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

Andrzej Dudek[*]

CLASSIFICATION OF LARGE DATA SETS. COMPARISON OF PERFORMANCE OF CHOSEN ALGORITHMS

Abstract. Researchers analyzing large (> 100,000 objects) data sets with the methods of cluster analysis often face the problem of computational complexity of algorithms, that sometimes makes it impossible to analyze in an acceptable time. Common solution of this problem is to use less computationally complex algorithms (like k-means), which in turn can in many cases give much worse results than for example algorithms using eigenvalues decomposition . The results of analysis of the actual sets of this type are therefore usually a compromise between quality and computational capabilities of computers. This article is an attempt to present the current state of knowledge on the classification of large datasets, and identify ways to develop and open problems.
Key words: Clustering, classification, large data sets.

I. INTRODUCTION

Researchers analyzing large (> 100,000 objects) data sets with the methods of cluster analysis often face the number of problems that make analysis very hard or even impossible. Computational complexity of algorithms, sometimes makes it impossible to analyze in an acceptable time. The other limitation is memory size of standard PC-like computers, which in many cases may be too small for necessary calculations on such data sets. Thus not all clustering algorithms may be used for those kind of data.

The article is divided into five parts with introduction. First part presents which clustering algorithms can and cannot be used for large data sets in popular statistical **R** framework. The second part describes known strategies of clustering of such data sets. The third and fourth part present computational simulation results on one million objects data matrices with known cluster structure for typical and untypical cluster shapes.

[*] Ph.D., Chair of Econometrics and Informatics, University of Economics, Wrocław.

II. LIMITATIONS OF LARGE DATA SETS CLASSIFICATION

To identify computational limitations of most popular clustering methods they were executed on data set with one million objects, formed into four partitions. All calculations have been made in **R** statistical environment, but parallel experiments identified the same limitations in other popular statistical packages as well as in source computer program written in C++ language. The following clustering algorithms have been examined:

- hierarchical agglomerative methods;
- hierarchical divisive method (diana);
- k-means algorithm;
- partition around medoids (pam, k-medois algorithm);
- spectral clustering approach (Ng, Jordan, Weiss (2002));
- ensemble approach (Dimitriadou, Weingessel, Hornik (2001));

To pass this preliminary test two simple conditions should been fulfilled:

I. method execution should not report any lack of memory error.

II. method should not run longer than five hours.

In this preliminary step clustering quality has not been measured, this will be done (for methods that passed the preliminary test) in chapter IV and V. The computer used for the simulation was Quad Core 2.6 GHz, 4 GB RAM, Windows 7 64b. The result of experiment was the following:

hierarchical agglomerative methods:

```
>library(mlbench)
>smileys<-mlbench.smiley(1000000)$x
> cutree(hclust(dist(smileys),"ward"),4)
Error in double(N * (N - 1)/2) : vector size specified is
too large
```

The algorithm (algorithms) has not passed due to memory lack error;

hierarchical divisive method (diana):

```
> diana(dist(smileys),3)
Error in double(N * (N - 1)/2) : vector size specified is
too large
```

The algorithm has not passed due to memory lack error;

k-means algorithm

The algorithm execution time was 6.6 s. with no memory errors;

partition around medoids (pam, k-medois algorithm)

```
> pam(smileys,4)
Error in double(1 + (n * (n - 1))/2) : vector size
specified is too large
```

The algorithm has not passed due to memory lack error;

spectral clustering approach

```
> speccl(smileys,4)
Error in double(nrow(x) * nrow(x)) : vector size cannot be NA
In addition: Warning message:
In nrow(x) * nrow(x) : NAs produced by integer overflow
```

The algorithm has not passed due to memory lack error;

ensemble approach

```
ens<-NULL
numberOfEssembles<-100
numberOfClusters<-4
for(i in 1:numberOfEssembles){
  if(sample(1:2,1)%%2 ==1){
        ens<-
c(ens,cl_ensemble(clara(smileys,numberOfClusters)))
  }
  else{
    ens<-c(ens,cl_ensemble(kmeans(smileys,numberOfClusters)))
  }
}
cons<-cl_consensus(ens)
clusters<-as.vector(cl_class_ids(cons))
```

The method execution time was longer than 5 hours (and after five hours it has been canceled).

III. STRATEGIES OF LARGE DATA SETS CLASSIFICATION

Four strategies of clustering large data sets can be distinguished:

Divide et impera **strategy** described in chapter 3 of Kaufman and Rousseeuw (1990) Compared to other partitioning methods such *as k-medoids*, it can deal with much larger datasets. Internally, this is achieved by considering sub-datasets of fixed size such that the time and storage requirements become linear rather than quadratic. Each sub-dataset is partitioned into k clusters using the same algorithm as in *k-medoids* method.

