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 255, 2011

Dorota Rozmus[*]

COMPARISON OF STABILITY OF ALGORITHMS IN CLASSICAL AND ENSEMBLE APPROACH IN TAXONOMY

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 resampling, 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 (Fern, Brodley 2003, Green *et al.* 2004). Although built upon unstable components, the ensemble is expected to be more accurate and robust than the individual clustering method. Here, the stability of the ensemble is looked at. This paper carries out an experimental study to examine whether cluster ensembles give more stable results than single clustering methods.

Key words: Cluster analysis, Cluster ensemble, Stability, Accuracy.

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

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

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 established 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; (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. Although built upon unstable components, the ensemble is expected to be more accurate and robust than the individual clustering method.

In this research stability of a clustering algorithm will be considered. The main aim is to compare the stability and accuracy of single and ensemble approach in taxonomy and also to study the relationship between accuracy and stability in cluster ensemble. For this purpose measures of accuracy and stability proposed by Kuncheva and Vetrov (2006) were used. All of them are based on adjusted Rand Index (*AR*).

1. Average ensemble accuracy:

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

where:
K – number of ensembles (k = 1, 2, …, 50),
AR – adjusted Rand Index,
P_k^{agr} – classification on the base of kth ensemble,
P^T – true class labels.

2. Individual accuracy:

$$A_i = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{J}\sum_{j=1}^{J} AR(P_j^k, P^T), \tag{2}$$

where:
J – number of ensemble members (j = 1, 2, …, 25),
P_j^k – classification on the base of jth member of kth ensemble.

3. Pairwise ensemble stability:

$$S_{agr} = \frac{2}{K \cdot (K-1)} \sum_{\substack{1 \le k,l \le K \\ k<l}}^{K} AR(P_k^{agr}, P_l^{agr}), \tag{3}$$

where:
P_l^{agr} – classification on the base of lth ensemble.

4. Pairwise individual stability:

$$S_i = \frac{1}{K} \sum_{k=1}^{K} \frac{2}{J(J-1)} \sum_{i<j} AR(P_i^k, P_j^k), \tag{4}$$

where:
P_i^k - classification on the base of ith member of kth ensemble,
P_j^k - classification on the base of jth member of kth ensemble.

III. EMPIRICAL RESULTS

The aim of empirical experiments was to compare the stability of single and cluster ensemble approach and also the relationship between accuracy and stability in cluster ensemble approach.

In the research there were used 50 ensembles, each of them based on 25 single members. Each single member was built on the bootstrap sample by means of k-means algorithm, and the results were aggregated by the method proposed by Hornik (2005). All were built on artificial generated data sets, whose short characteristics are shown on the Figure 1.

Looking at the results (Fig. 2 – Fig. 7) it can be seen that ensemble approach gives higher accuracy in comparison to the classical approach what is suggested by the diagrams on the left side. On the basis of the diagrams on the right side it can also be seen that cluster ensemble gives more stable results. The only exception is *Threenorm* data set where although cluster ensemble gives worse accuracy in comparison with single approach but still the aggregated approach is more stable than single algorithm.

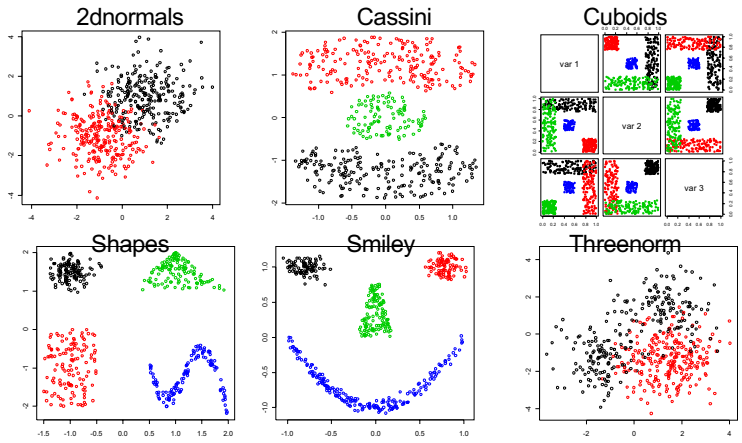

Fig. 1. Artificial generated data sets

Source: own work.

