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

Marcin Pełka[*]

CLUSTERING OF SYMBOLIC DATA WITH APPLICATION OF ENSEMBLE APPROACH

Abstract. Ensemble approaches based on aggregated models have been applied with success to discrimination and regression tasks. Nevertheless this approach can be applied to cluster analysis tasks. Many articles have proved that, by combining different clusterings, an improved solution can be obtained.

The article presents the possibility of applying ensemble approach based on aggregated models to cluster symbolic data. The paper presents also presents results of clustering obtained by applying ensemble approach.

Key words: cluster ensemble, co-associacion matrix, symbolic data.

I. INTRODUCTION

Symbolic objects, unlike classical objects, can be described by many different symbolic variable types. Besides well-known classical variables (metric or not) symbolic objects can be described by interval-valued variables, multivalued variables and multivalued variables with weights and also dependent variables [see for example Bock, Diday *et. al.* 2000, p. 2–3; Billard, Diday 2006, p. 7–8].

Generally ensemble approach means combining (aggregating) the results of M base models in to one aggregated model. The general aim of ensemble approach is to improve the performance (solution) of a model. This kind of approach has been applied with a success to regression and discrimination tasks.

Nevertheless the idea of ensemble approach, that is combining results of different models, can be successfully applied when clustering symbolic data [see for example Pełka 2012]. The aim of the article is to present the results of ensemble clustering of symbolic data. The empirical part of the paper presents results of clustering with application of different base models. In order to compare partitions obtained from ensemble clustering and known cluster structure adjusted Rand index was applied.

[*] Ph. D., Department of Econometrics and Informatics, University of Economics, Wroclaw.

II. SYMBOLIC DATA

Bock and Diday have defined five different symbolic variable types [Bock, Diday *et. al*. 2000, p. 2–3; Billard, Diday 2006, p. 8–30]:

1) single quantitative variable,
2) categorical variable,
3) quantitative variable of interval type (interval-valued variable),
4) set of values or categories (multivalued variable),
5) set of values or categories with weights (multivalued variable with weights),
6) modal interval-valued variable proposed in Billard and Diday (2006).

Regardless of the variable type symbolic variables also can be [Bock, Diday *et. al.* 2000, p. 2; Billard, Diday 2006, p. 30–34]:

- taxonomic dependent – with present *prior* known structure,
- hierarchically dependent – rules which decide if a variable is applicable or not have been defined,
- logically dependent – logical rules that affect variable's values have been defined.

There are two main types of symbolic objects in the symbolic data analysis. **First order objects** (simple objects) – single respondent, product, company (single individuals) described by symbolic variables. These objects are individuals that are called symbolic due to their nature. **Second order objects** (aggregate objects, super individuals) – more or less homogeneous groups, classes of classical objects or individuals described by symbolic variables.

III. ENSEMBLE CLUSTERING

The problem of clustering fusion can be generally described as follows – having a group of multiple partitions of the same data set, find an aggregated (combined) clustering with a better quality than simple partition [see for example: Fred 2001; Fred and Jain 2005; Jain *et. al.* 1999; Strehl and Gosh 2002; Gathemi *et. al.* 2009].

There are some ways to obtain different base clusterings that can be applied in symbolic data case [Fred and Jain 2005, p. 843, Rozmus 2010, p. 178; Gathemi *et. al.* 2009]:

1) combine results of different clustering algorithms,
2) produce different partitions by resampling data,
3) use different subsets of features (that can be disjoint or overlapping),
4) applying the same clustering algorithm with different values of parameters or initializations.

All of these approaches allow to obtain diverse base clustering that are essential to obtain good final ensemble clustering. The paper introduces the concept of evidence accumulation clustering proposed by Fred and Jain, that maps the individual data partition in clustering ensemble into a new similarity measure between patterns, that summarizes inter-pattern structure.

The result of applying any of methods that can produce different partitions is the set of N partitions $\mathbf{P} = \{P^1, P^2, \ldots, P^N\}$. The new partition P^* is obtained by combining N partitions with k^* clusters. This partition P^* should satisfy properties [Fred and Jain 2005, p. 842]:

- consistency with the clustering ensemble $\mathbf{P}$,
- robustness to small variations in $\mathbf{P}$,
- goodness of fit with ground truth information (true cluster labels of patterns), if such information is available.

