ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 269, 2012

Dorota Rozmus[*]

COMPARISON OF STABILITY OF CLASSICAL TAXONOMY BAGGING METHOD WITH BAGGING BASED ON CO-OCCURRENCE 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 in order to increase classification accuracy, robustness and stability of the clustering solutions. Research has proved that, by combining a collection of different clusterings, an improved solution can be obtained.

The stability of a clustering algorithm with respect to small perturbations of data (e.g., data subsampling or small variations in the feature values) or the parameters of the algorithm (e.g., random initialization) is a desirable quality of the algorithm. On the other hand, ensembles benefit from diverse clusterers. Although built upon unstable components, the ensemble is expected to be more accurate and robust than the individual clustering method. Here, we look at the stability of the ensemble methods based on bagging idea and co-occurrence matrix. This paper carries out an experimental study to compare stability of bagging method used to the classical data set with bagging based on co-occurrence matrix.

Key words: Cluster analysis, Cluster ensemble, Stability, Bagging in taxonomy, Co-occurrence matrix.

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. 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 the mean value of dependent variable *y*. The presumption in this approach is that using many models instead of one will give better results.

Recently, ensemble approach for cluster analysis has been suggested in order to increase the classification accuracy and robustness of the clustering solutions. The

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

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 established a new area in the conventional taxonomy (Fred 2002, Fred and Jain 2002, 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; (3) use different subsets of features; (4) run a given algorithm many times with different parameters or initializations.

II. STABILITY MEASURES AND CLUSTER ACCURACY

The stability of a clustering algorithm with respect to small perturbations of data and also different initializations is a desirable quality of the algorithm. Cluster ensembles, on the other hand, enforce and exploit some instability so that the ensemble is comprised of diverse clusterers (Fern, Brodley 2003, Green *et al.* 2004). Although built upon unstable components, the ensemble is expected to be more accurate and robust than individual clustering methods.

In this research stability of a clustering algorithm will be considered. The main aim is to compare the stability of ensemble approach based on bagging method used to the classical data set and to the co-occurrence matrix.

In the research there will be used measures of stability and accuracy proposed by Kuncheva and Vetrov (2006). Both are based on adjusted Rand Index (*AR*).

1. Pairwise ensemble stability:

$$S_{agr} = \frac{2}{K \cdot (K-1)} \sum_{\substack{1 \le k, l \le K \\ k<l}}^{\mathrm{K}} AR(P_k^{agr}, P_l^{agr}), \tag{1}$$

where:
K – number of ensembles,
AR – adjusted Rand Index,
P_k^{agr}, P_l^{agr} – classification on the base of – correspondingly – *k*th and *l*th ensemble.

2. Average ensemble accuracy:

$$A_{agr} = \frac{1}{K} \sum_{k=1}^{K} AR(P_k^{agr}, P^T), \tag{2}$$

where: P^T – true class labels.

III. BAGGING IN TAXONOMY AND THE CO-OCCURRENCE MATRIX

In bagging method in taxonomy (Hornik 2006) the first step is construction of bootstrap samples and running a cluster algorithm on them in order to get single partitions that are members of the cluster ensemble. The final partition is obtained by the *optimization approach* which formalizes the natural idea of describing consensus clusterings as the ones which "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 $(c_1,...,c_B)$ are the elements of the ensemble, the problem is solved by means of *least squares* consensus clusterings (generalized means):

$$\sum_{b=1}^{B} dist(c,c_b)^2 \Rightarrow \min_{c\in C} . \qquad (33)$$

The idea of using co-occurrence matrix as a data set has the source in provided by Kuncheva, Hadjitodorov and Todorova (2006) research in taxonomy where they got very promising results with co-occurrence matrix treated as a data set. The concept of co-occurrence matrix was proposed by Fred and Jain (2002) as 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 co-occurrence matrix construction can be described as follows:

First step - split. For a fixed number of cluster ensemble members *P*, 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 *P* 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 *P* runs of *k*-means are mapped into a $n\times n$ co-occurrence matrix:

$$co\_assoc(a,b) = votes_{ab} , \qquad (4)$$

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

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

IV. EMPIRICAL RESULTS

The aim of empirical experiments is to compare stability of bagging method used to the classical data and to the co-occurrence matrix. All computations are made on artificial generated sets, usually used in comparative studies in taxonomy, taken from `mlbench` package in **R**.

