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 235, 2010

*Dorota Rozmus**

COMPARISON OF CLUSTERING ACCURACY IN ENSEMBLE APPROACH BASED ON CO-OCCURENCE DATA

ABSTRACT. Ensemble approach has been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. Recently, analogous techniques for cluster analysis have been suggested. Research has proved that, by combining a collection of different clusterings, an improved solution can be obtained.

In the traditional way of learning from the data set the classifiers are 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 a matrix showing the similarities or distances to the rest of training samples.

This research has focused on exploiting the additional information provided by a collection of diverse clusterings to generate a co-association (co-occurrence) matrix (Fred and Jain, 2002). Taking the co-occurrences of pairs of patterns in the same cluster as votes for their association, the data partitions are mapped into a co-association matrix of patterns. This n × n matrix represents a new similarity measure between patterns. The final data partition is obtained by clustering this matrix. In the experiments, the behavior of partitions built on co-occurrence data with different clustering methods is studied.

Key words: Cluster analysis, Cluster ensemble, Co-association matrix, (Dis)similarity representation.

I. INTRODUCTION

Ensemble techniques based on aggregated models have been successfully applied in supervised learning (classification, discriminant analysis) and regression in order to improve the accuracy and stability of classification and regression algorithms (Breiman, 1996, Tsymbal *et al.* 2003). The concept of aggregation can be described as follows: instead of using one model for prediction, use many different models and then combine many theoretical values of dependent variable with some aggregation operator. In classification the most often used operator is majority voting: an observation is classified to the most often chosen class, in regression we often calculate mean of the theoretical values of dependent variable. The presumption in this approach is that using many models instead of one will give

* Ph.D., Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

better results. Among the most popular methods there are eg. *bagging* based on bootstrap samples (Breiman, 1996) and *boosting* based on giving higher weights to the wrong classified examples (Freund, 1990).

Recently, ensemble approach for cluster analysis has been suggested in order to increase the classification accuracy and robustness of the clustering solutions. The main idea of aggregation is to combine outputs of several clusterings. The problem of clustering fusion can be defined generally as follows: given multiple partitions of the data set, find a combined clustering with a better quality. Recently several studies on clustering combination methods have pioneered a new area in the conventional taxonomy (Fred, 2002; Fred and Jain, 2002; Jain *et al*., 1999; Strehl and Gosh, 2002). There are several possible ways to use the idea of ensemble approach in the context of unsupervised learning: (1) combine results of different clustering algorithms; (2) produce different partitions by resampling the data, such as in bootstrapping techniques, eg. *bagging*; (3) use different subsets of features (that can be disjoint or overlapping); (4) run a given algorithm many times with different parameters or initializations.

II. THE ALGORITHM

In this research the last approach is taken to some extend. Generally, this research has three sources. The first is proposed by Pekalska and Duin (2000) dissimilarity based approach. 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. The second source is proposed by Fred and Jain (2002) 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 third source is provided by Kuncheva, Hadjitodorov and Todorova (2006) research where they got very promising results with dissimilarity representation treated as a data matrix. Here similar split and merge approach is used. 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 eg. 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 taxonomic algorithm over this co-association matrix treated as dissimilarity representation of the original data.

The idea of ensemble approach is here used in the phase of preparing the data that should be clustered not in clustering immediately. There is prepared a special data description by using an aggregated approach and this matrix is then clustered by single run of the clustering algorithm (illustrated on Figure1.).

