A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 3(322)2016
http://dx.doi.org/10.18778/0208-6018.322.07

*Dorota Rozmus**

COMPARISON OF THE ACCURACY OF THE PROBABILISTIC DISTANCE CLUSTERING METHOD AND CLUSTER ENSEMBLES

Abstract: High accuracy of results is a very important aspect in any clustering problem *t* determines the effectiveness of decisions based on them. Therefore, literature proposes methods and solutions that aim to give more accurate and stable results than traditional clustering algorithms (e.g. *k*-means or hierarchical methods). Cluster ensembles (Leisch 1999; Dudoit, Fridlyand 2003; Hornik 2006; Fred, Jain 2002) or the distance clustering method (Ben-Israel, Iyigun 2008) are the examples of such solutions. Here, we carry out an experimental study to compare the accuracy of these two approaches.

Keywords: clustering, accuracy, distance clustering method, cluster ensemble
JEL: C38

1. INTRODUCTION

Recently, distance clustering methods have become increasingly popular, together with cluster ensemble methods for machine learning. Probability distance clustering method (in abbreviation d-clustering) is a relatively new clustering algorithm that was introduced by Ben-Israel and Iyigun (2008). This method is an iterative, distribution free, probabilistic, clustering method. D-clustering assigns units to a cluster according to the probability of their belonging to the cluster. The cluster ensemble approach can be defined generally as follows: given multiple partitions of the data set, find combined clustering with a better quality and stability.

The main aim of this research is to compare the accuracy of the distance clustering method and cluster ensembles.

2. DISTANCE CLUSTERING METHOD

D-clustering is a non hierarchical algorithm that assigns units to clusters according to their belonging probability to the cluster. The authors themselves

* Ph.D., Department of Economic and Financial Analysis, Faculty of Finance and Insurance, University of Economics in Katowice, drozmus@ue.katowice.pl

describe this method as:[1] *Given clusters, their centers and the distances of data points from these centers, the probability of cluster membership at any point is assumed inversely proportional to the distance from (the center of) the cluster in question.*

The algorithm of this method can be described as follows:

Initialization: given data ***D***, any two points $\mathbf{c}_1, \mathbf{c}_2$ and $\varepsilon > 0$.

Iteration:

Step 1. **Compute** distances $d1(\mathbf{x})$, $d2(\mathbf{x})$ for all $\mathbf{x} \in \mathbf{D}$ according to the expression:

$$d_k(\mathbf{x}_i, \mathbf{c}_k) = \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{x}_i - \mathbf{c}_k \right\|, \quad k = 1, 2. \tag{1}$$

Step 2. **Update** the centers $\mathbf{c}_1^+, \mathbf{c}_2^+$:

$$\mathbf{c}_k^+ = \sum_{i=1,\ldots,N} \left(\frac{u_k(\mathbf{x}_i)}{\sum_{j=1,\ldots,N} u_k(\mathbf{x}_j)} \right) \mathbf{x}_i, \tag{2}$$

where:

$$u_k(\mathbf{x}_i) = \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x_i}, \mathbf{c}_k)}, \tag{3}$$

and:

$$p_k(\mathbf{x}_i) = \frac{\prod_{j \neq k} d_j(\mathbf{x}_i)}{\sum_{t=1}^{K} \prod_{j \neq t} d_j(\mathbf{x}_i)}, \quad k = 1, 2. \tag{4}$$

Step 3. **If** $\left\| \mathbf{c}_1^+ - \mathbf{c}_1 \right\| + \left\| \mathbf{c}_2^+ - \mathbf{c}_2 \right\| < \varepsilon$, **stop**

else **return** to step 1.

[1] Quoted from: Iyigun, Ben-Israel (2007: 1).

3. BAGGING IN TAXONOMY

For the first time the bagging idea in taxonomy was used by Leisch (1999). The steps of the algorithm are as follows:

1. Construct B bootstrap samples.
2. Run the base clustering method (e.g. k-means) on each set, resulting $B \times K$ centers, where K is number of centers used in the base method.
3. Combine all centers into a new data set.
4. Run a hierarchical cluster algorithm on this set resulting in a usual dendrogram.
5. This dendrogram is cut at a particular level which is defined by a researcher in order to get groups of centers that are similar.
6. Each observation from the original data set is assigned to the group with the nearest center.

