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 286, 2013

*Dorota Rozmus**

COMPARISON OF ACCURACY OF AFFINITY PROPAGATION METHOD AND CLUSTER ENSEMBLES BASED ON CO-OCCURRENCE MATRIX

Abstract. High accuracy of results is a 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 whose 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 affinity propagation method. Here, we carry out an experimental study to compare accuracy of those two approaches.

Key words: clustering, accuracy, affinity propagation, cluster ensemble.

I. INTRODUCTION

Recently, affinity propagation method has 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. Affinity propagation is a relatively new clustering algorithm that has been introduced by Frey and Dueck (2007). The authors themselves describe affinity propagation as follows:[1] *"An algorithm that identifies exemplars among data points and forms clusters of data points around these exemplars. It operates by simultaneously considering all data point as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges."* Cluster ensemble approach can be defined generally as follows: given multiple partitions of the data set, find a combined clustering with a better quality. The main aim of this research is to compare accuracy of affinity propagation clustering and cluster ensembles based on co-occurrence matrix (Fred 2002; Fred and Jain 2002).

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

[1] Quoted from http://www.psi.toronto.edu/affinitypropagation/faq.html#def.

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.

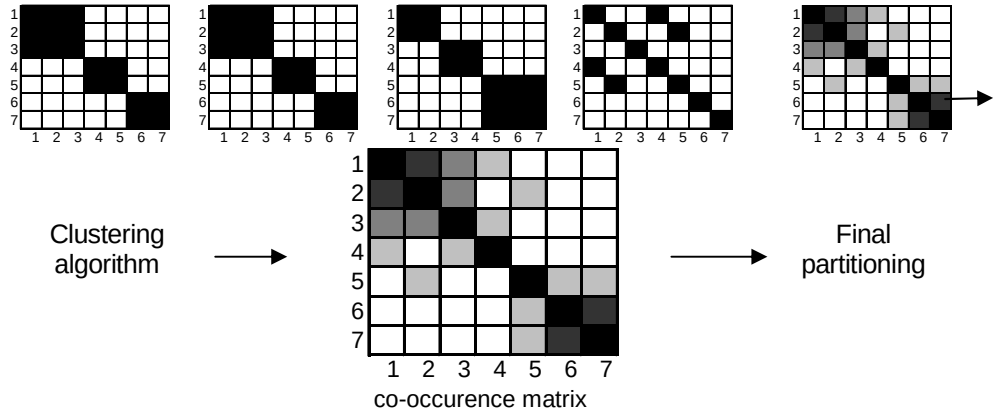

Fig. 1. Construction of the co-occurrence matrix and their final partitioning
Source: own work.

III. AFFINITY PROPAGATION

This method takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between all data points until a high-quality set of exemplars and corresponding clusters gradually emerges. The main aim of this method is to maximize the sum of similarities between points and their exemplars.

The particular steps of the algorithm are as follows:

1. Using negative squared error (Euclidean distance) find a matrix of similarities between points:

$$s(i,k) = -\left\|x_i - x_k\right\|^2 . \tag{2}$$

2. Find so called preferences which can be interpreted as the tendency of a data sample to become an exemplar:

$$s(k,k) = p, . \tag{3}$$

Two kinds of information are exchanged between points:

a. The "responsibility" $r(i,k)$, sent from data point x_i to candidate exemplar point x_k, reflects the accumulated evidence for how well-suited point x_k is to

serve as the exemplar for point x_i, taking into account other potential exemplars for point x_i.

b. The "availability" $a(i,k)$, sent from candidate exemplar point x_k to point x_i, reflects the accumulated evidence for how appropriate it would be for point x_i to choose point x_k as its exemplar, taking into account the support from other points that point x_k should be an exemplar.

