing the external cluster distance.

3. Bayesian Information Index [13]:

DOI 10.15588/1607-3274-2019-4-6

$$BIC = $$

$$= \sum_{i=1}^{z}\left(m_i \ln\frac{m_i}{m} - \frac{m_i \cdot n}{2}\ln(2\pi) - \frac{m_i}{2}\ln|\Sigma_i| - \frac{m_i - z}{2}\right) -$$

$$-\frac{1}{2} z \ln m,$$

$$\Sigma_i = \frac{1}{m_i - z}\sum_{j=1}^{m_i}\left\|\overline{c_j} - \overline{v_i}\right\|^2.$$

The higher the BIC index value is, the better is the clustering model.

If this index for a particular problem increases or decreases monotonously, then the data set structure can be detected using a difference function [11]:

$$\mathrm{DiffFun(z)} = \mathrm{BIC(z+1)} + \mathrm{BIC(z-1)} - 2\cdot\mathrm{BIC(z)}.$$

This sequential difference method uses the previous, subsequent and current index values at the same time. The points of the minimum *DiffFun* will determine the optimal number of clusters.

The prospects of application and development of this index is confirmed in [11, 21].

It should be noted: if after grouping a single-element cluster is formed, then it is impossible to use this index.

Group of fuzzy CVIs.

1. The Xie-Beni index [17] tries to find the balance point between fuzzy cluster compactness and separation of clusters to obtain optimal clustering results:

$$XB(z,m) = \frac{\frac{1}{m}\sum_{i=1}^{z}\sum_{j=1}^{m}(\mu_{ij})^r\left\|\overline{c_j} - \overline{v_i}\right\|^2}{\min\limits_{i,j,i\neq j}\left\|\overline{v_i} - \overline{v_j}\right\|^2}.$$

In the formula, the numerator is the average distance from the different objects to the centroids (used to measure the clusters compactness), and the denominator is the minimum distance between any two centroids, (defines the clusters separation). Less criterion value corresponds to a better partition.

2. Bensaid's index [18]. However, Bensaid et al. found that the size of each cluster had a great influence on the Xie-Beni index and proposed a new index that was insensitive to the number of objects in each cluster. The Bensaid's index is defined as

$$B(z,m) = \sum_{i=1}^{z}\frac{\sum_{k=1}^{m}(\mu_{ik})^r\left\|\overline{c_k} - \overline{v_i}\right\|^2}{\sum_{k=1}^{m}\mu_{ik}\sum_{j=1}^{z}\left\|\overline{v_i} - \overline{v_j}\right\|^2}.$$

If the number of clusters will approach the number of objects, then index B, like the XB, will monotonously decrease to 0. Therefore, in this case, it loses the ability to determine optimal clustering. The minimum value is achieved where the clustering is better conducted.

3. Min Ren et al. (R) [22] proposed an index that eliminates this problem:

$$R(z,m) = \sum_{k=1}^{z}\frac{\sum_{i=1}^{m}(\mu_{ik})^r\left\|\overline{c_i} - \overline{v_k}\right\|^2 + (1/z)\left\|\overline{v_k} - \overline{v*}\right\|^2}{\sum_{i=1}^{m}\mu_{ik}\cdot(1/(z-1))\cdot\sum_{j=1}^{z}\left\|\overline{v_j} - \overline{v_k}\right\|^2}.$$

Lower index value indicates better clustering schema.

The described indexes have their disadvantages and advantages. In particular, only the BIC index can determine the optimal single-cluster structure, but only in the case of its nonmonotonicity.

The general procedure for determining the optimal number of clusters for a given data set:

1) setting different clustering thresholds by clustering method based on fuzzy binary relations, we obtain crisp or fuzzy partitioning of the data set and fix the value of the selected CVI;

2) finding local extrema (minima or maxima according to the chosen index) for the fixed values.

4 EXPERIMENTS

Data sets.

Given that the validity of machine clustering is determined by its compliance with the in-person classification, the verification of the number of clusters was carried out on the two-dimensional data. This provides an additional visual opportunity to evaluate the clustering result and to find the optimal data set structure.