The next solution was proposed by Dudoit and Fridlyand (2003). The steps are as follows:

1. Apply the partitioning clustering procedure to the original learning set to obtain cluster labels for each observation $\mathbf{x}_i$.
2. Form the B bootstrap samples.
3. Apply the clustering procedure to the original data set and to the bootstrap samples.
4. Permute the cluster labels assigned observations from the bootstrap learning sets so that there is a maximum overlap with the labels assigned to the observations from the original data set.
5. In order to get the final clustering for each observation use the *majority vote*, that is, the cluster label corresponding to $\mathbf{x}_i$ is argmax $1 \le k \le K$.

In the bagging method by Hornik (2006) the first step is the construction of B bootstrap samples and running a partitioning cluster algorithm on them in order to get single partitions that are members of the cluster ensemble. The final partition is obtained with the *optimization approach*, which formalizes the natural idea of describing consensus clusterings as the ones that "optimally represent the ensemble" by providing a criterion to be optimized over a suitable set C of possible consensus clusterings. If *dist* is an Euclidean dissimilarity measure and $(c_1,\ldots,c_B)$ are the elements of the ensemble, the problem is solved by means of *least squares* consensus clustering (generalized means):

$$\sum_{b=1}^{B} dist(c,c_b)^2 \Rightarrow \min_{c \in C} . \tag{5}$$

4. CLUSTER ENSEMBLE BASED ON A CO-OCCURRENCE MATRIX

Fred and Jain (2002) proposed the idea of combining clustering results performed by transforming data partitions into a co-occurrence matrix that shows coherent associations. This matrix is then used as a distance matrix to extract the final partitions. The subsequent steps of the algorithm are as follows (Fig. 1):

Step One – split. For a fixed number of cluster ensemble members C cluster the data using e.g. the k-means algorithm, with different clustering results obtained by random initializations of the algorithm.

Step Two – combine. The underlying assumption is that patterns belonging to a "natural" cluster are very likely to be co-located in the same cluster among these C different clusterings. So taking the co-occurrences of pairs of patterns in the same cluster as votes for their association, the data partitions produced by C runs of k-means are mapped into a $n \times n$ co-occurrence matrix:

$$co\_assoc(a,b) = votes_{ab}, \tag{7}$$

where $votes_{ab}$ is the number of times when the pair of patterns (a, b) is assigned to the same cluster among the C clusterings.

Step Three – merge. In order to recover final clusters, apply any cluster algorithm over this co-occurrence matrix treated as the dissimilarity representation of the original data.

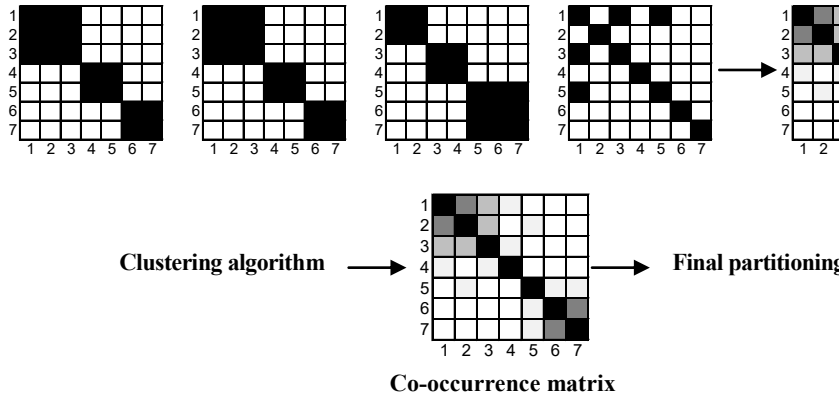

Figure 1. Construction of the co-occurrence matrix and their final partitioning

Source: own work.

