Evaluation of Two-Step Spectral Clustering Algorithm for Large Untypical Data Sets

Andrzej Dudek

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. In the article, the new algorithm from spectral clustering family is proposed and compared with other approaches.

Keywords Clustering · Classification · Large data sets · Spectral clustering

1 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 that kind of data.

The article is divided into four parts with introduction. First part presents which clustering algorithms can or cannot be used for large data sets in popular statistical **R** framework. The second part is a proposal of modification of spectral clustering procedure. Third part present computational simulation results on over 100,000 objects data matrices with known cluster structure for untypical cluster shapes against the proposed algorithm. The final part contains remarks and conclusions.

A. Dudek (✉)
Wrocław University of Economics and Business, Wrocław, Poland
e-mail: andrzej.dudek@ue.wroc.pl

K. Jajuga et al. (eds.), *Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-75190-6_1

2 Limitations of Large Data Sets Classification

Dudek (2013) has examined the following clustering algorithms on one million object multivariate normal distribution data set:

- hierarchical agglomerative methods,
- hierarchical divisive method (diana),
- k-means algorithm,
- partition around medoids (pam, k-medoids algorithm),
- spectral clustering approach (von Luxburg 2006),
- ensemble approach (Dimitriadou et al. 2001).

Only one algorithm (k-means) has passed the following requirements in **R** environment:

- method execution should not report any lack of memory error,
- method should not run longer than five hours.

But in further analysis for untypical cluster shapes, k-means has given the results that not meet the actual structure of clusters.

3 Proposal of New Algorithm

Spectral decomposition algorithm according to von Luxburg (2006) and Ng et al. (2002) can be stated in its general form in the following way:

Let X means data matrix with n rows and m columns, u—number of cluster to divide X (given by researcher before start of decomposition). Sample input data is presented on Fig. 1. Next figures will be showing the same data in transformed space.

Let **A** be similarity matrix of objects from **X**. **A** can be calculated in many ways but most often its elements a_{ij} are defied according to Eq. 1:

$$a_{ij} = e^{-\frac{\sum_{k=1}^{m}\left(x_{ik}-x_{jk}\right)^2}{\sigma}} \tag{1}$$

where: σ—scaling parameter. Most often it is calculated according to Ng et al. (2002) algorithm of iterative choosing of σ, minimalizing the with-class distances of random subset (random rows selected) of **X**: **X′** (this method requires processing of approximately few hundreds clustering procedures of objects in **X′**),

n—number of rows,

m—number of columns,

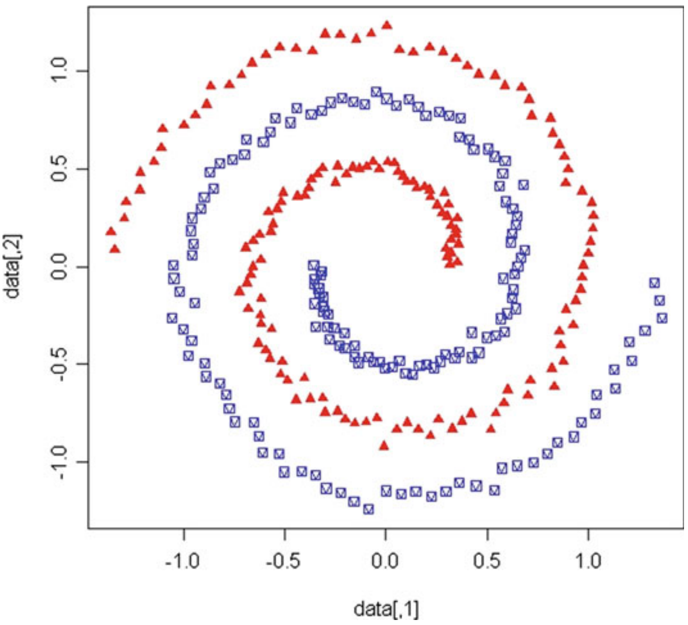

Fig. 1 Input data before spectral decomposition. *Source* Own elaboration with use of *mlbench* **R** library

$$i,j = 1,\ldots,n; k = 1,\ldots,m.$$

For **A** weights matrix **W** is constructed due to Eq. 2:

$$w_{ij} = \begin{cases} \sum_{j=1}^{n} a_{ij} & if \quad i = j \\ 0 & if \quad i \neq j \end{cases} \tag{2}$$

where: **W** = $[w_{ij}]$—weights matrix.