For the experiment, Gaussian synthetic data, with and without noise were generated. They are of special type because the problem of finding the optimal number of clusters on them can be solved differently. So apart from the optimal cluster structure for them, there is also a "natural" subcluster one.

The first data set "Purity" (Fig. 3a) consists of 135 points and contains 3 groups of clusters without emissions with different densities. In addition, this set is also characterized by subcluster structure of 6 clusters. In general, the natural for "Purity" is the partition into 3 and 6 clusters.

The dataset Noise (Fig. 3b) contains 150 points located with approximately similar density. "Natural" for it is clustering on 2 and 5 clusters.

Scheme of experiment.

For the experiments, a computer program that implements clustering by a single-level method based on fuzzy binary relations with a distance similarity measure and validity indices D, SSWC, BIC, XB, B, R was developed.

The input information for grouping objects is the numerical values n, m, $\mu^*_{R^V}$ and the coordinates of the objects' feature-vectors $\overline{c_i}$. By setting different clustering thresholds a crisp and fuzzy clustering of data was performed and the values of the indices D, SSWC, BIC, XB, B, R were recorded. Then, for each index, global and local extrema were recorder.

DOI 10.15588/1607-3274-2019-4-6

5 RESULTS

The results obtained in accordance with the described procedure of section 4 for the datasets "Purity" and "Noise" are listed in Tables 1 and 2.

Table 1 – A fragment of the CVIs value for the "Purity" dataset

The number of clusters	The clustering threshold	The cluster validity indices					
		D	SSWC	BIC	XB	B	R
2	0.6	0.76	0.68	–1151	0.24	0.49	0.50
3	0.85	**2.02**	**0.87**	–1069	**0.21**	**0.21**	**0.49**
4	0.88	1.05	0.86	–1041	1.99	0.29	0.91
5	0.91	0.64	0.63	–1034	6.25	0.24	0.86
6	0.92	0.73	0.85	–984	4.95	0.22	1.15
7	0.93	0.55	0.73	–980	9.21	0.24	1.46
8	0.94	0.63	0.72	–976	13.57	0.30	2.14
9	0.95	0.57	0.71	–966	20.07	0.37	3.03
10	0.955	0.56	0.67	–959	24.79	0.29	2.65

In tables, colored cells display values of the indexes containing extremes (local or global), and in bold type are global extremes (Fig. 1).

From the obtained results, it is evident that all CVIs have determined the optimal number of clusters for both sets. The indices D, SSWC, XB also correctly defined the subcluster structure, although the index D determined one false partition into 8 clusters for the "Purity" set. Indices B and R correctly determined only the data partition with the largest isolation of clusters. The index R is insensitive to the smaller divisions of both sets. Index B correctly defined partition of the "Noise" set into 5 clusters, but also recorded the false partition into 7 clusters.

Since BIC monotonically increases for both datasets (Fig. 2a), a difference function was constructed (Fig. 2b).

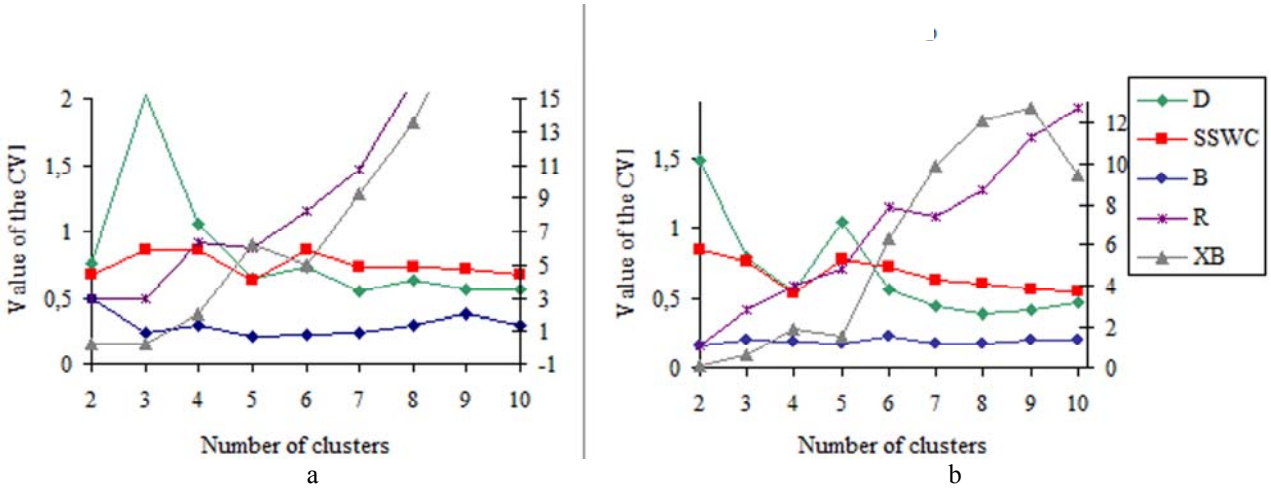
Fig. 2b shows that the minimum values of the DiffFun curve correctly determine the optimal cluster and subcluster structure for "Purity" and "Noise" data.

