PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 468 • 2017

Taksonomia 28
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: marcin.pelka@ue.wroc.pl

WIELOMODELOWA KLASYFIKACJA SPEKTRALNA DANYCH SYMBOLICZNYCH

ENSEMBLE SPECTRAL CLUSTERING FOR SYMBOLIC DATA

DOI: 10.15611/pn.2017.468.18
JEL Classification: C38, C87

Streszczenie: Klasyfikacja spektralna, którą zaproponowali Ng, Jordan i Weiss [2002], jest nie tyle nową metodą klasyfikacji, ile nowym podejściem do przygotowywania danych na potrzeby klasyfikacji, która wykorzystuje ideę dekompozycji spektralnej macierzy danych. Głównym celem artykułu jest zastosowanie klasyfikacji spektralnej na potrzeby podejścia wielomodelowego w analizie skupień danych symbolicznych oraz przeprowadzenie i analiza symulacji w tym zakresie. Klasyfikacja spektralna może znaleźć zastosowanie zarówno w przygotowaniu danych na potrzeby utworzenia macierzy współwystąpień (*co-association matrix*), jak i w samej klasyfikacji dokonywanej na podstawie tej macierzy, a także jako metoda przygotowywania danych na potrzeby adaptacji metody *boosting* w klasyfikacji. W części empirycznej artykułu zaprezentowano i zinterpretowano wyniki klasyfikacji wielomodelowej z zastosowaniem klasyfikacji spektralnej zarówno do przygotowania danych wejściowych, jak i samej klasyfikacji. Wykorzystano tu sztuczne zbiory danych o znanej strukturze klas.

Słowa kluczowe: klasyfikacja wielomodelowa, klasyfikacja spektralna, dane symboliczne.

Summary: Spectral clustering that was proposed by Ng, Jordn and Weiss [2002], is not a new clustering method, but a new approach to data preparation that uses the idea of data decomposition. The main aim of the paper is to present an application of spectral clustering in ensemble clustering for symbolic data. Spectral clustering can be used before building co-association matrix and to this matrix. Spectral clustering can be also combined with bagging for clustering. In the empircal part data sets with known cluster structures are used.

Keywords: ensemble clustering, spectral, clustering, symbolic data.

1. Wstęp

Podejście wielomodelowe może być z powodzeniem stosowane w zagadnieniach dyskryminacyjnych i regresyjnych (zob. np. [Gatnar 2008; Kuncheva 2014]). Niemniej jednak idea podejścia wielomodelowego, która polega na łączeniu wyników

otrzymanych z zastosowaniem różnych modeli (metod), może być z powodzeniem stosowana także w zagadnieniu analizy skupień danych symbolicznych.

Celem artykułu jest zaproponowanie zastosowania klasyfikacji spektralnej na potrzeby podejścia wielomodelowego w analizie skupień danych symbolicznych. Klasyfikacja spektralna może znaleźć zastosowanie zarówno w przygotowaniu danych na potrzeby utworzenia macierzy współwystąpień (*co-association matrix*), jak i w samej klasyfikacji dokonywanej na podstawie tej macierzy, a także jako metoda przygotowywania danych na potrzeby różnych adaptacji metody *boosting* w klasyfikacji.

W części empirycznej artykułu zaprezentowano wyniki klasyfikacji wielomodelowej z zastosowaniem klasyfikacji spektralnej zarówno do przygotowania danych wejściowych, jak i samej klasyfikacji. Wykorzystano tu sztuczne zbiory danych o znanej strukturze klas.

2. Dane symboliczne

Obiekty symboliczne, w przeciwieństwie do obiektów w ujęciu klasycznym, mogą być opisywane przez następujące rodzaje zmiennych [Bock, Diday (red.) 2000, s. 2-3; Billard, Diday 2006, s. 7-30; Dudek 2013, s. 35-36; Diday, Noirhomme-Fraiture 2008, s. 10-19] (przykładowe realizacje zmiennych symbolicznych zawarto w tabeli 1):

- zmienne nominalne, porządkowe, przedziałowe oraz ilorazowe,
- zmienne interwałowe – czyli przedziały liczbowe,
- zmienne wielowariantowe – czyli listy kategorii lub wartości,
- zmienne wielowariantowe z wagami – czyli listy kategorii z wagami,
- zmienne histogramowe – czyli listy wartości z wagami.

