ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 135, 2010

Andrzej Dudek[*]

CLASSIFICATION VIA SPECTRAL CLUSTERING

Abstract. Spectral clustering is known since end of twentieth century and is developing quite fast. This method gives very good empirical results on artificial and real data, despite lack of strong theoretical basement in few places. Article presents main steps of spectral clustering algorithm and points situations where spectral clustering gives much better results (measured by adjusted Rand index) than other clustering techniques. Finally some recommendations of usage of spectral clustering are given.

Key words: cluster analysis, spectral clustering

I. INTRODUCTION

Spectral clustering is not strictly new partitioning algorithm but rather new way of preparing data for known clustering methods such as k-means. Despite this fact spectral clustering (or clustering based on spectral decomposition) gives very promising result for cluster other than given by normal distribution. In this paper algorithm of spectral clustering is described (in most popular form) and some empirical result of use of this method for standard and non-standard cluster shapes are given

First chapter places spectral clustering among other clustering techniques.

Second part describes in details clustering decomposition, which is the heart of spectral clustering methods.

Part three to five describes results of empirical simulations investigating quality of spectral clustering vs. other clustering techniques in case of standard and non-standard cluster shapes, in case of data with noisy variables and in case of data set with nominal data.

Finally some conclusions and remarks are given

II. SPECTRAL CLUSTERING AS PARTITIONING ALGORITHM

There is big number of known clustering algorithm but few of them are most popular (and most widely used by researchers).

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

Hierarchical aggregative clustering methods (Gordon (1999) p. 79)):

- Ward hierarchical clustering,
- single link hierarchical clustering,
- complete link hierarchical clustering,
- average link hierarchical clustering,
- Mcquitty (1966) hierarchical clustering,
- centroid hierarchical clustering.

Optimization methods:

- *k*-means method with variants like Isodata algorithm, hard competitive learning or neural gas (soft competitive learning).
- Partitioning around medoids, also called *k*-medoids method (Kaufman, Rousseeuw (1990)).

Since the end of twentieth century two new branches has been developed quite fast:

- Model – based clustering
- Spectral clustering

Spectral clustering, based on spectral decomposition of distance matrix is not in fact new method but rather new way of preparing data for "classical" k-means procedure, but empirical results for different cluster shapes are very promising and this "branch" of clustering is worth separate studies. For understanding the idea of spectral methods one should start with the description of decomposition mentioned above.

III. SPECTRAL DECOMPOSITION

Spectral decomposition algorithm according to von Luxburg (2006) and Ng.*et al.* (2001) can be stated in 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 figure 1. Next figures will be showing the same data in transformed space.