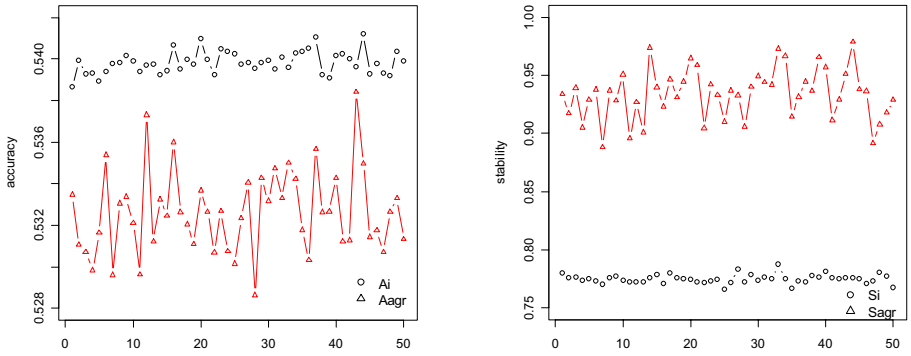

Fig. 2. Results for *Threenorm* data set

Source: own work.

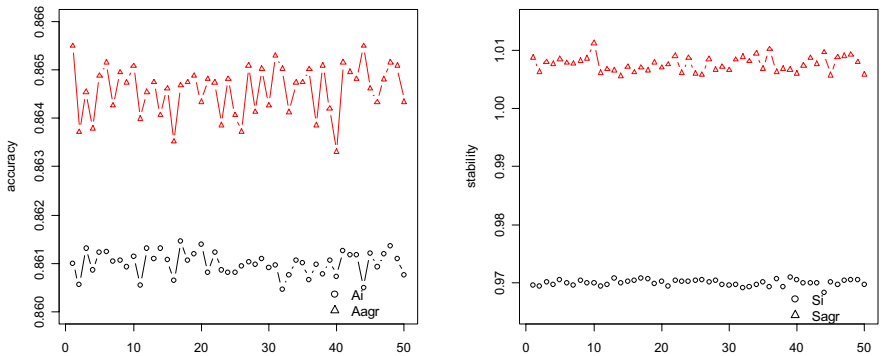

Fig. 3. Results for *2dnormals* data set

Source: own work.

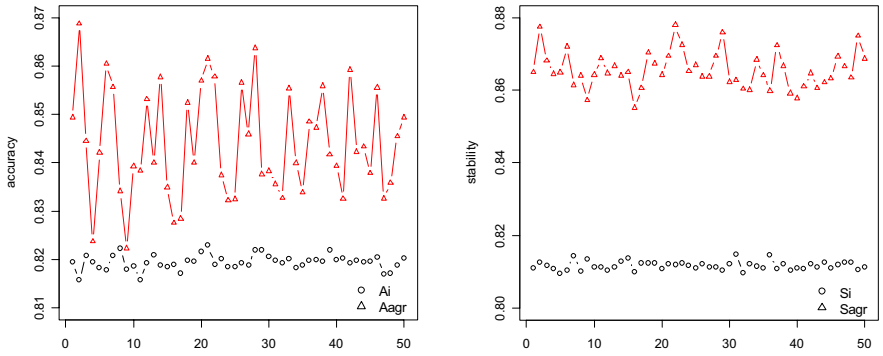

Fig. 4. Results for *Cassini* data set

Source: own work.

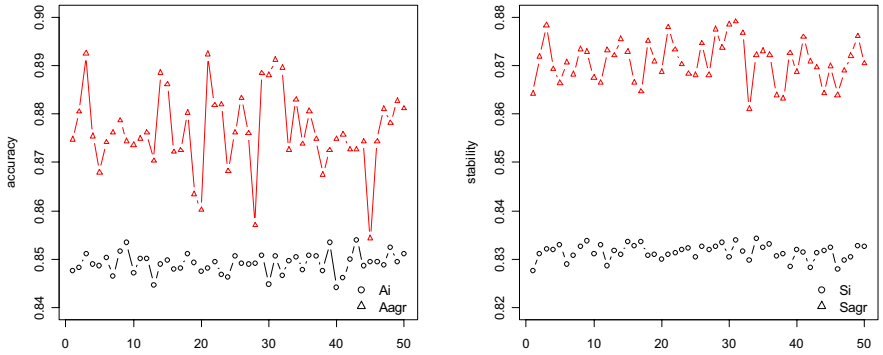

Fig. 5. Results for *Shapes* data set

Source: own work.