5. NUMERICAL EXPERIMENTS

In the study, artificially generated data sets are taken from mlbench library from **R**. Their short characteristics are shown in Table 1 and their structure is shown in Figure 2.

Table 1. Characteristics of used data sets

Data set	Number of objects	Number of variables	Number of classes
Cassini	500	2	3
Cuboids	500	3	4
Ringnorm	500	2	2
Shapes	500	2	4
Smiley	500	2	4
Spirals	500	2	2
Threenorm	500	2	2
2dnormals	500	2	2

Source: own work.

In the case of the bagging methods, the final partition was obtained as follows: in Leisch method, 50 bootstrap samples with *k*-means were used as a base clustering method and the final clusters were obtained with the Ward method; in Dudoit, Fridlyand and Hornik method, 50 bootstrap samples were used and *k*-means were applied to them.

The co-occurrence matrix was constructed on 10 components generated by the *k*-means algorithm and its further partitioning was conducted by *k*-means.

The accuracy of the methods was examined by the Rand Index. All computations were made in **R**.

Looking at the results (Fig. 3), we can see that in the case of the *2dnormals* and *Spirals* sets d-clustering renders very similar results to cluster ensembles. For the *Cuboids* and *Smiley* this method performs worse than cluster ensembles. For the *Cassini* data set it can be seen that d-clustering is comparable with the Leisch and Hornik bagging method. In the case of *Ringnorm*, d-clustering delivers only slightly worse results than the Hornik bagging method. For the *Shapes* data set only the co-occurrence method is worse than d-clustering, whereas the remaining methods give the same results. Finally, for the *Threenorm* data set d-clustering is comparable only with the Leisch bagging method.

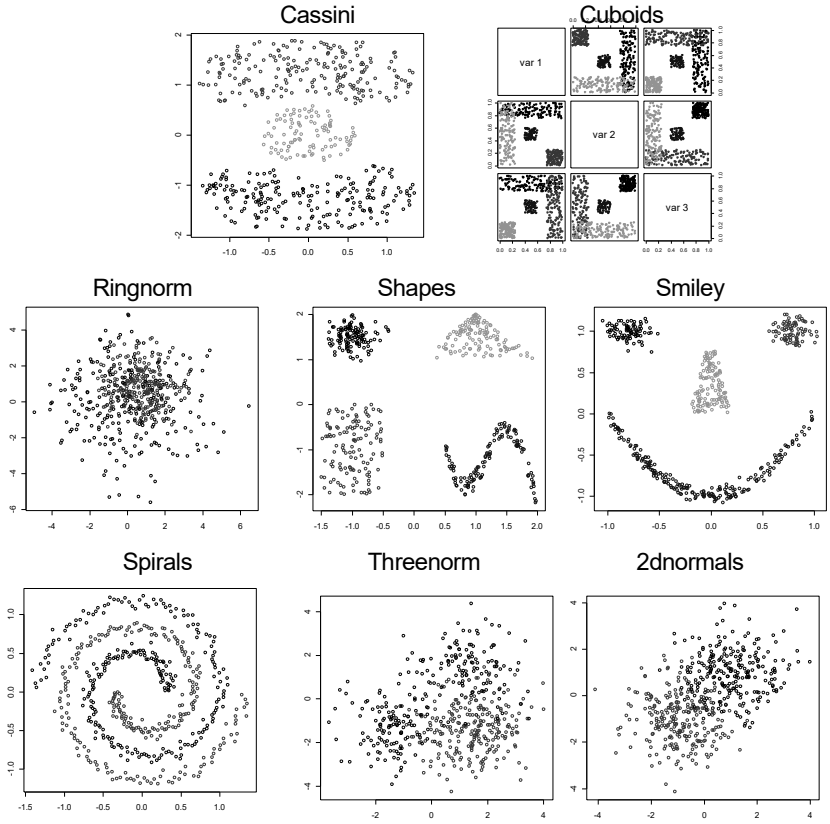

Figure 2. Structure of the used data sets

Source: own work on base of **R** program.