The algorithm of ensemble clustering using evidence accumulation for symbolic data can be described as follows [Fred and Jain 2005, p. 848]:

1) obtain different base partitions,

2) build the co-association matrix – the underlying assumption is that patterns belonging to a "natural" cluster are very likely co-located in the same cluster in different data partitions. The N partitions of n patterns are mapped into $n \times n$ co-association matrix:

$$C(i, j) = \frac{n_{ij}}{N}, \tag{1}$$

where: i, j – pattern numbers, n_{ij} – number of times pattern (i, j) is assigned to the same cluster among N partitions, N – number of partitions.

3) use the co-association matrix as the data matrix for single link method – of course any other clustering method can be applied in this part,

4) the final partition P^* is chosen as the one with the highest lifetime – it is defined as the range of the threshold values on the dendrogram that lead to identification of k^* clusters.

In the empirical part a combination of different clustering algorithms have been combined in order to obtain the co-association matrix. This matrix was then applied as the data matrix for single link method and *k*-means method. The Silhouette, Baker & Hubert and Hubert & Levine internal cluster quality indices were applied to determine number of clusters in the case of *k*-means clustering.

IV. EMPIRICAL RESULTS

For empirical part three different data sets where prepared:

1. **First data set** contains 110 objects divided into two elongated clusters described by two symbolic interval-valued variables. This data set contains no noisy variables or outliers.

A clustering ensemble with 11 partitions $(N = 11)$ was obtained by running: the partition around medoids (*pam*) algorithm, hierarchical clustering methods (ward, single, complete, mcquitty, median and centroid) – all based on Hausdorff distance for symbolic interval-valued data – with number of clusters k randomly chosen in the interval $[2;11]$. A co-association matrix was build, and then used as data matrix for single-link method. A cluster dendrogram is presented on figure 1.

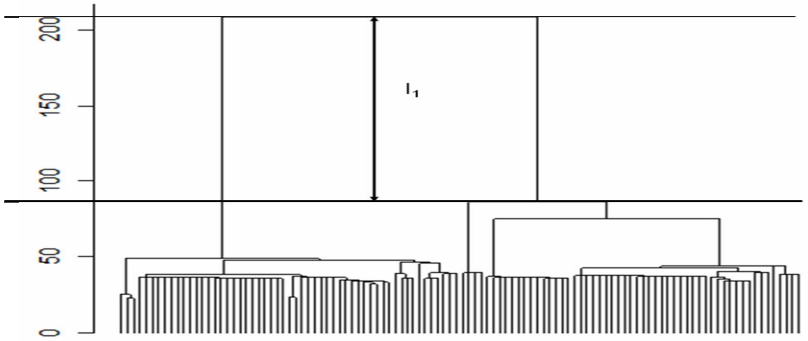

Figure 1. Cluster dendrogram for the first data set

Source: own computation with application of R software.

The longest lifetime that indicates number of clusters allowed to discover two cluster structure (indicated by l_1 on figure 1.). In order to compare partitions adjusted Rand index was calculated. $AR = 1$, this result means complete agreement between partitions.

2. **Second data set** contains five clusters in three dimensions that are not well separated – 150 objects. This data set also does not contain any outliers nor noisy variables.

A clustering ensemble with 11 partitions $(N = 11)$ was obtained by running: the partition around medoids (*pam*) algorithm, hierarchical clustering methods (ward, single, complete, mcquitty, median and centroid) – all based on Hausdorff distance for symbolic interval-valued data – with number of clusters

k randomly chosen in the interval $[2;15]$. A co-association matrix was build, and then used as data matrix for single-link method. A cluster dendrogram is presented on figure 2.

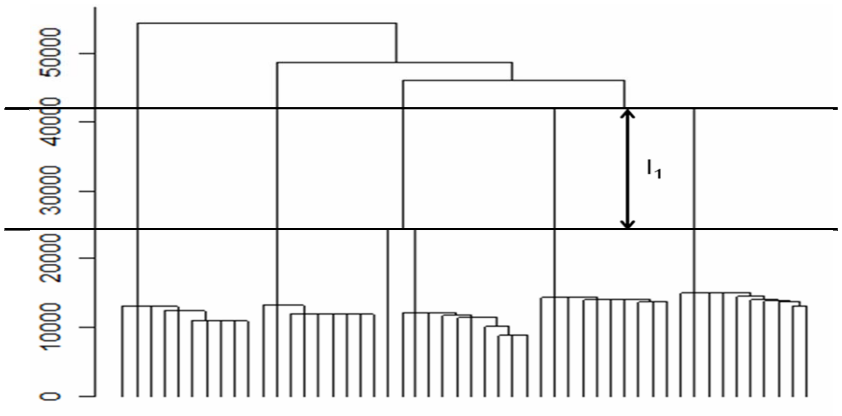

Figure 2. Cluster dendrogram for the second data set

Source: own computation with application of R software.