To verify the method based on fuzzy binary relations, the partitions of the "Purity" set for 3 and 6 clusters (Fig. 3a) and the "Noise" set for 2 and 5 clusters (Fig. 3b) are presented.

In Fig. 3 solid line conventionally labels the cluster structure of data sets, and dotted line – show a subcluster structure.

Table 2 – A fragment of the CVIs value for the "Noise" dataset

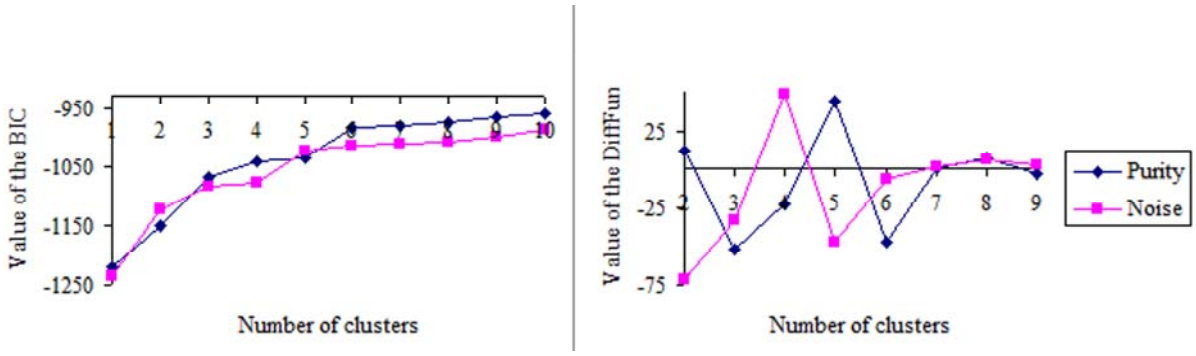
The number of clusters	The clustering threshold	The cluster validity indices					
		D	SSWC	BIC	XB	B	R
2	0.75	**1.49**	**0.84**	–1123	**0.08**	**0.16**	**0.16**
3	0.8	0.79	0.77	–1084	0.59	0.20	0.41
4	0.85	0.55	0.54	–1078	1.92	0.19	0.59
5	0.91	1.04	0.78	–1022	1.56	0.18	0.71
6	0.915	0.56	0.73	–1014	6.30	0.23	0.16
7	0.917	0.44	0.63	–1012	9.89	0.17	1.09
8	0.92	0.39	0.59	–1008	12.12	0.18	1.28
9	0.9203	0.42	0.56	–1003	12.76	0.20	1.64
10	0.9209	0.46	0.55	–987	9.41	0.21	1.85