In the classical bagging method, i.e. used to the classical data set, 50 ensembles were constructed, each of them based on 25 single members. In the first version of the experiment each single member was built on the bootstrap sample by means of *k*-means and in the second version- by *c*-means algorithm; with the *k* and *c* parameters equal to the true number of clusters.

In the experiments with bagging method used to the co-occurrence matrix, generally the matrix was based on 10 single partitions and bagging method was run 25 times on each of the final matrices. The number of ensembles was also equal 50. In one version of the experiments the co-occurrence matrix was built by means of *k*-means or *c*-means algorithms and the bagging method was run on bootstrap samples with *k*-means algorithm. In the second version – the co-occurrence matrix was built by means of the same algorithms (i.e. *k*-means or *c*-means) and bagging was run on them with *c*-means algorithm. All computations were repeated 20 times.

Looking at the results that refer to the stability (Fig. 1) it can be said that when for successive bootstrap samples *k*-means algorithm is used, then for *Threenorm, Ringnorm, 2dnormals* and *Shapes* data sets bagging method on co-occurrence matrix brings impairment of the results in comparison with bagging method used to the classical data sets. Especially it can be noticed for *Threenorm* data set for co-oc_k+bag_k method[1]. Only two data sets, i.e. *Cassini* and *Smiley* give improvement of the stability by using co-occurrence matrix, especially those built by means of *c*-means algorithm. In the case when to the successive bootstrap samples *c*-means algorithm is used then, for most of the data sets, bagging on co-occurrence matrix gives very similar results with classical bagging method, i.e. used for classical data sets. Visible exceptions are only *Cassini* and *Ringnorm* data sets, where co-oc_k+bag_c gives worse results in comparison with the classical bagging method (bag_cmeans); whereas co-oc_c+bag_c gives improvement of the results. The second exception is *Smiley* data set where in both cases using co-occurrence matrix gives worse results.

[1] The first part of the name pertains to the way of the construction of the co-occurrence matrix (k refers to *k*-means algorithm; c – to *c*-means); the second to the method used to the successive bootstrap samples (as above).