Tabela 1. Przykładowe zmienne symboliczne wraz z realizacjami

Zmienna symboliczna	Realizacje zmiennej	Typ zmiennej
Preferowana cena samochodu w tys. zł	<20; 35>, <27; 37>, <30; 45>, <28; 42>	interwałowa (przedziały nierozłączne)
Preferowana pojemność silnika w cm^3	<1000; 1200), <1200; 1400), <1400; 1600) <1600; 1800), <1800; 2000); <2000; 2200)	interwałowa (przedziały rozłączne)
Preferowany kolor	{czarny, czerwony, zielony}, {czerwony, żółty, różowy},{czarny}	wielowariantowa
Preferowana marka	{60% Honda, 40% Audi}, {100% Volkswagen}	wielowariantowa z wagami
Płeć respondenta	K, M	nominalna
Czas dojazdu do pracy	{60% <10; 20>, 40% <30; 40>} {20% <10; 20>, 80% <30; 40>}	histogramowa

Źródło: opracowanie własne (dane sztuczne).

Szerzej o obiektach i zmiennych symbolicznych, sposobach otrzymywania zmiennych symbolicznych z baz danych, różnicach i podobieństwach między obiektami symbolicznymi a klasycznymi piszą m.in.: [Bock, Diday (red.) 2000, s. 2-8; Dudek 2013, s. 42-43; 2004; Billard, Diday 2006, s. 7-66; Noirhomme-Fraiture, Brito 2011; Diday, Noirhomme-Fraiture 2008, s. 3-30].

3. Wielomodelowa klasyfikacja spektralna

Podstawowy algorytm klasyfikacji spektralnej zaproponowano w pracy Ng, Jordan i Weiss [2002]. Nazwa tej metody odnosi się do jednego z jej podstawowych kroków, w którym wyznacza się zbiór wartości własnych macierzy Laplace'a, nazywany w matematyce spektrum (widmem) (zob. np. [Walesiak 2013]). W literaturze przedmiotu zaproponowano wiele różnych modyfikacji klasyfikacji spektralnej – m.in. w pracach Shorteed [2006] czy Walesiaka i Dudka [2009].

W badaniach porównawczych klasyfikacja spektralna często osiąga znacznie lepsze rezultaty niż tradycyjne metody klasyfikacji. Wynika to z faktu, że nie przyjmuje się w niej żadnych założeń co do kształtu skupień. Dodatkowo klasyfikacja spektralna w większości przypadków daje lepsze rezultaty dla skupień o nietypowych kształtach (zob. np. [von Luxburg, Bousquet, Belkin 2005]).

Klasyfikacja spektralna dla danych symbolicznych składa się z następujących kroków[1]:

1. Konstrukcja tablicy danych symbolicznych $\mathbf{V} = \left[v_{ij} \right]$ o wymiarach $n \times m$ ($i = 1, ..., n$ – numer obiektu, $j = 1, ..., m$ – numer zmiennej).
2. Zastosowanie estymatora jądrowego do obliczenia macierzy podobieństw (*affinity matrix*) pomiędzy obiektami. Najczęściej do wyznaczania macierzy podobieństw stosowany jest estymator gaussowski:

$$A_{ik} = \exp\left(-\sigma \cdot d_{ik}\right), \quad i,k = 1,\ldots,n, \tag{1}$$

gdzie: d_{ik} – odległość pomiędzy i-tym i k-tym obiektem symbolicznym, σ – parametr skali (szerokość pasma).

W części empirycznej wykorzystano znormalizowaną miarę odległości Ichino-Yaguchiego oraz parametr $\sigma = 2$ dla wszystkich symulacji.

3. Obliczenie diagonalnej macierzy **D**, w której na głównej przekątnej znajdują się sumy każdego z wierszy macierzy **A**, a poza nią zera.
4. Wyznaczenie znormalizowanej macierzy Laplace'a:

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{2}$$

[1] Opracowanie własne na podstawie prac: [Walesiak, Dudek 2009, s. 12-14; Dudek 2013, s. 78; Walesiak 2013, s. 34-35].