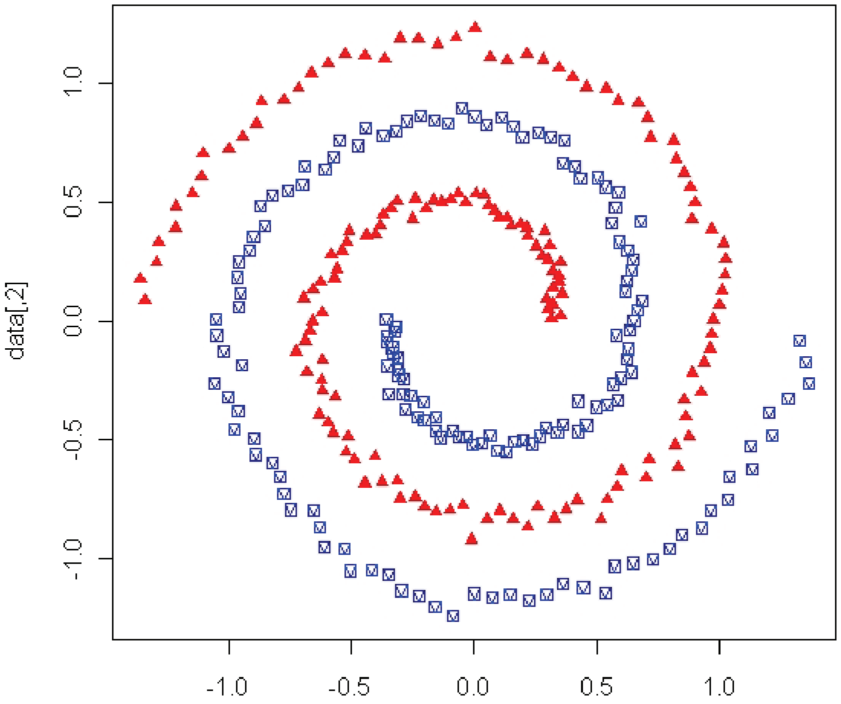

Figure 1. Input data before spectral decomposition.

Source: Own research with use of *mlbench* **R** library.

Let $\mathbf{D}$ be similarity matrix of objects from $\mathbf{X}$. $\mathbf{D}$ can be calculated in many ways but most often its elements d_{ij} are defied according to (1)

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

where: σ^2 – scaling parameter. Most often its calculated according to Ng, Jordan and Weiss algorithm of iterational choosing of σ^2, minimalizing the inner class distances of random subset of $\mathbf{X}$: $\mathbf{X}'$. (this method requires proceedings of approximately few hundreds clustering procedures of objects of $\mathbf{X}'$).

For $\mathbf{D}$ weights matrix $\mathbf{W}$ is constructed due to (2)

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

where: $\mathbf{W} = [w_{ij}]$ – weights matrix,

and Laplacian $\mathbf{L}$ according to (3)

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

In graph theory $\mathbf{L}$ is treated as algebraical representation of graph created from objects of $\mathbf{X}$

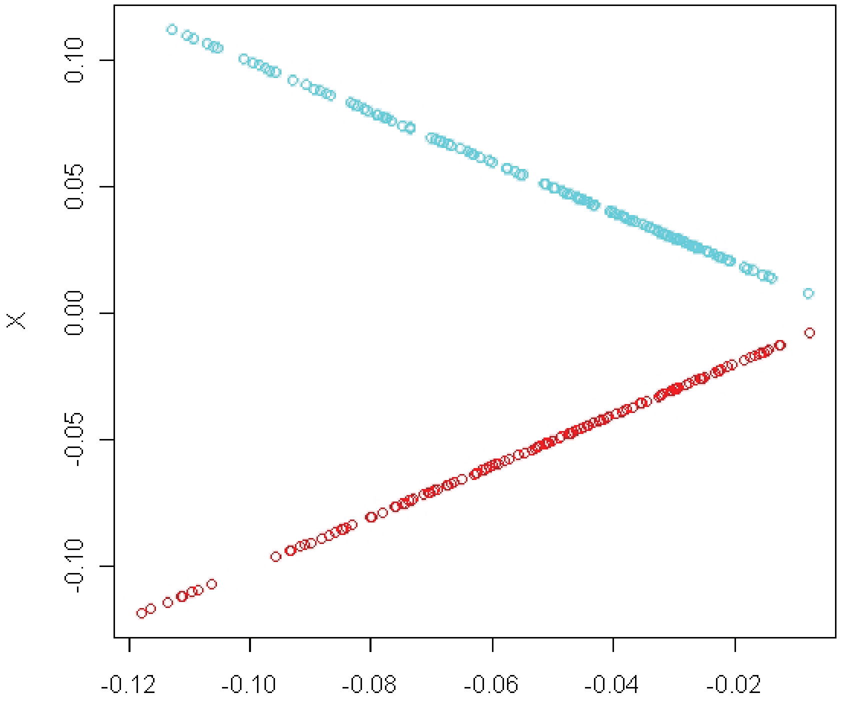

Figure 2. Data in transformed space

Source: Own research with use of *mlbench* **R** library.