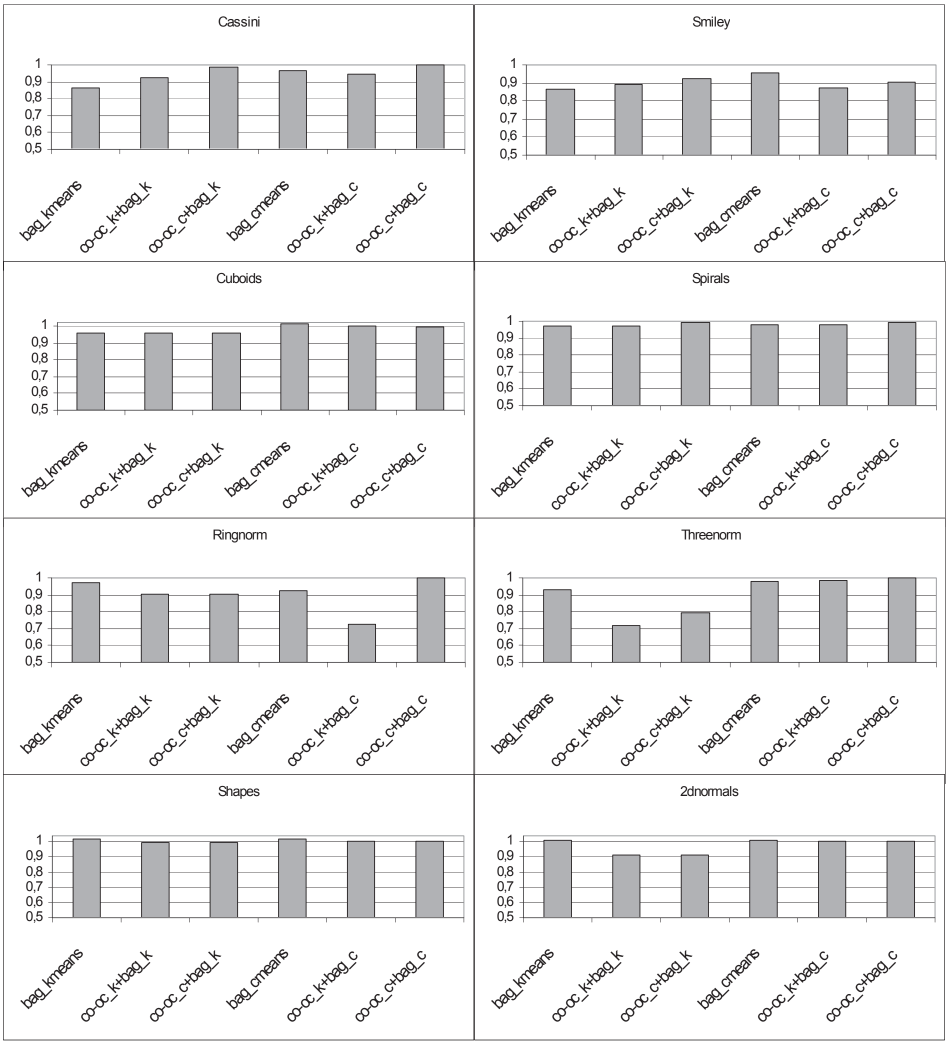

Fig. 1. Stability for different data sets

Source: own computations.

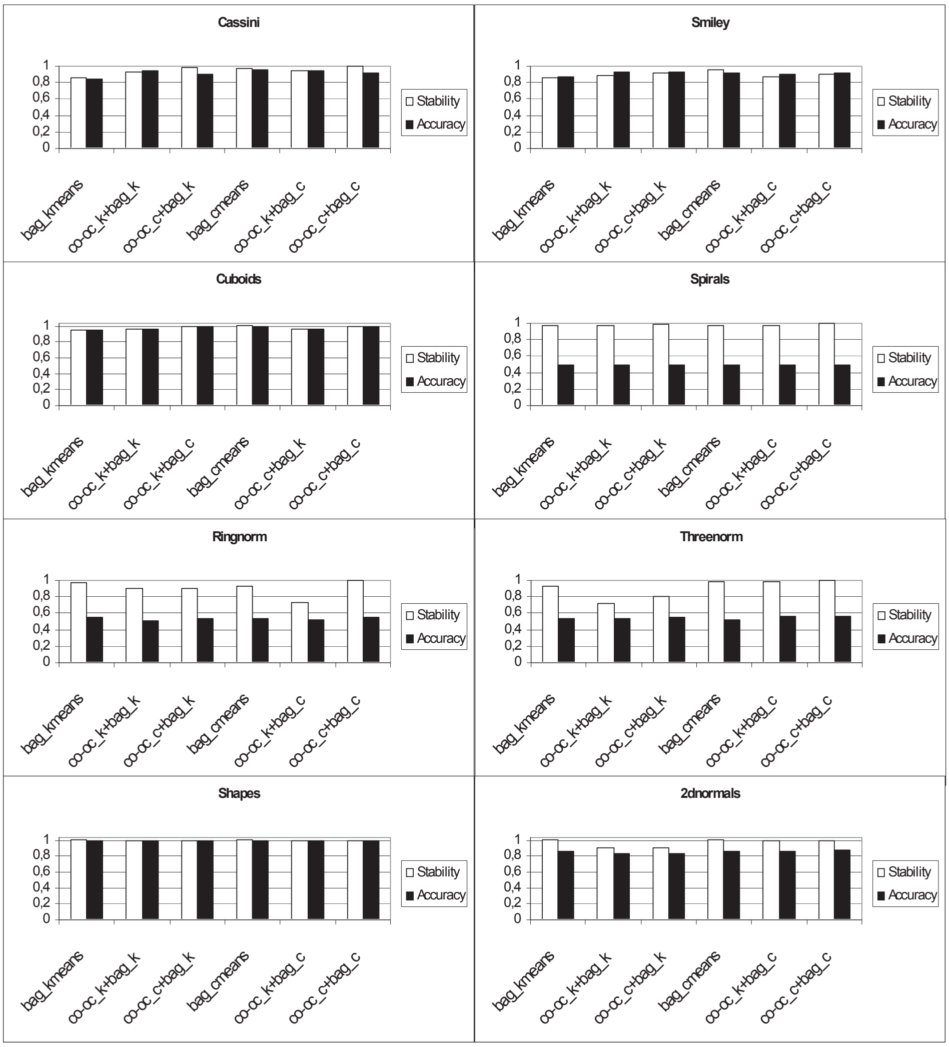

Fig. 2. Relationship between stability and accuracy for different data sets

Source: own computations.

