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 285, 2013

*Dorota Rozmus**

COMPARISON OF ACCURACY OF SPECTRAL CLUSTERING AND CLUSTER ENSEMBLES BASED ON CO-OCCURRENCE MATRIX

Abstract. High accuracy of the results is very important task in any grouping problem (clustering). It determines effectiveness of the decisions based on them. Therefore in the literature there are proposed methods and solutions that main aim is to give more accurate results than traditional clustering algorithms (e.g. *k*-means or hierarchical methods). Examples of such solutions can be cluster ensembles or spectral clustering algorithms. Here, we carry out an experimental study to compare accuracy of spectral clustering and cluster ensembles.
Key words: spectral clustering, cluster ensembles, clustering.

I. INTRODUCTION

Recently, spectral methods have become increasingly popular, together with cluster ensemble methods for machine learning. They may be applied especially in cases where simple algorithms such as *k*-means fail. Spectral clustering uses eigenvectors from spectral decomposition of an affinity matrix derived from the data. Then the dominant eigenvalues and the corresponding eigenvectors are used for clustering the original data. Several algorithms have been proposed in the literature (Kannan et al. 2004, Ng et al. 2001, Shi and Malik 2000), each using the eigenvectors in slightly different ways. In this paper, we focus on the method proposed by Ng et al. (2001). Cluster ensemble approach can be defined generally as follows: given multiple partitions of the data set, find a combined clustering with a better quality. Here we consider cluster ensembles based on co-occurrence matrix (Fred 2002; Fred and Jain 2002). The main aim of this research is to compare accuracy of spectral clustering and cluster ensembles.

* Ph.D., Department of Statistics, University of Economics, Katowice.

II. CLUSTER ENSEMBLE BASED ON CO-OCCURRENCE MATRIX

Generally, the main source of the idea of co-occurrence matrix is proposed by Pekalska and Duin (2000) dissimilarity based approach in discriminant analysis. In the conventional way of learning from examples of observations the classifier is built in a feature space. However, an alternative way can be found by constructing decision rules on dissimilarity representations. In such a recognition process each object is described by its distances (or similarities) to the rest of training samples. Classifier is built on this dissimilarity representation that is on a matrix describing similarities between used examples of objects for training.

Based on this Fred and Jain (2002) proposed the idea of combination of clustering results performed by transforming data partitions into a co-occurrence matrix which shows coherent associations. This matrix is then used as a distance matrix to extract the final partitions. The particular steps of the algorithm are as follows:

First step – 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.

Second step – 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-association matrix:

$$co\_assoc(a,b) = votes_{ab}, \tag{1}$$

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.

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

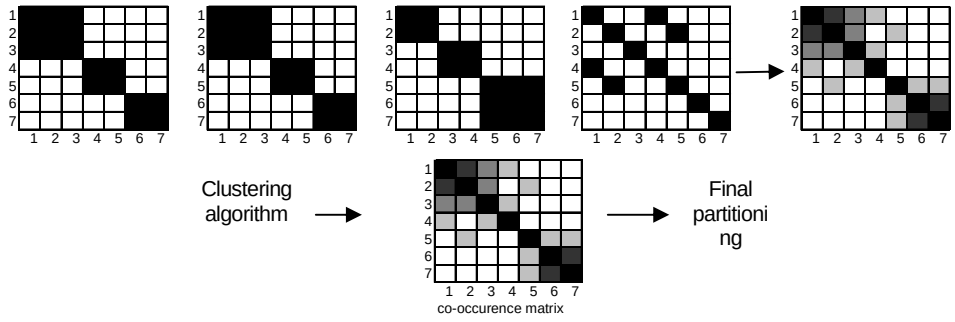

Figure 1. Construction of the co-occurrence matrix and their final partitioning

Source: own work.