4. To begin with, the availabilities are initialized to zero: $(a(i,k) = 0)$.

5. The responsibilities and availability are computed using the rules:

$$r(i,k) \leftarrow s(i,k) - \max_{k':k' \neq k} \{a(i,k') + s(i,k')\}, \tag{4}$$

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum_{i':i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}, & \text{when } i \neq k \\ \sum_{i':i' \neq i} \max\{0, r(i',k)\}, & \text{when } i = k \end{cases}. \tag{5}$$

6. The message-passing procedure may be terminated after:
- a fixed number of iterations,
-changes in the messages fall below a threshold,
- the local decisions stay constant for some number of iterations.

7. Partition of data points to clusters $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_N)$ is done according the rule:

$$\hat{c}_i = \arg\max_k [a(i,k) + r(i,k)], \tag{6}$$

where $\hat{c}_i$ is an exemplar of those cluster, where observation x_i is assigned.

IV. NUMERICAL EXPERIMENTS

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

$$Acc = \frac{1}{Z} \sum_{z=1}^{Z} R(P_z, P^T), \tag{7}$$

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 artificially generated data sets 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	# 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.

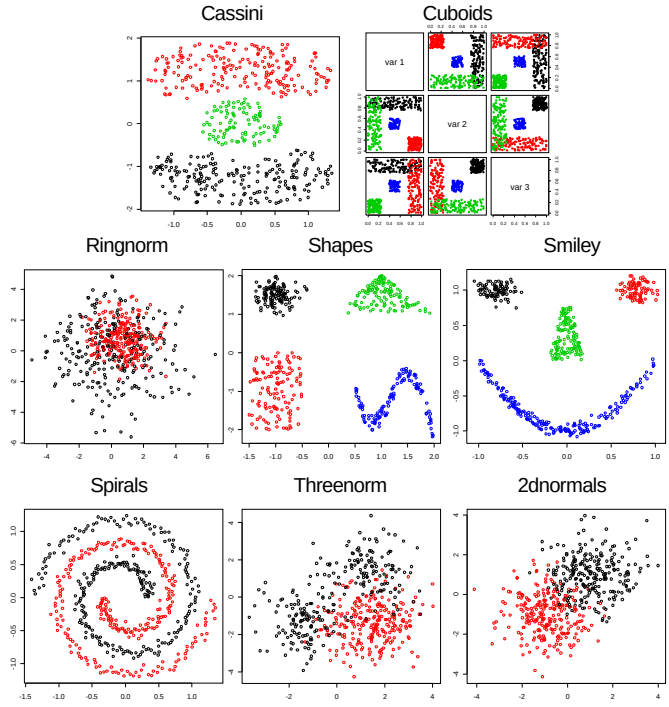

Fig. 2. Structure of the used data sets

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

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.

In the case when the co-occurrence matrix was built by means of *k*-means method it can be said that in most cases aggregated approach and affinity propagation method gives very similar results. It can be seen especially for *Cassini, Ringnorm, Spirals* and *Threenorm* data sets. Higher differences in accuracy can be noticed for *Cuboids*, *Shapes* and *2dnormals* where the least accurate is aggregated variant kmeans_kmeans and for *Smiley* data set where the most accurate are aggregated variants kmeans_pam and kmeans_clara.

Similar conclusions bring the results in case of comparison affinity propagation method with co-occurrence matrix built by means of *c*-means. That means very similar results for both approaches especially for *Cassini*, *Ringnorm*, *Spirals* and *Threenorm* data sets. Higher differences can be noticed for *Cuboids*, *Shapes* and *2dnormals* where the least accurate are aggregated variants kmeans_pam and kmeans_clara.