5. Obliczenie wartości własnych oraz odpowiadających im wektorów własnych dla macierzy **L**. Uporządkowanie wektorów własnych według malejących wartości własnych. Pierwsze u z nich (gdzie: u – liczba klas) tworzy macierz **E**.

6. Znormalizowanie macierzy **E** zgodnie ze wzorem:

$$y_{ij} = e_{ij} \Big/ \sqrt{\sum_{j=1}^{u} e_{ij}^{\,2}}, \tag{3}$$

gdzie: $i = 1, ..., n$ – numer obiektu, $j = 1, ..., m$ – numer zmiennej, u – liczba klas.

Macierz **Y** stanowi punkt wyjścia do zastosowania jednej z metod klasyfikacji (najczęściej jest to metoda k-średnich).

W literaturze przedmiotu zaproponowano wiele różnorodnych podejść, które umożliwiają zastosowanie klasyfikacji spektralnej w ramach podejścia wielomodelowego, które można podzielić na dwie główne grupy (zob. [Huang i in. 2015; Liu i in. 2015; Jia i in. 2011; Jia, Liu, Jiao 2011]):

1. Zastosowanie podejścia spektralnego do macierzy danych lub tablicy danych symbolicznych, a następnie:
- wykorzystanie jednej z adaptacji metody *bagging* (jeden z trzech algorytmów: Leischa, Hornika lub Doudoit i Fridlyandy) na potrzeby klasyfikacji,
- utworzenie macierzy współwystąpień i dokonanie klasyfikacji na jej podstawie,
- utworzenie macierzy współwystąpień i ponowne zastosowanie podejścia spektralnego do tej macierzy.

2. Utworzenie na podstawie tablicy danych symbolicznych macierzy współwystąpień i zastosowanie podejścia spektralnego do tej macierzy.

4. Badania symulacyjne

Na potrzeby badań symulacyjnych przygotowano pięć zbiorów danych o znanej strukturze klas (z wykorzystaniem funkcji `clusterGen` z pakietu `clusterSim` programu R (zob. [Walesiak, Dudek 2015]). W tabeli 2 zawarto charakterystykę pięciu przygotowanych zbiorów danych.

Na podstawie każdego modelu wygenerowano 20 zbiorów danych, a następnie przeprowadzono klasyfikację wielomodelową z wykorzystaniem znormalizowanej odległości Ichino-Yaguchiego, które okazały się skuteczne w klasyfikacji spektralnej danych symbolicznych interwałowych dla zbiorów danych różnych typów (zob. [Pełka 2014]). Do ustalenia ostatecznej liczby klas zastosowano indeks Calińskiego-Harabasza (w klasyfikacji spektralnej można zastosować różne indeksy jakości klasyfikacji – zob. np. [Walesiak 2013]), za każdym razem rozważano podziały od 2 do 10 klas. Wyniki porównano ze znaną strukturą klas za pomocą skorygowanego indeksu Randa (zob. [Hubert, Arabie 1985]).

Na rysunku 1 przedstawiono graficzną prezentację zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` dla danych

Tabela 2. Charakterystyka modeli w analizie symulacyjnej

lp	*nm*	*m*	*u*	*lo*	Środki ciężkości klas	Macierz kowariancji **Σ**
1	3	2	2	50	(0, 0), (1, 5)	$\sigma_{jj} = 1$, $\sigma_{jl} = -0{,}9$
2	4	2	3	50	(0, 0), (1,5, 7), (3, 14)	$\sigma_{jj} = 1$, $\sigma_{jl} = -0{,}9$
3	6	2	5	55	(5, 5), (–3, 3), (3, –3), (0, 0), (–5, –5)	$\sigma_{jj} = 1$, $\sigma_{jl} = 0{,}9$
4	7	3	5	60, 80, 35, 45, 30	(5, 5, 5), (–3, 3, –3), (3, –3, 3), (0, 0, 0), (–5, –5, –5)	$\sigma_{jj} = 1$ $(1 \leq j \leq 3)$, $\sigma_{jl} = 0{,}9$ $(1 \leq j \neq l \leq 3)$,
5	9	3	5	60	(0, 0, 0), (10, 10, 10), (–10, –10, –10), (10, –10, 10), (–10, 10, 10)	$\sigma_{jj} = 3$ $(1 \leq j \leq 3)$, $\sigma_{jl} = 2$ $(1 \leq j \neq l \leq 3)$,