III. SPECTRAL CLUSTERING

Spectral clustering is a promising alternative to classical algorithms. In this approach one uses the top eigenvectors of a matrix created by some distance measure between the points. Then the top k eigenvectors (where k is the number of clusters to be found) of the affinity matrix are used to form an $n \times k$ matrix $\boldsymbol{Y}$. Treating each row of this matrix as a data point, clustering algorithm (usually k-means) is finally used to cluster the points. The algorithm can be described as follows (Ng et al. 2001). Given data set $G = \{x_1,...,x_n\} \in R^l$ that should be clustered into k groups:

1. Form the affinity matrix $A \in R^{n \times n}$ whose elements are defined as:

$$A_{ij} = \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right), \tag{2}$$

for $i \neq j$ and $A_{ii} = 0$. σ is a scaling parameter chosen by the user.

2. Define $\boldsymbol{D}$ as a diagonal matrix with (i, i)-element equal to the sum of the i-th row of $\boldsymbol{A}$ and construct the matrix:

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{3}$$

3. Find the first k eigenvectors of $\boldsymbol{L}$ and form the matrix:

$$\mathbf{Z} = \left[\mathbf{z_1},...,\mathbf{z_k}\right] \in R^{n \times k}. \tag{4}$$

by stacking them in columns.

4. Renormalize each row of $\boldsymbol{Z}$ to have unit length according to transformation:

$$y_{ij} = z_{ij} / \left(\sum_j z_{ij}^2\right)^{1/2}. \tag{5}$$

5. Treating each row of $\boldsymbol{Y}$ as a point in R^k, cluster them into k groups by means of k-means (or another algorithm).

6. In order to get final partition assign each original point x_i to the j-th cluster if i-th row of the matrix $\boldsymbol{Y}$ was assigned to j-th cluster.

IV. EMPIRICAL EXPERIMENTS

In order to compare accuracy of the methods there was used measure based on Rand index:

$$Acc = \frac{1}{Z}\sum_{z=1}^{Z} R(P_z, P^T), \tag{6}$$

where:

Z – number of partitions,

R – Rand index,

P_z – clusters get on the base of z-th partition.

In the research there were used artificial generated data sets taken from `mlbench` library from **R**. Their short characteristics are shown in the Table 1 and their structure is shown on Fig. 2.

Table 1. Characteristics of used data sets

Data set	# of objects	# of variables	# 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.

The co-occurrence matrix was constructed on 10 components with two algorithms, i.e. *k*-means and *c*-means and its further partitioning was made by *k*-means, *c*-means, pam and clara algorithms. Matrix $\boldsymbol{Y}$ in spectral approach was clustered by *k*-means algorithm. Each approach was used 50 times and its accuracy was then examined by measure given by formula 6.