Results that refer to the relationship between accuracy and stability (Fig. 2) show two general patterns. For *Cassini*, *Cuboids*, *Shapes*, *Smiley* and *2dnormals* data sets stability and accuracy measures reach almost the same level regardless of the bagging method was used to the classical or co-occurrence data sets. An insignificantly higher difference can be noticed only for *2dnormals* data set where for all the methods accuracy is always lower than stability measures.

In turn for *Ringnorm, Spirals* and *Threenorm* data sets accuracy is on almost the same level for all the methods but stability behaves very differently, i.e. it achieves almost the same level for *Spirals* data set regardless of the used method and different levels for *Ringnorm* and *Threenorm* data sets.

IV. SUMMARY

To sum up it is worth to notice that many clustering algorithms, also those based on ensemble approach rely on a random component. So the stability of a clustering algorithm with respect to small perturbations of the data, or the parameters of the algorithm is a desirable quality. On the other hand it is also known that diversity within an ensemble is of vital importance for its success. Although built upon unstable components, the ensemble is expected to be more accurate and robust than the individual clustering method. Here, we have look at the stability of the ensemble. The main aim of this research was to compare the stability of classical bagging method (i.e. used to the classical data set) in taxonomy with bagging used to the co-occurrence matrix. From the empirical results it appears that generally there is no clear pattern – for some data sets using bagging to the co-occurrence matrix gives better stability than running it on classical data sets; and for some – it give worse results. No clear pattern can be also noticed for the relationship between stability and accuracy for compared methods.

REFERENCES

Fern X. Z., Brodley C. E. (2003), Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach, *Proceedings of the 20th International Conference of Machine Learning*, pages: 186–193.

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 16th International Conference on Pattern Recognition*, pages: 276–280, ICPR, Canada.

Greene D., Tsymbal A., Bolshakova N. and Cunningham P. (2004), Ensemble Clustering in Medical Diagnostics, *Technical Report TCD-CS-2004-12*, Trinity College, Dublin, Ireland.

Hornik K. (2006), A CLUE for CLUster Ensembles*, Journal of Statistical Software*, 14:65–72.

Kuncheva L. I., Hadjitodorov S. T., Todorova L. P. (2006), Experimental Comparison of Cluster Ensemble Methods, *19th International Conference on Information Fusion*, pages: 1 - 7, Florence.

Kuncheva L., Vetrov D. (2006), Evaluation of Stability of *k*-Means Cluster Ensembles with

Respect to Random Initialization, *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 28, No. 11, pages: 1798–1808.

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

Dorota Rozmus

PORÓWNANIE STABILNOŚCI KLASYCZNEJ TAKSONOMICZNEJ METODY BAGGING Z METODĄ BAGGING OPARTĄ NA MACIERZY WSPÓŁWYSTĄPIEŃ

Podejście wielomodelowe dotychchzas z dużym powodzeniem stosowane było w dyskryminacji w celu podniesienia dokładności klasyfikacji. Analogiczne propozycje pojawiły się także w taksonomii, aby zwiększyć poprawność i stabilność wyników grupowania.

Stabilność algorytmu taksonomicznego w odniesieniu do niewielkich zmian w zbiorze danych (np. wybór podzbioru zmiennych), czy też parametrów algorytmu (np. losowa inicjalizacja algorytmu) jest pożądaną cechą algorytmu. Głównym punktem zainteresowania tego referatu jest stabilność w podejściu zagregowanym taksonomii. Zasadniczym celem jest przeprowadzenie badań empirycznych, które mają za zadanie porównać stabilność metody bagging stosowanej do klasycznego zbioru danych oraz do tzw. macierzy współwystąpień.