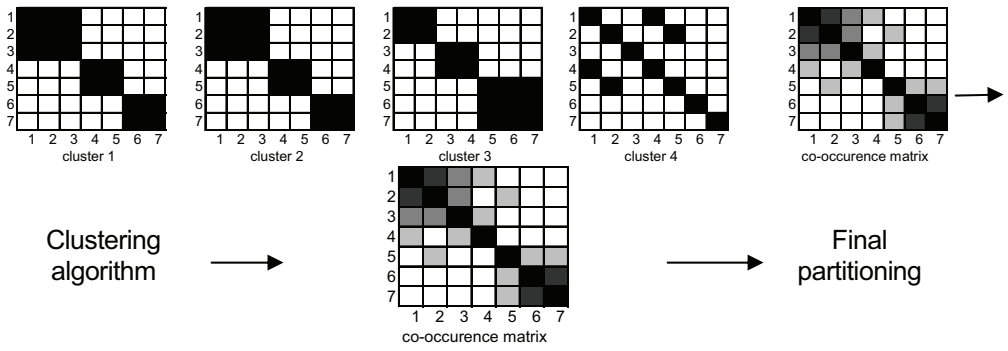

Fig. 1. Construction of the co-occurence matrix and their final partitioning

Source: own work.

III. EMPIRICAL RESULTS

The aim of empirical experiments was to compare the ability to recognize the right class structure of the proposed cluster ensemble approach with using two cluster algorithms for their construction and their later partitioning with different algorithms.

In the step of building the co-occurrence matrix there were used k-means algorithm and developed by Bezdek (1981) c-means, which is the fuzzy version of the k-means algorithm. The number of cluster ensemble members C was set to equal 10, and the values of parameters c and k were equal to the number of class.

Assuming the right number of classes for *c* and *k* parameters is often used approach by the researchers from the field of taxonomy.

Among used methods for further partitioning of the co-occurence matrix there were: *k*-means, *c*-means, partition among medoids (*k*-medoids), which is a more robust version of *k*-means (Rousseeuw and Kaufmann, 1990) and clara, which compared to other partitioning methods such as *k*-medoids can deal with much larger datasets (Rousseeuw and Kaufman,1990). As a measure of correctness of the algorithm a popular Rand Index was used (Rand, 1971). Most computations were made in **R**[1]. Among used algorithms there were kmeans from library *stats*, cmeans from library *e1071*, pam and clara algorithms from library *cluster*.

In the research there were used real and artificial generated data sets, their short characteristics are shown in the Table 1.

Tab. 1. Data sets used in the experiments

Data set	# of objects	# of variables	# of class
Boston	506	13	4
Ecoli	336	8	8
Glass	214	10	6
Cassini	500	2	3
Cuboids	500	3	4
Shapes	500	2	4
Smiley	500	2	4

Source: own work.

The first three are real data and the rest are artificial generated sets. Among the real data there were used sets that are usually applied in classification for model building and its evaluation. These are data sets where the object's class adherence is known. This information is treated as an *a priori* information about the number of clusters. Such an approach is also often used by researches from the field of taxonomy. The presented real data sets are usually used in benchmarking researches in classification, and they are made available by UCI Repository (Blake *et al*., 1988). Among artificial generated data, there were sets that are usually used in comparative studies in taxonomy. Their structure is presented on Figure 2. The *Cassini, Shapes* and *Smiley* are two dimensional data sets with clearly separated classes, the *Cuboids* is a problem where there are uniformly distributed on a 3-dimensional space within 3 cuboids and a small cube in the middle of them.

[1] This program is free available on web site: http://www.r-project.org.