a b

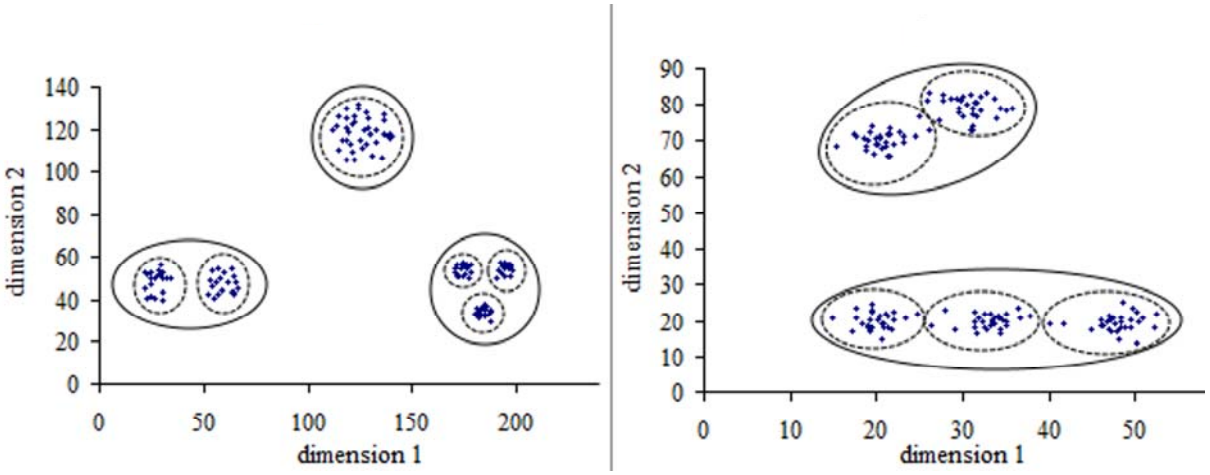
Figure 1 – A fragment of the geometric interpretation of the D, SSWC, B, R, XB values for the "Purity" (a) and the "Noise" (b) datasets

DOI 10.15588/1607-3274-2019-4-6



a b

Figure 2 – A fragment of the geometric interpretation of the BIC (a) and the DiffFun (b) values for the both dataset



a b

Figure 3 – Geometric interpretation of clustering the “Purity” set (a) and the “Noise” set (b) by means of a method based on fuzzy binary relations

6 DISCUSSION

All studied CVIs correctly identified the optimal partition of “Purity” and “Noise” data with the highest cluster separation (isolation). The SSWC, BIC, D, and XB indices also discovered sub-cluster structure in both datasets with a lower separation index. It should be noted that index D also identified one false split of the “Purity” set.

Despite the results of experiments, to ensure the effectiveness of clustering quality estimation and to obtain objective results, it is appropriate to take into account not only one index, but several of them.

To determine the natural number of clusters, most methods involve identifying only the global or first extremum of the corresponding index. In the conducted research it is shown that the identifying of local extremes for the SSWC, BIC, D and XB indices is also important. It’s can record the optimal sub-cluster structure with less separation.

CONCLUSIONS

The problem of studying the effectiveness of determining the natural structure of data by crisp and fuzzy cluster validity indices and conducting their comparative analysis is being solved.

The scientific novelty of the obtained results is the following.

1. Set of studies were analysed and the most effective CVIs were identified in a crisp and fuzzy group.

2. A study of crisp and fuzzy validity indices applied to clustering method based on fuzzy binary relations by a distance similarity measure has been conducted for the first time.

3. A comparative analysis of the obtained data considering the determining natural cluster and subcluster structure for synthetically generated Gaussian dataset with and without noise is made. The best indices were – SSWC, BIC, XB. Index D has also shown good results, but it can also detect false structures.

4. In practice, for the SSWC, BIC, D, and XB indices it is important not only to find the global extremum but also local ones. They record the optimal sub-cluster data structure with a smaller separation.

The practical significance of obtained results is that the created software implements the clustering method based on fuzzy binary relations and different types of CVIs. It makes it possible to conduct crisp and fuzzy data clustering and to determine dataset natural structure. Experiments showed its effectiveness in solving some classes of cluster analysis problems. Datasets with cluster

DOI 10.15588/1607-3274-2019-4-6

and sub-cluster structure were also generated in the work. They can be used by other researchers in their experiments.

This work is a continuation and development of previous research [2, 19, 20]. In the future it is supposed to use the obtained results for:

– development of a combined criterion that would join the SSWC, XB, BIC indices and would determine optimal structure clustering applied to a method based on fuzzy binary relations by distance similarity measure;

– construction of a generalized cluster validity index for any similarity measures of fuzzy binary relations method;

– development of a decision support system that would ensure the automatic grouping of objects by concentric spheres, cones, ellipses clusters without the preliminary determination of the clustering threshold.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Uzhgorod National University "Development of mathematical models and methods for data processing and data mining" (state registration number 0115U004630).