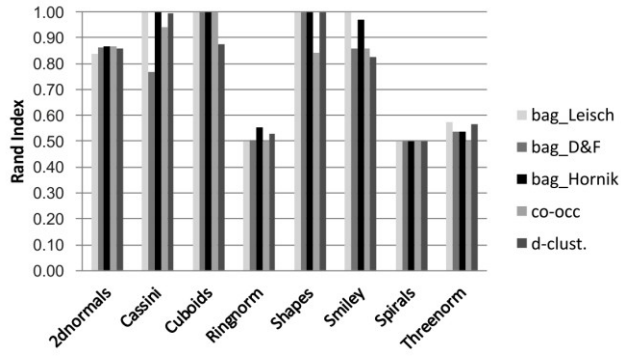

Figure 3. Accuracy of d-clustering and cluster ensemble methods

Source: own work.

6. CONCLUSIONS

The main aim of the study was to compare the accuracy of the d-clustering method and cluster ensembles. Looking at the results we cannot say unambiguously that d-clustering is better than cluster ensembles. In some cases, it performed better than some of the variants of cluster ensembles (e.g. *Ringnorm*, *Threenorm*), while in others d-clustering yielded worse results than cluster ensembles (e.g. *Cuboids*, *Smiley*) or the result were very comparable (*2dnormals*, *Spirals*).

Looking at the structure of used data sets we can conclude that d-clustering is a better choice than cluster ensemble, especially in case of sets with overlapping groups. In cases when data has clearly separated group, it will be less risky to choose any variant of bagging, especially that proposed by Leisch.

Recently there have been more and more discussions about stability[2] of clustering algorithm that can be an extra criterion in choosing the best partitioning method. Therefore, further experiments should be carried out in order to compare stability of these two approaches. Finally, a criterion that combines accuracy and stability, may be the right way in deciding which algorithm should be chosen.

REFERENCES

Ben-Israel A., Iyigun C. (2008), *Probabilistic d-clustering*, "Journal of Classification", 25(1), pp. 5–26.

Dudoit S., Fridlyand J., (2003), *Bagging to improve the accuracy of a clustering procedure*, "Bioinformatics", vol. 19, no. 9, pp. 1090–1099.

Fred A., Jain A. K. (2002), *Data clustering using evidence accumulation*, "Proceedings of the Sixteenth International Conference on Pattern Recognition", pp. 276–280.

Hornik K., (2005), *A CLUE for CLUster ensembles*, "Journal of Statistical Software", 14, pp. 65–72.

Leisch F. (1999), *Bagged clustering*, "Adaptive Information Systems and Modeling in Economics and Management Science", Working Papers, SFB, 51.

Dorota Rozmus

PORÓWNANIE DOKLADNOŚCI METODY ODLEGŁOŚCI PROBABILISTYCZNEJ I PODEJŚCIA ZAGREGOWANEGO W TAKSONOMII

Streszczenie: Stosowanie metod taksonomicznych w jakimkolwiek zagadnieniu grupowania wymaga jednocześnie zapewnienia wysokiej dokładności wyników podziału. Ona bowiem warunkuje skuteczność wszelkich decyzji podjętych na podstawie uzyskanych rezultatów. Dlatego też w literaturze wciąż proponowane są nowe rozwiązania, których zadaniem jest poprawa dokładności grupowania w stosunku do tradycyjnie stosowanych metod (np. *k*-średnich, hierarchicznych). Przykładami mogą tu być metody polegające na zastosowaniu podejścia

[2] Stability means invariance of results to the random initializations of the algorithm.

zagregowanego (Leisch 1999; Dudoit, Fridlyand 2003; Hornik 2006; Fred, Jain 2002), czy niedawno zaproponowana metoda odległości probabilistycznej (Ben-Israel, Iyigun 2008).

Głównym celem artykułu jest porównanie dokładności omawianej metody z dokładnością podejścia zagregowanego w taksonomii.

Słowa kluczowe: grupowanie, dokładność, metoda odległości probabilistycznej, podejście zagregowane w taksonomii