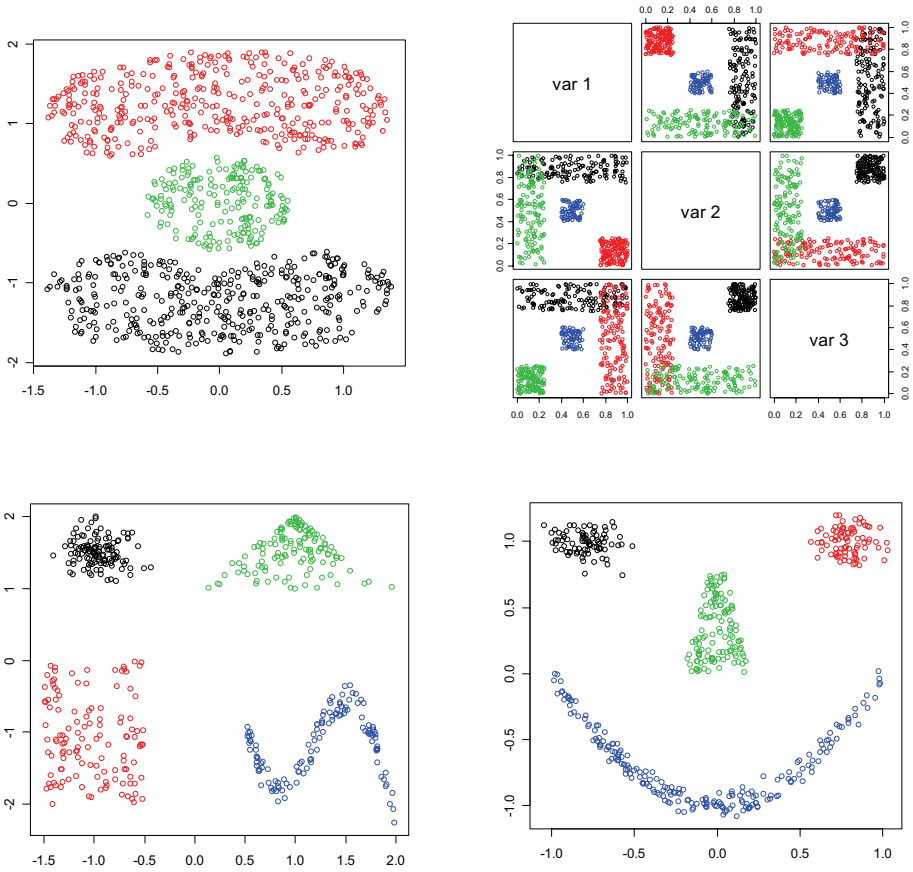

Fig. 2. Artificial generated data sets; upper – *Cassini* and *Cuboids*, bottom – *Shapes* and *Smiley*

Source: own work.

Looking at the empirical results (Table 2) it can be noticed that when as a method for co-occurrence matrix construction the *k*-means method was chosen, then the best method for their latter partitioning is *k*-means. Quite good results can be also obtained with clara algorithm. In turn *c*-means is not recommended in this case.

When the co-occurence matrix was constructed by means of *c*-means algorithm, then in three cases the best method for the final partitioning is the *k*-means algorithm, and in three cases – pam and clara. Again it can be noticed that *c*-means can't be recommended.

Tab. 2. Values of Rand Index

Matrix	kmeans	cmeans	kmeans	cmeans	kmeans	cmeans	kmeans	cmeans
Cluster algorithm / Data set	kmeans	kmeans	cmeans	cmeans	pam	pam	clara	clara
Cassini	0,791	0,973	0,791	0,971	0,784	0,972	0,784	0,972
Cuboids	0,918	0,939	0,916	0,874	0,915	0,876	0,915	0,876
Smiley	0,675	0,903	0,824	0,875	0,981	0,879	0,981	0,789
Shapes	0,845	0,997	0,998	0,997	0,997	0,998	0,999	0,997
Boston	0,680	0,622	0,613	0,617	0,655	0,675	0,668	0,675
Ecoli	0,824	0,782	0,799	0,788	0,788	0,796	0,809	0,796
Glass	0,677	0,706	0,714	0,710	0,718	0,724	0,718	0,724

Source: own computations.

Looking from the point of view of chosen taxonomic algorithm for the final partitioning it can be seen that for *k*-means and *k*-medoids algorithms the best results can be obtained when the co-occurrence matrix was constructed with *c*-means method. For *c*-means and clara algorithms an ambiguity of the results can be noticed.

IV. SUMMARY

To sum up it is worth to notice that choosing a good taxonomic method is much more difficult than choosing a good classifier. It is so because in discrimination there is a situation where class membership for the observations is known in advance; there is a problem of supervised learning. In the taxonomy on the other hand, the class adherence for objects isn't known so the right structure that should be found by the algorithm is unknown. So, in order to omit the risk of a wrong algorithm selection, the ensemble approach can be used in order to combine some of them. Since each of different clustering methods has different strengths and weaknesses it can be expected that their joint contribution will have a compensatory effect.