REFERENCES

1. Kondruk N. E. Decision Support System for automated diets, *Management of Development of Complex Systems*, 2015, Issue. 23(1), pp. 110–114.
2. Kondruk N. Clustering method based on fuzzy binary relation, *Eastern-European Journal of Enterprise Technologies*, 2017, Vol. 2, No. 4(86), pp. 10–16. DOI: 10.15587/1729–4061.2017.94961
3. Ghosh A., De Rajat K. Identification of certain cancer-mediating genes using Gaussian fuzzy cluster validity index *Journal of biosciences*, 2015, Vol. 40, No. 4, pp. 741–754. DOI: 10.1007/s12038-015-9557-x
4. Vendramin L., Campello R. J. G. B., Hruschka E. R. Relative clustering validity criteria: A comparative overview, *Statistical analysis and data mining: the ASA data science journal*, 2010, Vol. 3, No. 4, pp. 209–235. DOI: 10.1002/sam.10080
5. Meroufel H., Mahi H., Farhi N. Comparative Study between Validity Indices to Obtain the Optimal Cluster, *International Journal of Computer Electrical Engineering*, 2017, Vol. 9, No. 1, pp. 1–8. DOI: 10.17706/IJCEE.2017.9.1.343-350
6. Tomasini C. A Study on the Relationship between Internal and External Validity Indices Applied to Partitioning and Density-based Clustering Algorithms, *ICEIS (1)*, 2017, pp. 89–98. DOI: 10.5220/0006317000890098
7. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, 1987, Vol. 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7
8. Luna-Romera J. M., García-Gutiérrez J., Martínez-Ballesteros M., Santos J. An approach to validity indices for clustering techniques in Big Data, *Progress in Artificial Intelligence*, 2018, Vol. 7, No. 2, pp. 1–14. DOI: 10.1007/s13748-017-0135-3
9. Sivogolovko E. V. Methods for assessing the quality of clear clustering, *Computer tools in education*, 2011, No. 4 (96), pp. 14–31.
10. Dunn J. C. Well-separated clusters and optimal fuzzy partitions, *Journal of cybernetics*, 1974, Vol. 4, No. 1, pp. 95–104. DOI: 10.1080/01969727408546059
11. Estiri H., Omran B. A., Murphy S. N. Kluster: An Efficient Scalable Procedure for Approximating the Number of Clusters in Unsupervised Learning, *Big data research*, 2018, Vol. 13, pp. 38–51. DOI: 10.1016/j.bdr.2018.05.003
12. Zhao Q., Hautamaki V., Fränti P. Knee point detection in BIC for detecting the number of clusters, *International conference on advanced concepts for intelligent vision systems: ACIVS 2008, LNCS 5259*, 2008, pp. 664–673. DOI: 10.1007/978-3-540-88458-3_60
13. Fraley C., Raftery A. E. How many clusters? Which clustering method? Answers via model-based cluster analysis, *The computer journal*, 1998, Vol. 41, No. 8, pp. 578–588. DOI: 10.1093/comjnl/41.8.578
14. Gamarra D. F. T. Fuzzy image segmentation using validity indexes correlation, *International Journal of Computer Science and Information Technology*, 2015, Vol. 7, No. 3, pp. 15–26.
15. Zhou K., Ding S., Fu C., Yang S. L. Comparison and weighted summation type of fuzzy cluster validity indices, *International Journal of Computers Communications & Control*, 2014, Vol. 9, No. 3, pp. 370–378. DOI: 10.15837/ijccc.2014.3.237
16. Meroufel H., Mahi H., Farhi N. Comparative Study between Validity Indices to Obtain the Optimal Cluster, *International Journal of Computer Electrical Engineering,* 2017, Vol. 9, No. 1, pp. 1–8.
17. Xie X. L., Beni B G. A validity measure for fuzzy clustering, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1991, Vol. 13, No. 8, pp. 841–847. DOI: 10.1109/34.85677
18. Bensaid A. M., Hall L. O., Bezdek J. C. Validity-guided (re)clustering with applications to image segmentation, *IEEE Transactions on Fuzzy Systems*, 1996, Vol. 4, No. 2, pp. 112–123. DOI: 10.1109/91.493905
19. Kondruk, N. E. Some methods of automatic grouping of objects, *Eastern-European Jornal of Enterprise Technologies*, 2014, Vol. 2, Issue № 4 (68), pp. 20–24.
20. Kondruk N. Use of length-based similarity measure in clustering problems, *Radio Electronics, Computer Science, Control*, 2018, No. 3 (46), pp. 98–105. DOI: 10.15588/1607-3274-2018-3-11.
21. Teklehaymanot F. K., Muma M., Zoubir A. M. Novel Bayesian cluster enumeration criterion for cluster analysis with finite sample penalty term, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, 15–20 April 2018*, Calgary, ICASSP, 2018, pp. 4274–4278. DOI: 10.1109/ICASSP.2018.8462172
22. Ren M., Peiyu L., Zhihao W., Jing Y. A self-adaptive fuzzy c-means algorithm for determining the optimal number of clusters, *Computational intelligence and neuroscience*, 2016, Vol. 2016, pp. 1–12.