Laplacian **L** is calculated next according to Eq. 3:

$$\boldsymbol{L} = \boldsymbol{W}^{-\frac{1}{2}} \times \mathbf{A} \times \boldsymbol{W}^{-\frac{1}{2}} \tag{3}$$

L can be treated as algebraic representation of graph created from objects of **X**.

First *u* eigenvectors of Laplacian **L** creates **E** matrix. Each eigenvector is treated as column of **E** (thus **E** ma has dimensions $n \times u$). The main aim of this step is to widen data in transformed space (see Fig. 2).

Optional matrix **E′** is a result of normalization of **E** due to Eq. 4. This step is narrowing data in transformed space (it can be observed on Fig. 3).

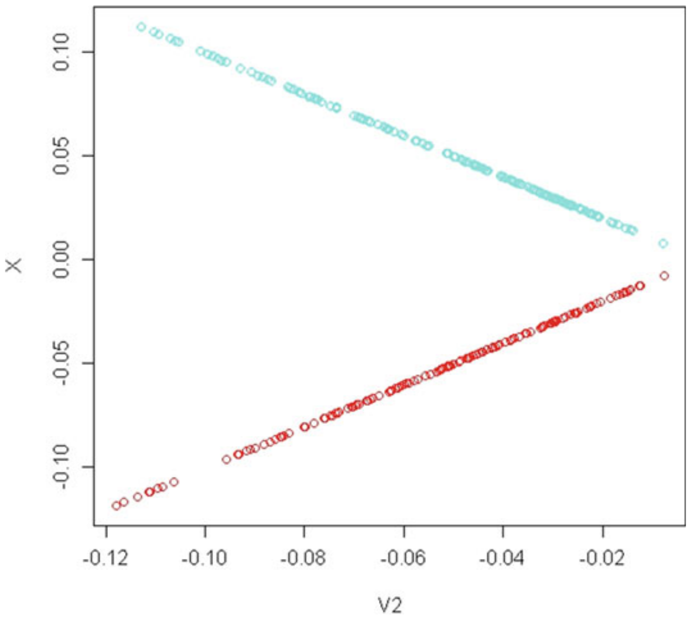

Fig. 2 Data in transformed space. *Source* Own elaboration with use of *mlbench* **R** library

$$E'_{ij} = \frac{E_{ij}}{\sqrt{\sum_{k=1}^{n} E_{kj}^2}} \tag{4}$$

In last stage, **E**′ (or **E** if normalization step is omitted) is clustered with one of "standard" algorithms. Most often *k*-means is used for this purpose.

The two-step spectral clustering algorithm (TSSC) is a try-out of avoiding computational limitations of spectral clustering algorithm. The approach behind it is similar to that used by Shinnou and Sasaki (2008) and Kong et al. (2011).

It can be stated in four steps:

1. Pre-cluster the data set with k-means into given number (500 or 1000) of small clusters (inter-clusters).
2. Calculate the medoids of each cluster.
3. Run normal clustering procedure on cluster medoids from first step.
4. Finally assign object from original data set to the cluster to which belongs the medoid of inter-cluster.

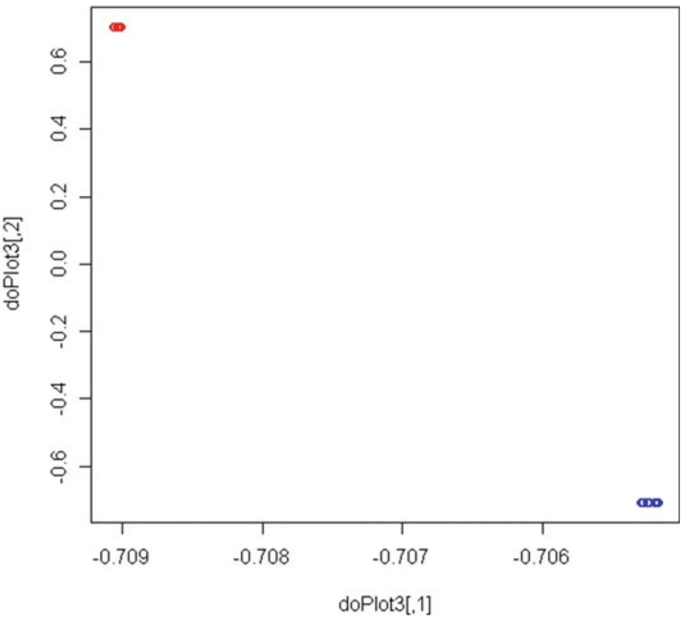

Fig. 3 Data in transformed space after normalization step. *Source* Own elaboration with use of *mlbench* **R** library

4 Simulation Experiment Results

For measuring the quality of methods based on spectral decomposition, an experiment has been carried out. In all simulations, the adjusted Rand (Hubert and Arabie 1985) index has been used for measuring the quality of clustering. The experiment examines this technique in case of non-standard cluster shapes.