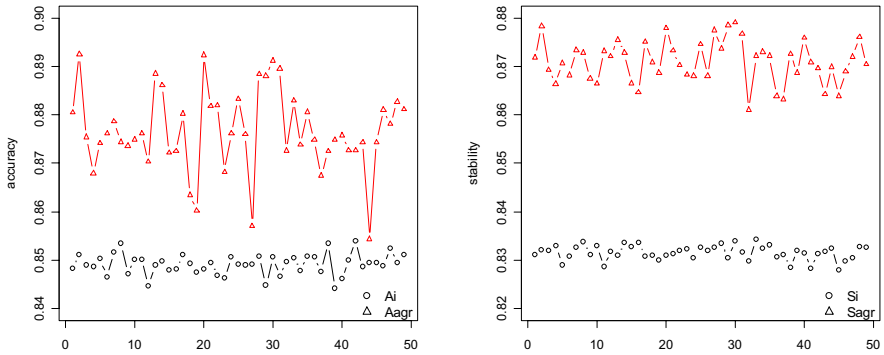

Fig. 6. Results for *Smiley* data set

Source: own work.

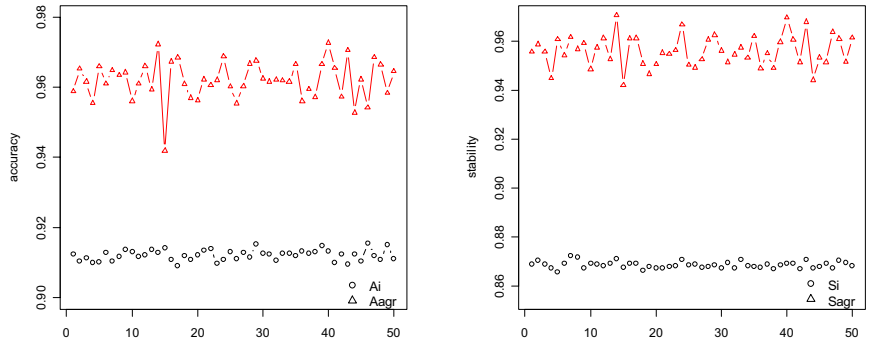

Fig. 7. Results for *Cuboidsy* data set

Source: own work.

More insightful study of the diagrams allows also conjecturing that higher stability goes together with higher accuracy. To confirm this suppose scatter plots for stability and accuracy are presented on fig. 8 and fig. 9. Almost all of them confirm this supposition. The most it is seen for *Cuboids* and the least for *Cassini* data set (it was supported by Pearson's linear coefficient).

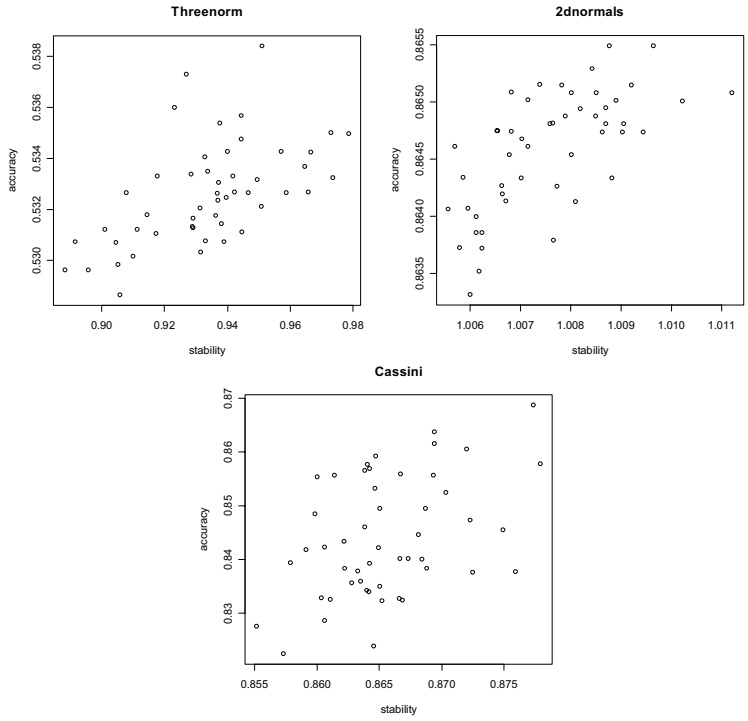

Fig. 8. Scatterplots for stability and accuracy for *Threenorm, 2dnormals* and *Cassini* data sets

Source: own work.