Received 09.06.2019.
Accepted 25.09.2019.

DOI 10.15588/1607-3274-2019-4-6

УДК 519.237.8

ПОРІВНЯЛЬНЕ ДОСЛІДЖЕННЯ ПОКАЗНИКІВ ОЦІНКИ ЯКОСТІ КЛАСТЕРИЗАЦІЇ

Кондрук Н. Е. – канд. техн. наук, доцент, доцент кафедри кібернетики і прикладної математики Ужгородського національного університету, Ужгород, Україна.

АНОТАЦІЯ

Актуальність. Кластерний аналіз є методом класифікації без учителя, тобто в умовах коли попередня інформація про кількість кластерів заздалегідь невідома. Тому, знаходження оптимальної кількості кластерів і перевірка результатів розбиття наборів даних є складною задачею і потребує додаткових досліджень.

Метою дослідження є вивчення ефективності знаходження природної структури даних чіткими та нечіткими індексами якості кластеризації реалізованої методом кластеризації основаним на нечітких бінарних відношеннях та проведення їх порівняльного аналізу.

Методи. Для розбиття наборів даних використано метод заснований на нечітких бінарних відношеннях, який дозволяє одночасно проводити чітку та нечітку кластеризацію об'єктів за різними видами мір подібності. В роботі використана міра подібності «відстань», яка розбиває дані на еліпсоїдні кластери. Згенеровано два синтетичні набори двовимірних даних спеціального виду, природна кластеризація яких можлива двома способами. Обидва набори є гаусівськими. Описано найбільш ефективні та використовувані групи чітких та нечітких індексів якості кластеризації, що дозволяють виявити оптимальну структуру даних.

Результати. Проведено дослідження оцінки якості кластеризації методом заснованим на нечітких бінарних відношеннях шістьма індексами на двох наборах даних. Зроблено порівняльний аналіз ефективності визначання індексами якості кластерної та підкластерної структури даних.

Висновки. На практиці для деяких індексів достовірності розбиття важливим є знаходження не тільки глобального екстремуму, а й локальних. Вони можуть фіксувати оптимальну підкластерну структуру даних із меншим показником розділення. Для забезпечення ефективності оцінки якості кластеризації та отримання об'єктивного результату доцільним є врахування не одного індексу, а декількох. В перспективних дослідженнях передбачається побудова комбінованого критерію, що поєднував би найефективніші індекси оцінки кластеризації методом заснованим на нечітких бінарних відношеннях за відстаневою мірою подібності; створення узагальненого індексу якості кластеризації за будь-якою мірою подібності методу нечітких бінарних відношень; розробка програмної системи, що забезпечить автоматичне групування об'єктів на кластери концентричними сферами, конусами, еліпсами без попереднього визначення порогу кластеризації.

КЛЮЧОВІ СЛОВА: індекси оцінки якості кластеризації, кластер, кластеризація.