lp – numer modelu; *nm* – numer modelu w funkcji `cluster.Gen` pakietu `clusterSim`; *m* – liczba zmiennych; *u* – liczba klas; *lo* – liczba obiektów w klasach (jedna liczba oznacza klasy równoliczne).

Źródło: opracowanie własne.

symbolicznych interwałowych. Osie przedstawiają zakres wartości zmiennych symbolicznych interwałowych w każdym z modeli.

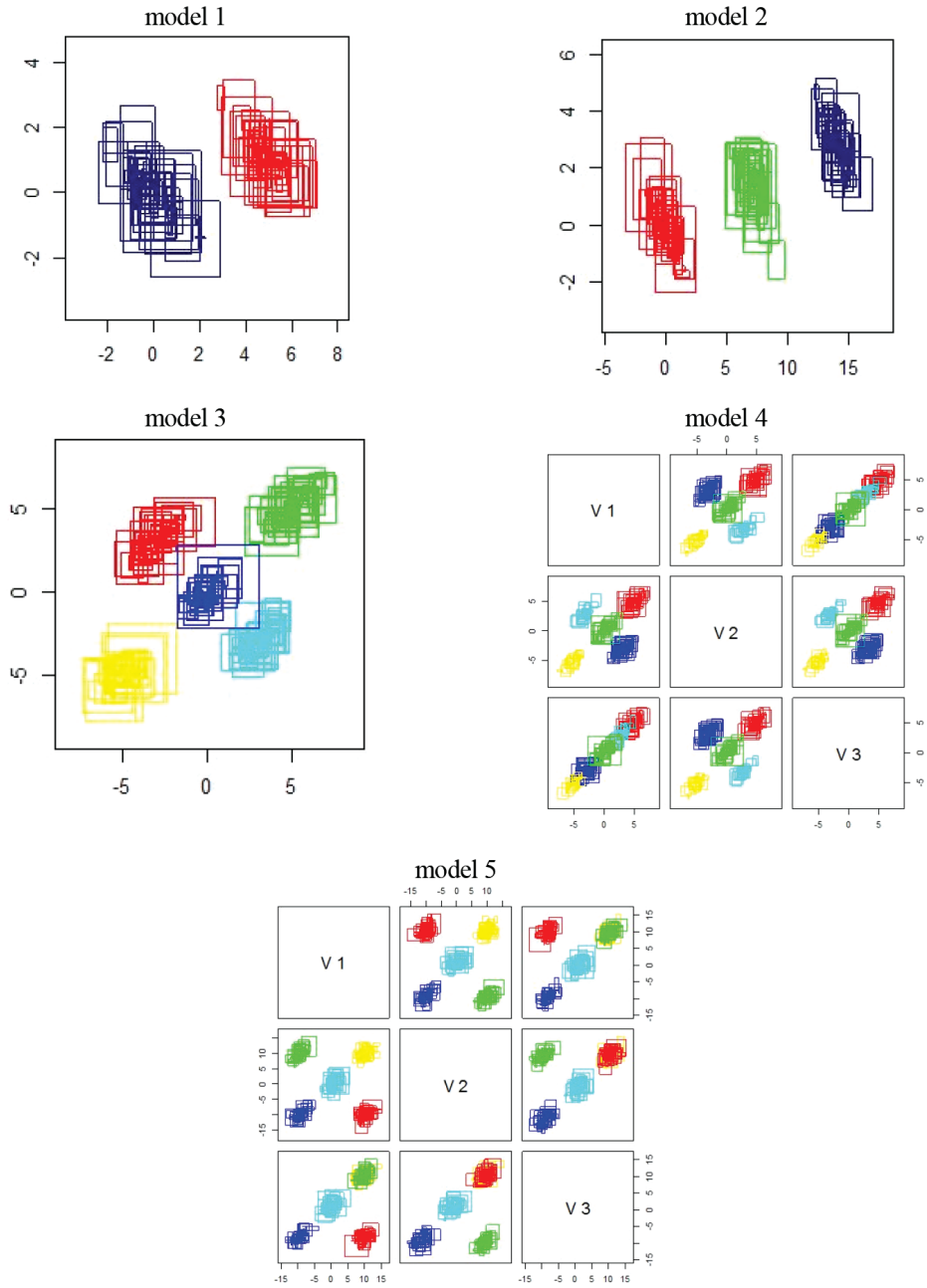
Tabela 3 prezentuje średnie, obszary zmienności oraz odchylenia standardowe wartości skorygowanego indeksu Randa dla sześciu różnych rozwiązań w zakresie wykorzystania klasyfikacji spektralnej w podejściu wielomodelowym i pięciu zbiorów danych oraz średnią wartość indeksu Randa dla każdej z metod w przekroju wszystkich zbiorów danych.