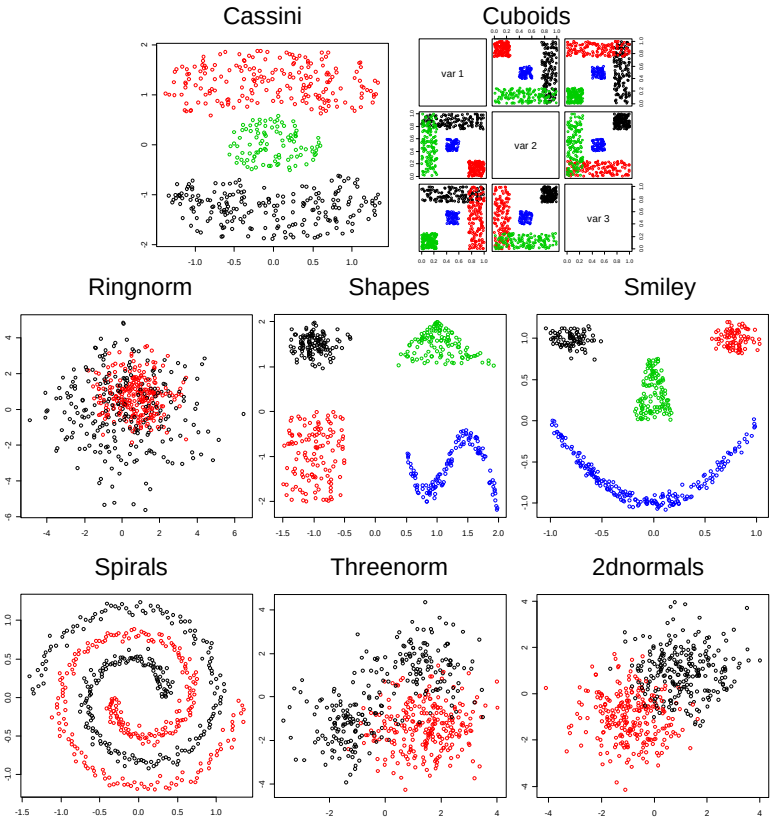

Figure 2. Structure of the used data sets

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

In the case when co-occurrence matrix was constructed by *k*-means algorithm (Fig. 3) it can be noticed that for *Cuboids*, *Shapes*, *Threenorm* and *2dnormals* data sets spectral clustering (specc) gives higher accuracy than cluster ensemble with *k*-means used for partitioning of the co-occurrence matrix (kmeans_kmeans) but lower when *c*-means (kmeans_cmeans), pam (kmeans_pam) and clara (kmeans_clara) algorithms used for its further partitioning. For *Smiley* data set spectral clustering is more accurate than kmeans_kmeans and kmeans_cmeans but less accurate than kmeans_pam and kmeans_clara. Spectral clustering gives the highest accuracy among all the methods for two data sets, i.e. *Cassini* and *Spirals*. The lowest accuracy in comparison with all variants of aggregated approach we can notice only for *Ringnorm* data set.

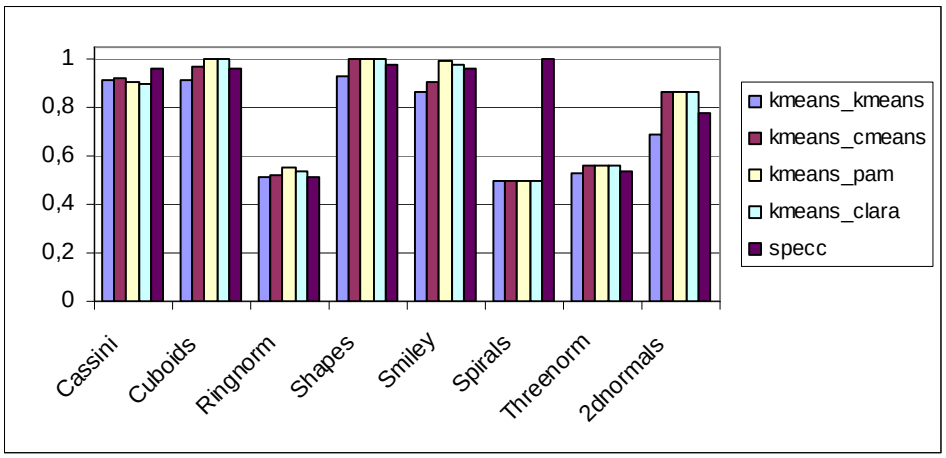

Figure 3. Accuracy of spectral clustering and cluster ensemble based on co-occurrence matrix with *k*-means used for its construction

Source: own work.