УДК 519.237.8

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ ПОКАЗАТЕЛЕЙ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ

Кондрук Н. Э. – канд. техн. наук, доцент, доцент кафедры кибернетики и прикладной математики Ужгородского национального университета, Ужгород, Украина.

АННОТАЦИЯ

Актуальность. Кластерный анализ является методом классификации без учителя, то есть в условиях, когда предварительная информация о количестве кластеров заранее неизвестна. Поэтому, нахождение оптимального количества кластеров и проверка результатов разбиения наборов данных является сложной задачей и требует дополнительных исследований.

Целью исследования является изучение эффективности нахождения естественной структуры данных четкими и нечеткими индексами качества кластеризации основанной на нечетких бинарных отношениях и проведения их сравнительного анализа.

Методы. Для разбиения наборов данных использован метод основанный на нечетких бинарных отношениях, который позволяет одновременно проводить четкую и нечеткую кластеризацию объектов по различным видам мер сходства. В работе использована мера сходства «расстояние», которая разбивает данные на эллипсоидные кластеры. Сгенерировано два синтетические наборы двумерных данных специального вида естественная кластеризация которых возможна двумя способами. Оба набора являются гауссовскими. Описаны наиболее эффективные и используемые группы четких и нечетких индексов качества кластеризации, позволяющие выявить оптимальную структуру данных.

Результаты. Проведено исследование оценки качества кластеризации методом основанным на нечетких бинарных отношениях шестью индексами на двух наборах данных. Сделан сравнительный анализ эффективности определения индексами качества кластерной и подкластерной структуры данных.

Выводы. На практике для некоторых индексов достоверности разбиения важным является нахождение не только глобального экстремума, но и локальных. Они могут фиксировать оптимальную подкластерную структуру данных с меньшим показателем разделения. Для обеспечения эффективности оценки качества кластеризации и получения объективного результата целесообразно учитывать не один индекс, а несколько. В перспективных исследованиях предполагается построение комбинированного критерия, сочетающего эффективные индексы оценки кластеризации методом основанным на нечетких бинарных отношениях по степени сходства «расстояние»; создание обобщенного индекса качества кластеризации по любой степени сходства метода нечетких бинарных отношений; разработка программной системы, которая обеспечит автоматическое группирование объектов на кластеры концентрическими сферами, конусами, эллипсами без предварительного определения порога кластеризации.

КЛЮЧЕВЫЕ СЛОВА: индексы оценки качества кластеризации, кластер, кластеризация.