Once k representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid. The sum of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset for which the sum is minimal, is retained. A further analysis is carried out on the final partition. Each sub-dataset is forced to contain the medoids obtained from the best sub-dataset until then.

Clustering Large Arrays (CLARA) algorithm is probably best known representative of this strategy.

Iterative linear calculation strategy contains classical algorithm as k-means clustering and neural networks (self-organizing maps). The common feature of all algorithms in this group is sequential processing of data in each iteration step. Methods from this group have linear calculation time and memory usage, but they may give insufficient results for untypical cluster shapes.

Sampling strategy randomly generates subsets of large data set. Those subsets are clustered with one of clustering algorithm like k-means and final results are achieved by aggregation of partial models.

Symbolic approach strategy aggregates objects from data sets in so-called symbolic objects (see Bock, Diday (2000), Diday, Noirhome-Fraiture (2008)). Clustering of symbolic objects requires special, more effective than traditional algorithm but final partitions contains also symbolic objects and its verification and interpretation is different than in classical case. For this reason this strategy will be omitted in next chapters.

IV. STANDARD CLUSTERS SHAPES EXAMPLE

In the experiment *CLARA* (first strategy) , k-means method (second strategy) and k-means with sampling (third strategy) results and performance time have been compared on artificial data set with known cluster structure, generated from `clusterSim` package (Walesiak, Dudek (2011)). Table 1 contains average Rand Index and average execution time for 50 simulations.

Table 1. Results of experiment I (clusters given from multidimensional normal distribution)

Strategy	Adjusted Rand	Average execution time
I	0,7999866	6,05 s.
II	0,7495783	9,80 s.
III	0,337106	13 m. 51 s.

Source: own calculations.

First strategy gives best clustering adequacy measured in corrected Rand Index values as well as shortest running time. The second ("pure" k-means) gives results not much worse. We can consider these results as acceptable. The sampling strategy gives much worse results and average corrected Rand Index value at 0,33 level shows very unstable resulting cluster structure, thus we can qualify this strategy as insufficient.

Despite fact that third strategy gives unacceptable results in relatively long time, first two strategies behaves quite well and in reasonable time. So, in case of standard clusters shapes clustering of data sets with 100 000 or even

1 000 000 is no problem and "classical" clustering methods are fully sufficient. Situation complicates when clusters have untypical, not given from multidimensional normal distribution, shapes. In next experiment performance of three described earlier strategies will be compared on those kind of data sets.

V. UNTYPICAL CLUSTER SHAPES EXAMPLE

In following experiment CLARA (first strategy), k-means method (second strategy) and k-means with sampling (third strategy) results and performance time have been compared on artificial data set with known cluster structure and with untypical cluster shapes: spirals and smileys generated from `mlbench` package and own data set – MSA. The sample representations of each model are presented on figure 1.

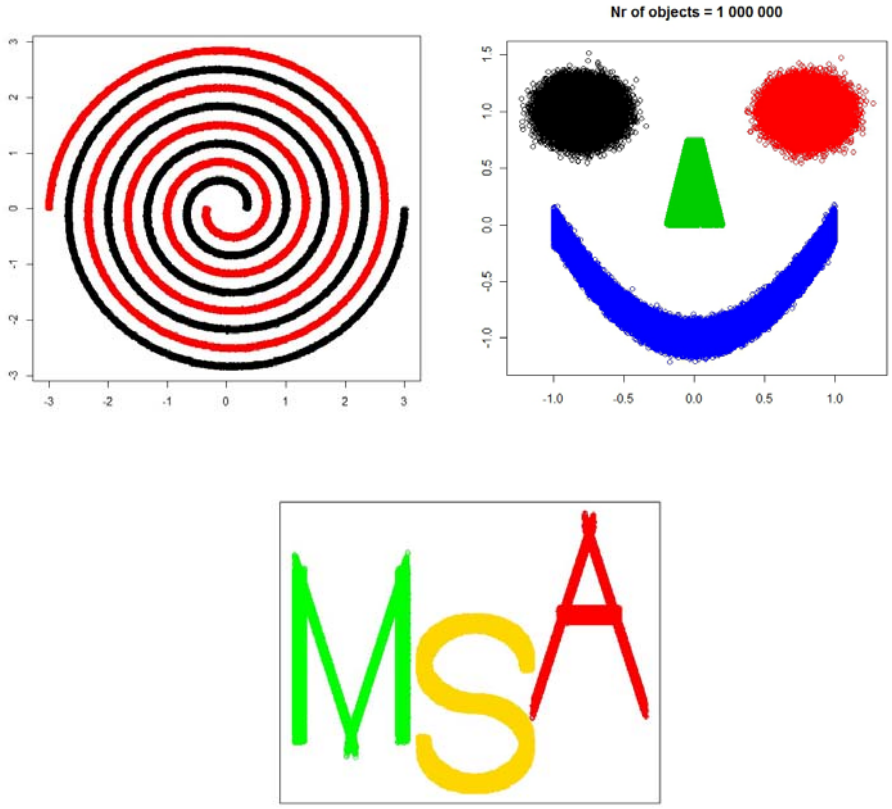

Figure 1. Sample data sets used in experiment II

Source: `mlbench` package and own source.