In the experiments, the results of clustering with use of spectral decomposition, two-step spectral clustering (TSSC), *k*-means algorithm, *k*-medoids algorithm, Ward clustering, and complete link clustering have been compared on five data models: *Spirals, Worms, WWW, Smiley, Cassini*. For each model, fifty random realizations have been generated with use of *mlbench* **R** library. Each data set consists of more than 100,000 objects and is untypical in sense that they are not generated from any distribution mixture. Figure 4 shows sample data sets generated from each model.

Table 1 shows the results of simulation. In all cases, clustering based on spectral decomposition has found the more accurate class structure.

Only two from compared methods have not crushed due to memory limits. From those two, the newly proposed TSSC algorithm gave better results in every case. In two cases (spirals and w3), the difference of performance is at very high level, while for three other, the results measured in average adjusted Rand are comparable, but against with significant TCSS advantage.

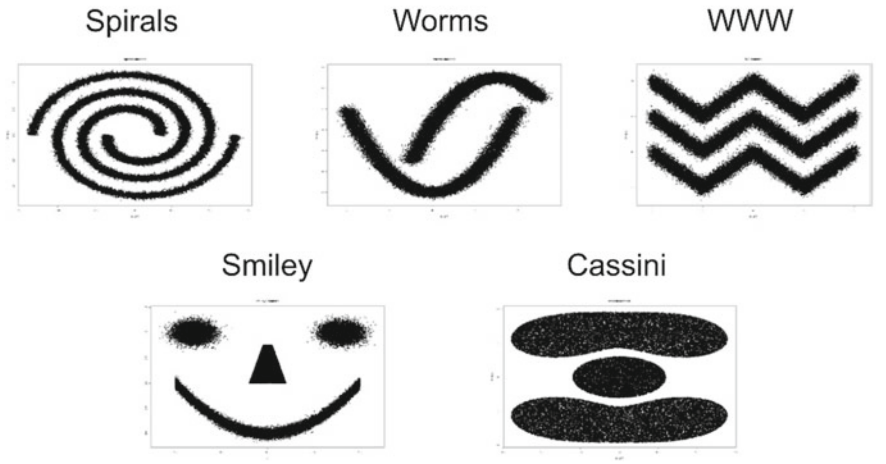

Fig. 4 Data sets used in first experiment. *Source* Own elaboration with use of *mlbench* **R** library

Table 1 Average adjusted Rand values from 50 simulations

	Nr of objects	k-means	k-medoids	Ward	Complete link	Spectral clust	TSSC
Spirals	200,000	0.0350	*	*	*	*	1
Worms	200,000	0.5116	*	*	*	*	0.9999
w3	300,000	0.0049	*	*	*	*	0.9359
Smiley	120,000	0.7981	*	*	*	*	0.8534
Cassini	300,000	0.8157	*	*	*	*	0.9917

*—Calculational/memory complexity exceeded

5 Final Remarks and Conclusions

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. In the paper, new algorithm is evaluated for such data sets, and the results of experimental analysis are very promising.

Author is aware that "there is no free lunch" and is far for acclaiming this method as "the best clustering algorithm" but sometimes it can behave better that standard well-known methods of cluster analysis.

References

Dimitriadou E, Weingessel A, Hornik K (2001) Voting-merging: an ensemble method for clustering. In: Dorffner G, Bischop H, Hornik K (eds) Artificial neural networks—ICANN 2001. Lecture notes in computer science, vol 2130. Springer, Heidelberg, pp 217–224

Dudek A (2013) Classification of large data sets. Comparison of performance of chosen algorithms. Acta Universitatis Lodziensis. Folia Oeconomica 285:71–78

Hubert LJ, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Kong T, Tian Y, Shen H (2011) A fast incremental spectral clustering for large data sets, pp 1–5. https://doi.org/10.1109/PDCAT.2011.4

Ng A, Jordan M, Weiss Y (2002) On spectral clustering: analysis and an algorithm. In: Dietterich T, Becker S, Ghahramani Z (eds) Advances in neural information processing systems 14. MIT Press, pp 849–856

Shinnou H, Sasaki M (2008) Spectral clustering for a large data set by reducing the similarity matrix size. In: Proceedings of the sixth international conference on language resources and evaluation (LREC), pp 201–2014

von Luxburg U (2006) A tutorial on spectral clustering. Max planck institute for biological cybernetics, Technical Report TR-149