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

Optional matrix $\mathbf{E}^{'}$ is a result of normalization of $\mathbf{E}$ due to (4). This step is narrowing data in transformed space (it can be observed on figure 3)

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

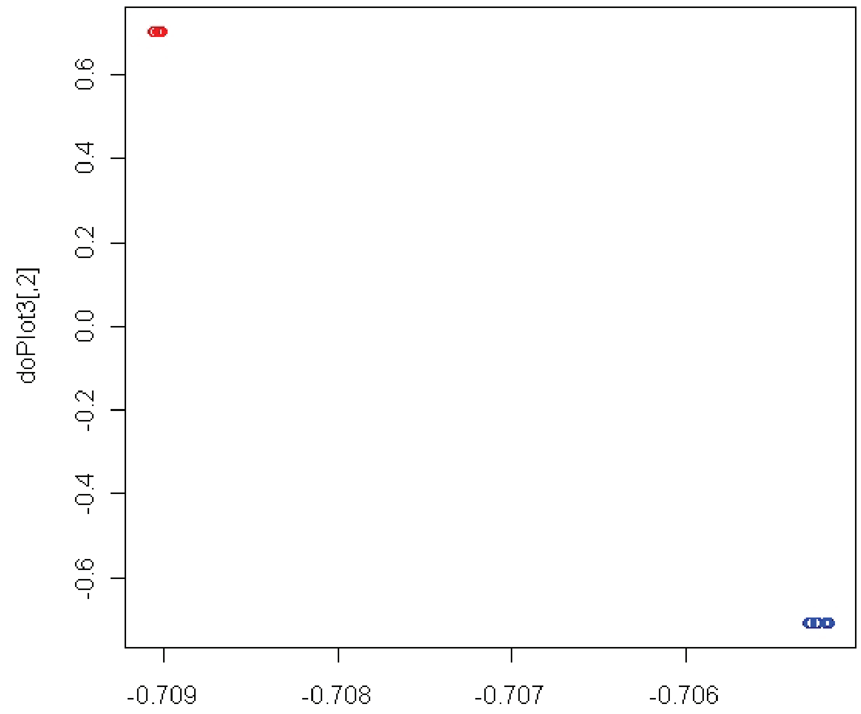

Figure 3. Data in transformed space after normalization step.
Source: Own research with use of *mlbench* **R** library.

In last stage $\mathbf{E}^{'}$ (or $\mathbf{E}$ if normalization step is omitted) is clustered with one of "standard" algorithm. Most often *k*-means is used for this purpose.

IV. SPECTRAL CLUSTERING IN CASE OF TYPICAL AND NON-TYPICAL CLUSTER SHAPES

For measuring the quality of methods based on spectral decomposition several experiments have been carried out. In all simulations the adjusted Rand

(Hubert, Arabie (1985)) index has been used for measuring the quality of clustering. First two experiments examines this technique in case of non-standard and standard (given from multivariate normal distribution) cluster shapes.

In first experiments the results of clustering with use of spectral decomposition, *k*-means algorithm, *k*-medoids algorithm, Ward clustering, and complete link clustering have been compared on 5 data models: *Spirals, Worms, WWW, Smiley*, *Cassini*. For each model fifty realizations has been generated with use of *mlbench* **R** library. Figure 4 shows sample datasets generated from each model.