The next advantage of this approach is the possibility to make the results independent from selected methods or some their parameters, eg. the initial values of the *k* parameter for *k*-means algorithm. This means that aggregation make it possible to stabilize the results of clustering solutions.

The next strength of an ensemble approach is robustness; this means lower sensitivity to noise, outliers and sampling variability.

And the last conclusions that flows from the empirical experiments are that when the co-occurrence matrix is constructed by means of *k*-means algorithm then the best method of their latter partitioning is *k*-means and clara algorithms; for co-

occurrence matrix prepared by *c*-means the best methods are *k*-means, pam and clara. From the point of view of chosen classification method for *k*-means and pam the best method for co-occurrence matrix construction is *c*-means, in turn for *c*-means and clara classification methods the results are not straightforward.

REFERENCES

Bezdek J. C. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York.

Blake C., Keogh E., Merz C. J. (1988), *UCI Repository of Machine Learning Databases,* Department of Information and Computer Science, University of California, Irvine.

Breiman L. (1996), Bagging Predictors, *Machine Learning*, 26 (2): 123 – 140.

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, LNC.

Fred A., Jain A. K. (2002), Data Clustering Using Evidence Accumulation, *Proceedings of the Sixteenth International Conference on Pattern Recognition*, pages 276-280, ICPR, Canada.

Jain A., Murty M. N and Flynn P. (1999), Data Clustering: A Review, *ACM Computing Surveys,* 31 (3): 264 – 323.

Kaufman L., Rousseeuw P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis,* Wiley, New York.

Freund Y. (1990), Boosting a weak learning algorithm by majority. *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages: 202 – 216.

Kuncheva L. I., Hadjitodorov S. T., Todorova L. P. (2006), Experimental Comparison of Cluster Ensemble Methods, *Nineteenth International Conference on Information Fusion*, pages: 1 - 7, Florence.

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.

Rand W. M. (1971), Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, 66: 846 – 850.

Strehl A., Ghosh J. (2002), Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions, *Journal of Machine Learning Research*, 3: 583 - 618.

Tsymbal A., Pechenizkiy M., Cunningham P. (2003), *Diversity in Ensemble Feature Selection,* Technical Report, Trinity College Dublin.

Dorota Rozmus

PORÓWNANIE DOKŁADNOŚCI METOD TAKSONOMICZNYCH W PODEJŚCIU WIELOMODELOWYM OPARTYM NA MACIERZY WSPÓŁWYSTĄPIEŃ

Podejście wielomodelowe dotychczas z dużym powodzeniem stosowane było w klasyfikacji i regresji w celu podniesienia dokładności predykcji. W ostatnich latach analogiczne propozycje pojawiły się także w taksonomii, a liczne badania wykazały, że agregacja różniących się między sobą wyników wielokrotnego grupowania, pozwala na poprawę dokładności klasyfikacji.

W badaniu uwaga została skupiona na pozyskaniu dodatkowej informacji dostarczanej przez zbiór wyników wielokrotnie dokonanej klasyfikacji w celu konstrukcji tzw. macierzy współwystąpień. Biorąc pod uwagę jednoczesne wystąpienie pary obiektów w tej samej klasie jako wskazówkę istnienia

związku między nimi, pierwotny zbiór obserwacji przekształcany jest w $n \times n$ – wymiarową macierz, która opisuje podobieństwo między obiektami. Ostateczne grupowanie dokonywane jest na podstawie uzyskanej macierzy współwystąpień.

Celem referatu jest porównanie dokładności rozpoznawania poprawnej struktury klas zaproponowanego podejścia wielomodelowego z zastosowaniem różnych algorytmów taksonomicznych do konstrukcji macierzy współwystąpień oraz jej późniejszego podziału na klasy obiektów podobnych do siebie.