The longest lifetime that indicates number of clusters allowed to discover five cluster structure (indicated by l_1 on fig. 2.). In order to compare partitions adjusted Rand index was calculated. $AR = 1,$ this result means complete agreement between partitions.

3. **Third data set** is an adaptation of well-known cuboids data set (from `mlbench` package of R software) – 100 objects in three dimensions, divided into four not well separated clusters.

A clustering ensemble with 15 partitions $(N = 15)$ was obtained by running: the partition around medoids (*pam*) algorithm, hierarchical clustering methods (ward, single, complete, mcquitty, median and centroid) – all based on Hausdorff distance for symbolic interval-valued data – with number of clusters k randomly chosen in the interval $[2;18]$. A co-association matrix was build, and then used as data matrix for the *k*-means method. The number of clusters *k* varied from 2 to 11, and for each cluster structure the Silhouette, Baker & Hubert and Hubert & Levine internal cluster quality indices were calculated. The best results were obtained for four cluster structure. Adjusted Rand index for this cluster structure was equal to $0{,}99996$.

V. FINAL REMARKS

Ensemble approach based on co-association matrix can be applied to cluster symbolic data of different types. In empirical part an ensemble technique which combines results of different clustering algorithms (with different cluster numbers chosen at random) was applied. The obtained clustering where used to build co- association matrix. This matrix was then applied as data matrix for single-link and *k*-means methods.

For purposes of the research an R source code was prepared. It allows to obtain different base clusterings and build co- association matrix. For all data sets used in the empirical part ensemble clustering reached high values of adjusted Rand index.

The aim for the future research will be to compare the efficiency of ensemble clustering methods when dealing data with noisy variables and outliers. Also other ways to obtain different base clusterings will be under study.

REFERENCES

Bock H.-H., Diday E. (red.) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*, Springer Verlag, Berlin-Heidelberg.

Billard L., Diday E. (2006), *Symbolic data analysis: conceptual statistics and data mining*, John Wiley & Sons Inc., Chichester.

Fred A.L.N. (2001), *Finding consistent clusters in data partitions*, [w:] Kittler J., Roli F. (eds) *Multiple Classifier Systems*, Vol. 1857 of Lecture Notes in Computer Science, Springer-Verlag, Berlin-Heidelberg, p. 78–86.

Fred A.L.N., Jain A.K. (2005), *Combining multiple clustering using evidence accumulation*, „IEEE Transaction on Pattern Analysis and Machine Intelligence", Vol. 27, p. 835–850.

Gahemi R., Sulaiman N., Ibrahim H., Mustapha N. (2009), *A survey: Clustering ensemble techniques* [w:] *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 38, p. 636–645.

Jain A.K., Murty M.N., Flynn P.J. (1999), Data clustering: A review, "ACM Computing Surveys", Vol. 31, No. 3, p. 264–323.

Pełka M. (2012), *Ensemble approach for clustering of interval-valued symbolic data*, Statistics in Transition, Volume 13, Number 2, s. 335–342.

Rozmus D. (2010), *Comparison of clustering accuracy in ensemble approach based on co-occurence data*, [in:] Fink A., Lausen B., Seidel W., Ultsch A. (eds), *Advances in Data Analysis, Data Handling and Business Intelligence*, Springer-Verlang, Berlin-Heidelberg, p. 174–184.

Strehl A., Gosh J. (2002), *Cluster ensembles – a knowledge reuse framework for combining multiple partitions*, "Journal of Machine Learning Research", Vol. 3 (Dec), p. 587–617.

Marcin Pełka

KLASYFIKACJA DANYCH SYMBOLICZNYCH Z WYKORZYSTANIEM PODEJŚCIA WIELOMODELOWEGO

Podejście wielomodelowe oparte na agregacji modeli jest z powodzeniem wykorzystywane w zagadnieniach dyskryminacyjnych i regresyjnych. Niemniej jednak podejście to może zostać także zastosowane w zagadnieniu klasyfikacji. W wielu artykułach wskazuje się, że połączenie wielu różnych klasyfikacji pozwala otrzymać lepsze wyniki.

Artykuł przedstawia możliwość zastosowania podejścia wielomodelowego w klasyfikacji danych symbolicznych. W artykule przedstawiono także wyniki klasyfikacji z wykorzystaniem podejścia wielomodelowego.