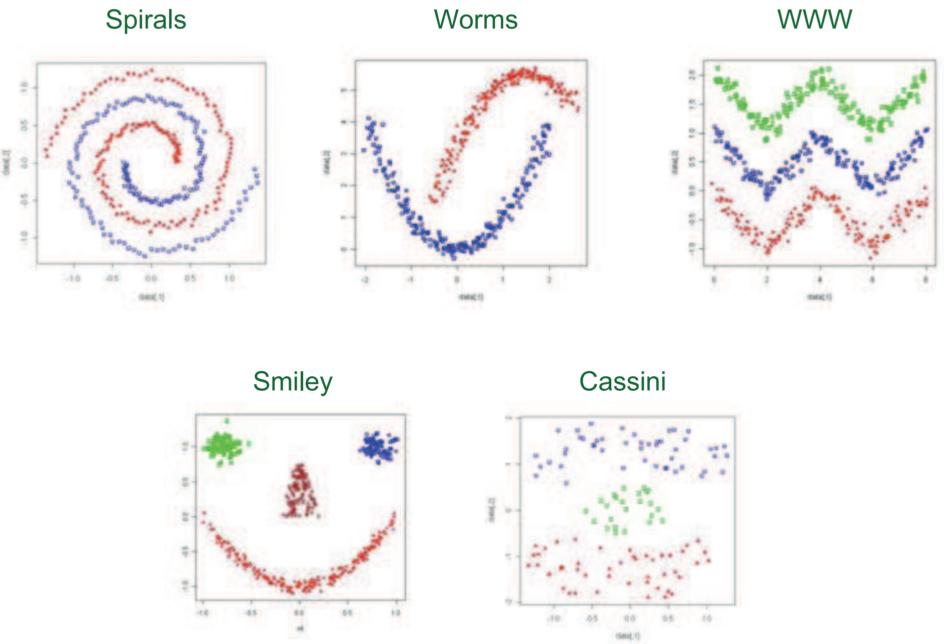

Figure 4. Datasets used in first experiment .

Source: Own research with use of *mlbench* **R** library.

Table 1 shows the results of simulation. In 4 of 5 cases clustering based on spectral decomposition gives the most accurate class structure. In fifth case spectral clustering loses only with Ward method.

Table 1. Average adjusted Rand values from 50 simulations

	k–means	*k–medoids*	Ward	complete link.	spectral clust.
spirals	0,032917	0,025898	0,040443	0,052412	**1**
worms	0,519405	0,505834	0,53292	0,491536	**1**
w3	–0,00271	–0,00157	0,008503	0,010088	**0,949382**
smiley	0,712618	0,808584	**1**	0,645165	0,952397
cassini	0,583186	0,80901	0,875653	0,583428	**0,896623**

Source: Own research.

The second simulation compares clustering results of model corners from *mlbench* **R** library (3-dimensional spherical normal distribution with equal standard deviation and means at the corners of a 3-dimensional hypercube. The number of classes is *8*).

Table 2. Average adjusted Rand values from 50 simulations for *corners* dataset

	k-means	*k-medoids*	Ward	complete link.	spectral clust.
	0,8345	**1**	**1**	**1**	0,9377

Source: Own research.

One can observe that while clustering based on spectral decomposition gives very good results in case of cluster shapes (experiment 1) its performance is getting worse for ellipsoidal clusters given from normal distribution (experiment 2).

V. SPECTRAL CLUSTERING VS. NOISY VARIABLES

In third experiment for five models of *clusterSim* **R** library (Walesiak Dudek (2008) – models 8 to 12) fifty realizations have been generated with one, two, thee, and four noisy variables. The results of this simulation presents table 3.

Table 3. Average adjusted Rand values from 50 simulations in experiment 3

Model	noisy variables	Clustering methods				
		k-means	*k-medoids*	Ward	complete link.	spectral clust.
VIII	1	0,629077	0,787007	0,735939	0,394949	**0,803979**
	2	0,456183	0,599806	0,573116	0,273773	**0,73285**
	3	0,398277	0,451854	0,444257	0,192093	**0,637395**
	4	0,339029	0,299987	0,391039	0,18647	**0,579096**
IX	1	0,737759	**0,991415**	0,97134	0,533588	0,868137
	2	0,684393	**0,910343**	0,903599	0,399407	0,798072
	3	0,62989	0,758684	**0,832027**	0,400814	0,759928
	4	0,645731	0,64405	0,727011	0,393227	**0,744193**
X	1	0,645402	**0,978026**	0,947038	0,350163	0,900259
	2	0,50338	**0,882869**	0,834369	0,096393	0,840437
	3	0,300671	0,744025	0,720378	0,128239	**0,928347**
	4	0,334466	0,51415	0,538178	0,107489	**0,832613**
XI	1	0,844613	0,998363	**1**	0,406477	0,8784
	2	0,844928	**0,994847**	0,994404	0,351526	0,923779
	3	0,807329	0,946388	**0,981421**	0,362097	0,859247
	4	0,750618	0,818999	**0,94802**	0,28082	0,877913
XII	1	0,465162	0,527189	0,39321	0,318616	**0,845875**
	2	0,377781	0,372313	0,35115	0,271025	**0,632185**
	3	0,312744	0,253072	0,294639	0,178948	**0,428952**
	4	**0,285008**	0,197342	0,232718	0,182356	0,260269

Source: Own research.