DOI 10.15588/1607-3274-2019-4-6

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Kondruk N. E. Decision Support System for automated diets / N. E. Kondruk // Management of Development of Complex Systems. – 2015. – Issue 23 (1). – P. 110–114.
2. Kondruk N. Clustering method based on fuzzy binary relation / N. Kondruk // Eastern-European Journal of Enterprise Technologies. – 2017. – Vol. 2, № 4 (86). – P. 10–16. DOI: 10.15587/1729–4061.2017.94961
3. Ghosh A. Identification of certain cancer-mediating genes using Gaussian fuzzy cluster validity index / A. Ghosh, K. De Rajat // Journal of biosciences. – 2015. – Vol. 40, № 4. – P. 741–754. DOI: 10.1007/s12038-015-9557-x
4. Vendramin L. Relative clustering validity criteria: A comparative overview / L. Vendramin, R. J. G. B. Campello, E. R. Hruschka // Statistical analysis and data mining: the ASA data science journal. – 2010. – Vol. 3, №. 4. – P. 209–235. DOI: 10.1002/sam.10080
5. Meroufel H. Comparative Study between Validity Indices to Obtain the Optimal Cluster / H. Meroufel, H. Mahi, N. Farhi // International Journal of Computer Electrical Engineering. – 2017. – Vol. 9, № 1. – P. 1–8. DOI: 10.17706/IJCEE.2017.9.1.343-350
6. Tomasini C. A Study on the Relationship between Internal and External Validity Indices Applied to Partitioning and Density-based Clustering Algorithms / C. Tomasini // ICEIS (1). – 2017. – P. 89–98. DOI: 10.5220/0006317000890098
7. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis / P. J. Rousseeuw // Journal of computational and applied mathematics. – 1987. – Vol. 20. – P. 53–65. DOI: 10.1016/0377-0427(87)90125-7
8. An approach to validity indices for clustering techniques in Big Data / [J. M. Luna-Romera, J. García-Gutiérrez, M. Martínez-Ballesteros, J. Santos] // Progress in Artificial Intelligence. – 2018. – Vol. 7, № 2. – P. 1–14. DOI: 10.1007/s13748-017-0135-3
9. Sivogolovko E. V. Methods for assessing the quality of clear clustering / E. V. Sivogolovko // Computer tools in education. – 2011. – № 4 (96). – P. 14–31.
10. Dunn J. C. Well-separated clusters and optimal fuzzy partitions / J. C. Dunn // Journal of cybernetics. – 1974. – Vol. 4, № 1. – P. 95–104. DOI: 10.1080/01969727408546059
11. Estiri H. Kluster: An Efficient Scalable Procedure for Approximating the Number of Clusters in Unsupervised Learning / H. Estiri, B. A. Omran, S. N. Murphy // Big data research. – 2018. – Vol. 13. – P. 38–51. DOI: 10.1016/j.bdr.2018.05.003
12. Zhao Q. Knee point detection in BIC for detecting the number of clusters / Q. Zhao, V. Hautamaki, P. Fränti // International conference on advanced concepts for intelligent vision systems: ACIVS 2008, LNCS 5259, 2008. – P. 664–673. DOI: 10.1007/978-3-540-88458-3_60
13. Fraley C. How many clusters? Which clustering method? Answers via model-based cluster analysis / C. Fraley, A. E. Raftery // The computer journal. – 1998. – Vol. 41, №. 8. – P. 578–588. DOI: 10.1093/comjnl/41.8.578
14. Gamarra D. F. T. Fuzzy image segmentation using validity indexes correlation/ D. F. T. Gamarra // International Journal of Computer Science and Information Technology. – 2015. – Vol. 7, №. 3. – P. 15–26.
15. Zhou K. Comparison and weighted summation type of fuzzy cluster validity indices / K. Zhou, S. Ding, C. Fu, S. L. Yang // International Journal of Computers Communications & Control. – 2014. – Vol. 9, № 3. – P. 370–378. DOI: 10.15837/ijccc.2014.3.237
16. Meroufel H. Comparative Study between Validity Indices to Obtain the Optimal Cluster / H. Meroufel, H. Mahi, N. Farhi // International Journal of Computer Electrical Engineering. – 2017. – Vol. 9, № 1. – P. 1–8.
17. Xie X. L. A validity measure for fuzzy clustering / X. L. Xie, B G. Beni // IEEE Transactions on Pattern Analysis & Machine Intelligence. – 1991. – Vol. 13, № 8. – P. 841–847. DOI: 10.1109/34.85677
18. Bensaid A. M. Validity-guided (re)clustering with applications to image segmentation/ A. M. Bensaid, L. O. Hall, J. C. Bezdek // IEEE Transactions on Fuzzy Systems. – 1996. – Vol. 4, № 2. – P. 112–123. DOI: 10.1109/91.493905
19. Kondruk N. E. Some methods of automatic grouping of objects / N. E. Kondruk // Eastern-European Jornal of Enterprise Technologies. – 2014. – Vol. 2, Issue № 4 (68). – P. 20–24.
20. Kondruk N. Use of length-based similarity measure in clustering problems/ N. Kondruk // Radio Electronics, Computer Science, Control. – 2018. – № 3 (46). – P. 98–105. DOI: 10.15588/1607-3274-2018-3-11.
21. Teklehaymanot F. K. Novel Bayesian cluster enumeration criterion for cluster analysis with finite sample penalty term / F. K. Teklehaymanot, M. Muma, A. M. Zoubir // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, 15–20 April 2018 – Calgary: ICASSP, 2018. – P. 4274–4278. DOI: 10.1109/ICASSP.2018.8462172
22. A self-adaptive fuzzy c-means algorithm for determining the optimal number of clusters / [M. Ren, L. Peiyu, W. Zhihao, Y. Jing // Computational intelligence and neuroscience. – 2016. – Vol. 2016. – P. 1–12.

DOI 10.15588/1607-3274-2019-4-6