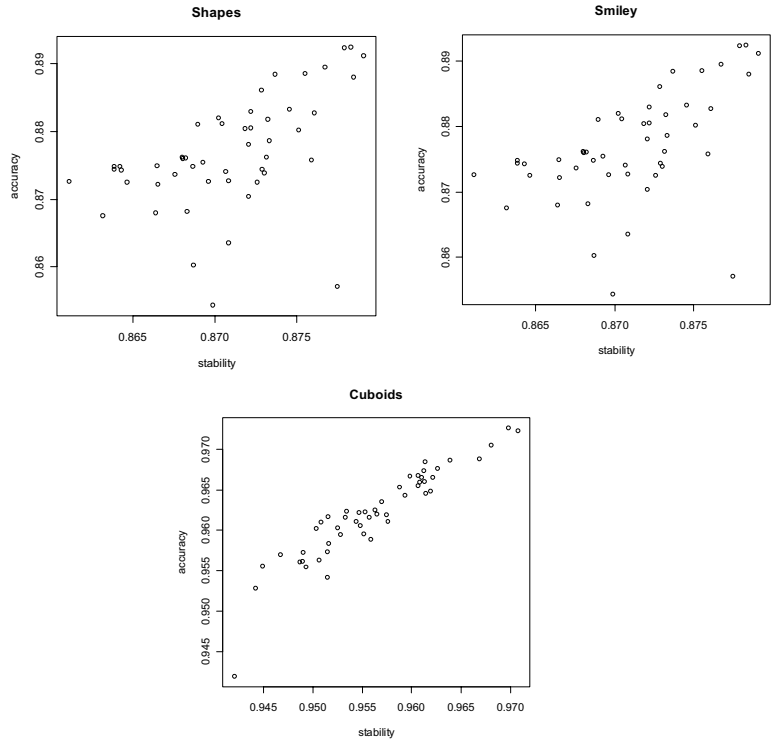

Fig. 9. Scatterplots for stability and accuracy for *Shapes, Smiley, Cuboids* data sets
Source: own work.

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 single and cluster ensemble approach and also the relationship between accuracy and stability in cluster ensemble. From the empirical results it appears that ensemble approach gives more stable results than classical approach and that often higher stability went together with a higher accuracy of an ensemble.

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 Learn*ing, 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. (2005), A CLUE for CLUster Ensembles*, Journal of Statistical Software*, 14:65–72

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

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 ALGORYTMÓW W PODEJŚCIU KLASYCZNYM I ZAGREGOWANYM W TAKSONOMII

Podejście wielomodelowe dotychczas z dużym powodzeniem stosowane było w dyskryminacji w celu podniesienia dokładności klasyfikacji. W ostatnich latach analogiczne propozycje pojawiły się w taksonomii, aby zapewnić większą poprawność i stabilność wyników grupowania. Liczne badania wykazały, że agregacja różniących się między sobą wyników wielokrotnego grupowania, pozwala na poprawę dokładności klasyfikacji.

Stabilność algorytmu taksonomicznego w odniesieniu do niewielkich zmian w zbiorze danych, czy też parametrów algorytmu jest pożądaną cechą algorytmu. Z drugiej jednak strony, podejście wielomodelowe czerpie korzyści ze zróżnicowanych klasyfikacji składowych, których połączenie przynosi bardziej dokładne i stabilne rozwiązanie niż pojedynczy algorytm.

Głównym punktem zainteresowania tego badania była stabilność w podejściu zagregowanym w taksonomii. Przeprowadzone badania empiryczne pokazały, że podejście zagregowane daje bardziej stabilne rezultaty niż pojedyncze algorytmy taksonomiczne oraz, że często wyższa stabilność idzie w parze z wyższą dokładnością klasyfikacji w podejściu zagregowanym.