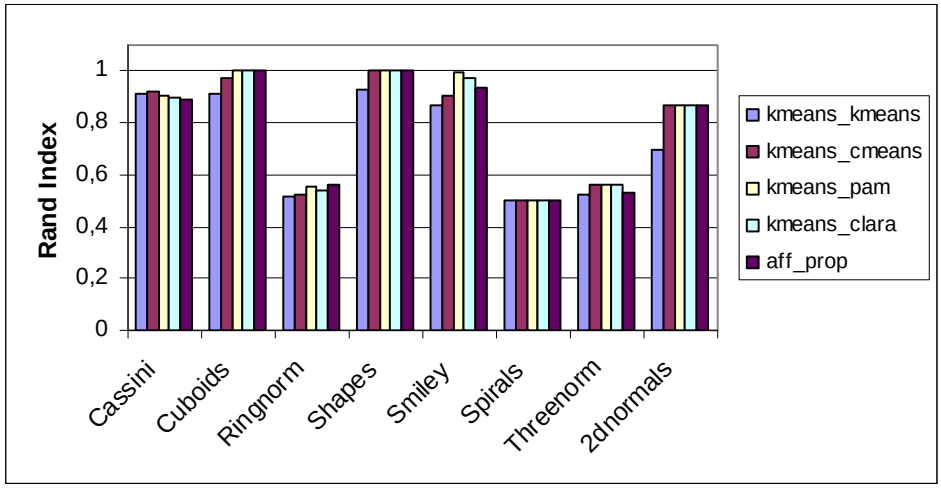

Fig. 3. Accuracy of affinity propagation and cluster ensemble based on co-occurrence matrix with *k*-means used for its construction

Source: own work.

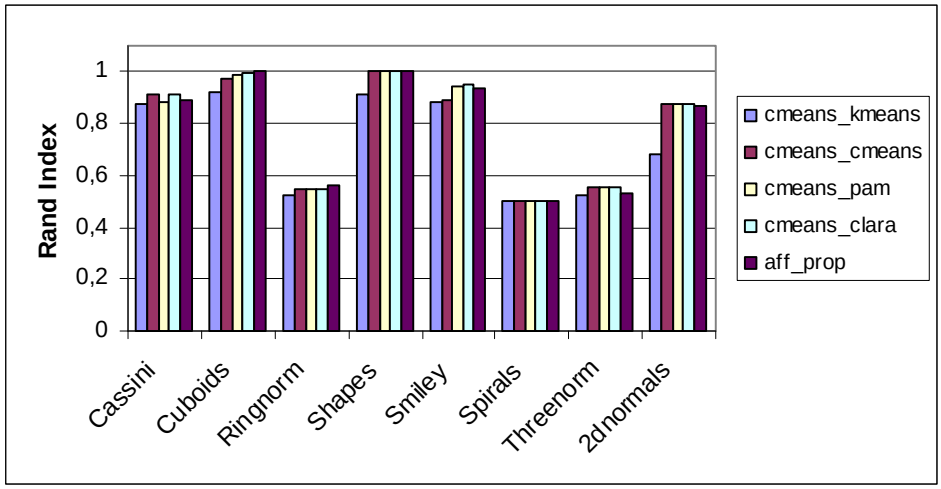

Fig. 4. Accuracy of affinity propagation and cluster ensemble based on co-occurrence matrix with *c*-means used for its construction

Source: own work.

V. CONCLUSIONS

To sum up all the numerical experiments of this research it can be said that in most cases affinity propagation method and cluster ensemble based on co-occurrence matrix give very similar results, especially for *Cassini*, *Ringnorm*, *Spirals* and *Threenorm* data sets. Only aggregated variants kmeans_pam and kmeans_clara for *Cuboids*, *Shapes* and *2dnormal* data sets are noticeably better than affinity propagation method.

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.

Frey B. J., Dueck D., (2007), Clustering by passing messages between data points, *Science*, 315, 972-976. DOI: 10.1126/science.1136800.

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.

Dorota Rozmus

PORÓWNANIE DOKŁADNOŚCI TAKSONOMICZNEJ METODY PROPAGACJI PODOBIEŃSTWA ORAZ ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA IDEI 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 ich podstawie. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod (np. *k*-średnich, metod hierarchicznych). Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego, czyli łączenia wyników uzyskanych w wyniku wielokrotnego grupowania (ang. *cluster ensemble*) oraz taksonomiczna metoda propagacji podobieństwa (ang. *affinity propagation clustering*).

Głównym celem tego artykułu jest porównanie dokładności taksonomicznej metody propagacji podobieństwa zaproponowana przez Frey i Duecka (2007) oraz zagregowanych algorytmów taksonomicznych opartych idei macierzy współwystąpień (Fred, Jain 2002).