In general the results of this experiment lead to conclusion: The more noisy variables the better behavior of spectral clustering.

VI. SPECTRAL CLUSTERING FOR NOMINAL DATA

Finally the quality of spectral clustering for datasets containing variables measured on nominal scales has been examined. Once again *cluster.Gen* models of *clusterSim* **R** library has been used, this time for generating datasets with variables measured for nominal scales (each variable in each dataset has seven categories). The results of clustering with use of spectral decomposition, *k*-means algorithm, *k*-medoids algorithm, Ward clustering, and complete link clustering are presented in table 4.

Table 4. Average adjusted Rand values from 50 simulations in experiment 4

	k-means	*k-medoids*	Ward	complete link.	spectral clust.
Model 5	0,937	0,977599	1	0,573738	0,952191
Model 6	0,859297	0,938542	0,898959	0,79931	0,820793
Model 7	0,774414	0,969944	0,946314	0,827362	0,884087
Model 8	0,759305	0,991232	0,962294	0,947591	0,848093
Model 9	0,874804	0,996221	0,992574	0,978128	0,798206
Model 10	0,902385	1	1	1	0,892882

Source: Own research.

As we can observe for such kind of data the usage of spectral techniques is rather useless. In general it gives much worse results than *k-medoids* and hierarchical methods.

VII. FINAL REMARKS

Results of experiments empower to state thesis that methods based on spectral decomposition gives better results than "traditional" algorithms in case of untypical cluster shapes and for datasets with noisy variables while they don't improve (and sometimes decrease) results of classification for ellipsoidal clusters given from normal distribution and for datasets with variables measured on nominal scales.

In real research problems regular shapes are rather rare while non-standard clusters appear more often, so we can also state that this method may be for researchers a good alternative to widely used k-means, *k*-medoids or hierarchical clustering algorithms.

REFERENCES

Climescu-Haulica A. (2006) How to choose the number of clusters. The Cramer Multiplicity Solution, W: H.H.-J. Lenz, R. Decker (Eds.), *Advances in Data Analysis*, Berlin, pp. 15–23.

Cristianini N., Kandola J. (2001), Spectral Methods for Clustering*, Neural Information Processing Symposium*, available at http://www.nips.cc/NIPS2001/papers/psgz/AA35.ps.gz

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

Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE, Wrocław.

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

Hartigan J.A. (1975), *Clustering algorithms*, Wiley, New York, London, Sydney, Toronto.

Hubert L.J., Arabie P. (1985), Comparing partitions, *Journal of Classification*, no. 1, 193-218.

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

von Luxburg U. (2006), *A tutorial on spectral clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.

Ng A., Jordan I., Weiss Y (2001), On Spectral Clustering: Analysis and an Algorithm *Neural Information Processing Symposium*, available at http://www.nips.cc/NIPS2001/ papers/ psgz/AA35.ps.gz

Walesiak M., Dudek A. (2008), *ClusterSim,* **R** package available at http://wgrit.ae.jgora.pl/ keii/clusterSim

Andrzej Dudek

ANALIZA SKUPIEŃ METODAMI KLASYFIKACJI SPEKTRALNEJ

Klasyfikacja spektralna to rozwijająca się od końca poprzedniego wieku metoda analizy skupień. Metoda ta, mimo niekiedy niezbyt rozbudowanej podbudowy teoretycznej, daje bardzo dobre wyniki empiryczne zarówno na zbiorach testowych jak i na rzeczywistych zbiorach danych. Artykuł przedstawia najważniejsze kroki algorytmu klasyfikacji spektralnej, wskazuje sytuacje, w których stosowanie algorytmu daje duże lepsze rezultaty (mierzone indeksem Randa) niż inne metody analizy skupień. W zakończenie przedstawione są rekomendacje dotyczące sytuacji, w których warto stosować tą technikę klasyfikacji.