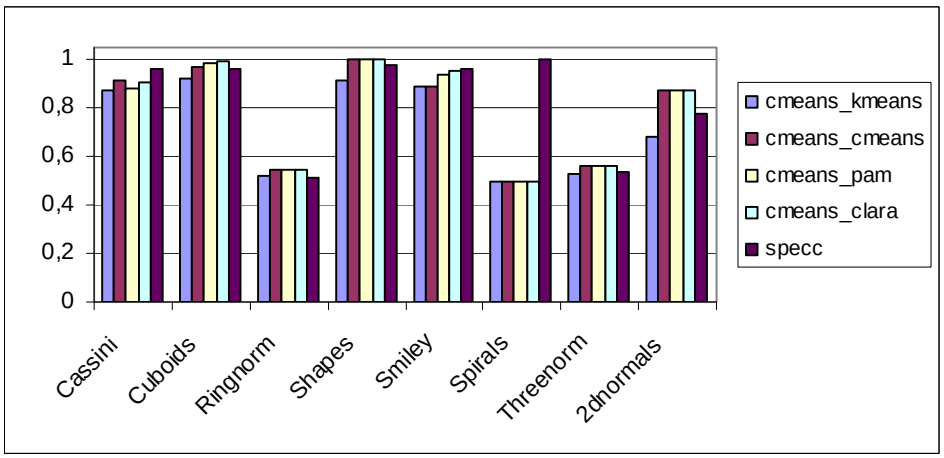

Figure 4. Accuracy of spectral clustering and cluster ensemble based on co-occurrence matrix with *c*-means used for its construction

Source: own work.

For co-occurrence matrix with *c*-means used for its construction, similarly as before for *Cuboids*, *Shapes*, *Threenorm* and *2dnormals* data sets, its further partitioning with *k*-means algorithm (cmeas_kmeans) gives lower accuracy than spectral clustering; but using *c*-means (cmeas_cmeans), pam (cmeas_pam) and clara (cmeas_clara) algorithms bring higher accuracy than spectral clustering. Spectral clustering is the most accurate in comparison with all variants of

aggregated approach for *Cassini*, *Spirals* and *Smiley* data sets. Once again spectral clustering is the least accurate for *Ringnorm* data set.

V. CONCLUSIONS

To sum up all the numerical experiments of this research it can be said that using *k*-means algorithm for further partitioning of the co-occurrence matrix usually leads to lower accuracy in comparison with spectral clustering. Using clara, pam and often also *c*-means algorithm for majority of the data sets in cluster ensemble approach gives better results than spectral approach. For data sets with the structure similar to *Spirals* data set spectral clustering seems to be the best solution.

REFERENCES

Fred A. (2002), Finding consistent clusters in data partitions, in Roli F., Kittler J., editors, *Proceedings of the International Workshop on Multiple Classifier Systems*, pages: 309 – 318.

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

Kannan R., Vempala S., Vetta A. (2004), On clustering – good, bad and spectral, *Journal of the ACM*, Vol. 51, No.3, pages 497–515.

Ng A. Y., Jordan M. I., Weiss Y. (2001), On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems*, pages 849–856.

Pekalska E., Duin R. P. W. (2000), Classifiers for dissimilarity-based pattern recognition, in Sanfeliu A., Villanueva J. J, Vanrell M., Alquezar R., Jain A. K. and Kittler J., editors, *Proceedings of the Fifteenth International Conference on Pattern Recognition*, pages 12–16, IEEE Computer Society Press, Los Alamitos.

Shi J., Malik J. (2000), Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, 888–905, (http://www-2.cs.cmu.edu/ »jshi/Grouping/.)

Dorota Rozmus

PORÓWNANIE DOKŁADNOŚCI TAKSONOMII SPEKTRALNEJ ORAZ ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA MACIERZY WSPÓŁWYSTĄPIEŃ

Stosując metody taksonomiczne w jakimkolwiek zagadnieniu klasyfikacji ważną kwestią jest zapewnienie wysokiej poprawności wyników grupowania. Od niej bowiem zależeć będzie skuteczność wszelkich decyzji podjętych na tej podstawie. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod. Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz algorytmy spektralne.

Głównym celem tego artykułu jest porównanie dokładności zagregowanych i spektralnych algorytmów taksonomicznych. W badaniach pod uwagę wzięta zostanie tylko specyficzna klasa metod agregacji, która oparta jest na macierzy współwystąpień (Fred, Jain 2002). Natomiast jako algorytm spektralny zastosowana będzie metoda zaproponowana przez Ng i in. (2001).