Table 2 contains average Rand Index and average execution time for common 50 simulations on 3 models.

Table 2. Results of experiment II (untypical cluster shapes)

Strategy	Adjusted Rand	Average execution time
I	0,513475	8,14 s.
II	0,475634	9,94 s.
III	0,296519	14 m .23 s.

Source: own calculations.

The order of clustering results and performance time is similar to previous experiment but it is hard consider the results as satisfying. The average adjusted Rand index value at 0,51 level indicates rather not stable clusters. We can observe on figure two, representing clustering results of one of the simulations steps that resulting clusters are not adequate to real cluster structure.

Figure 2. Partial clustering results on MSA data sets in first iteration step.
Source: own source.

VI. FINAL REMARKS AND OPEN PROBLEMS

The problem of classification of large data may be divided into two groups. Classification of large data sets with typical, given from normal distribution, shapes and classification of large data sets with untypical non ellipsoid-like clusters. While in first case “standard” clustering algorithms give satisfying results in acceptable time, the second type of classification needs further

development. We can state that at the moment there is no effective clustering algorithm for these kind of data. From previous experiment (see for example Walesiak, Dudek (2009)) we can assume that spectral clustering approach may be appropriate, but due to need of calculation of eigenvalues in spectral clustering process it has very large memory and time requirements and (see chapter II) is not working for data sets with million or more objects in standard statistical environments.

From the other hand, CLARA algorithm and "*divide et impera*" strategy gives best (but not sufficient) results from examined method. So we can assume that future development should direct to combine features of CLARA algorithm and spectral approach.

REFERENCES

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

Diday E., Noirhomme-Fraiture M. (eds.) (2008), *Symbolic Data Analysis with SODAS Software*, John Wiley & Sons, Chichester.

Dimitriadou E., Weingessel A., Hornik K. (2001), Voting-Merging: An Ensemble Method for Clustering. [in:] G. Dorffner, H. Bischop, K. Hornik (eds.), *Artificial Neural Networks – ICANN 2001,* Lecture Notes in Computer Science volume 2130 Springer, Berlin / Heidelberg, 217–224

Everitt B.S., Landau S., Leese M. (2001), *Cluster analysis*, Edward Arnold, London.

Gordon A.D. (1999), *Classification*, Chapman & Hall/CRC, London.

Hubert L.J., Arabie P. (1985), *Comparing partitions*. „Journal of Classification", no. 2, 193–218.

Kaufman L., Rousseeuw P.J. (1990), *Finding groups in data: an introduction to cluster analysis*, Wiley, New York.

Ng A., Jordan M., Weiss Y. (2002), *On spectral clustering: analysis and an algorithm*, [w:] T. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, 849–856.

Walesiak M., Dudek A. (2010), *Klasyfikacja spektralna z wykorzystaniem odległości GDM,* Prace Naukowe UE we Wrocławiu nr 107, 161–171.

Walesiak M., Dudek A. (2011), `clusterSim` *package*, URL http://www.R-project.org.

Andrzej Dudek

KLASYFIKACJA DUŻYCH ZBIORÓW PORÓWNANIE WYDAJNOŚCI WYBRANYCH ALGORYTMÓW

Badacze analizujący przy pomocy metod analizy skupień duże (> 100.000 obiektów) zbiory danych, stają często przed problemem złożoności obliczeniowej algorytmów, uniemożliwiającej niekiedy przeprowadzenie analizy w akceptowalnym czasie. Jednym z rozwiązań tego problemu jest stosowanie mniej złożonych obliczeniowo algorytmów (*hierarchiczne aglomeracyjne, k-średnich*), które z kolei mogą w wielu sytuacjach dawać zdecydowanie gorsze rezultaty niż np. algorytmy wykorzystujące dekompozycję względem wartości własnych. Rezultaty rzeczywistych analiz tego typu zbiorów są więc zazwyczaj kompromisem pomiędzy jakością a możliwościami obliczeniowymi komputerów. Artykuł jest próbą przedstawienia aktualnego stanu wiedzy na temat klasyfikacji dużych zbiorów danych oraz wskazania dróg rozwoju i problemów otwartych.