Tabela 3. Wyniki symulacji dla poszczególnych zbiorów danych

Zbiór danych (*lp*)	Ścieżka klasyfikacji					
	D-S-M-K	*D-S-M-S-K*	*D-M-S-K*	*D-S-BH*	*D-S-BL*	*D-S-BDF*
1	*2*	*3*	*4*	*5*	*6*	*7*
1	0,9767	0,9898	0,2032	0,9863	0,9871	0,9655
2	0,9883	0,9814	0,3092	0,9882	0,9798	0,9876
3	0,9563	0,9465	0,1102	0,9453	0,9567	0,9509
4	0,9209	0,9330	0,1200	0,9400	0,9514	0,9439
5	0,9754	0,9678	0,3303	0,9654	0,9831	0,9873
Średnia	0,9635	0,9637	0,2146	0,9650	0,9716	0,9670
Odchylenie standardowe	0,0264	0,0237	0,0102	0,0224	0,0164	0,0202
Obszar zmienności	0,0674	0,0568	0,2201	0,0482	0,0357	0,0437

nm – numer zbioru danych w funkcji `cluster.Sim` pakietu `clusterGen` programu R; *D* – tablica danych symbolicznych; *S* – podejście spektralne; *M* – macierz współwystąpień; *K* – metoda *k*-średnich; *BH* – adaptacja metody *bagging* Hornika; *BL* – adaptacja metody *bagging* Leischa; *BDF* – adaptacja metody *bagging* Dudoit i Fridlyandy.

Źródło: opracowanie własne.



Rys. 1. Graficzna prezentacja zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` (dane symboliczne interwałowe).

Źródło: opracowanie własne z wykorzystaniem programu R.

Najlepsze wyniki w sensie uśrednionego skorygowanego indeksu Randa dla wszystkich zbiorów danych uzyskało podejście spektralne połączone z adaptacją metody *bagging* Leischa. Dość zbliżone wyniki uzyskały także inne adaptacje metod bagging – Hornika oraz Dudoit i Fridlyandy.

Także biorąc pod uwagę obszar zmienności (rozstęp) skorygowanego indeksu Randa najlepsze wyniki otrzymano dla adaptacji metody *bagging* Leischa oraz adaptacji *bagging* Hornika.

Najgorsze wyniki, w sensie średniego skorygowanego indeksu Randa, dla każdego ze zbiorów danych uzyskało podejście spektralne zastosowane do macierzy współwystąpień. Wyniki te mają jednocześnie duży rozstęp wartości oraz dość niskie odchylenie standardowe.

Zastosowanie podejścia spektralnego do danych symbolicznych (tablicy danych symbolicznych), a także połączenie zastosowania podejścia spektralnego do tablicy danych symbolicznych wraz z zastosowaniem klasyfikacji spektralnej do macierzy współwystąpień otrzymały zbliżone wyniki.

5. Podsumowanie

Przeprowadzone symulacje wskazują, że klasyfikacja spektralna może znaleźć z powodzeniem zastosowanie w klasyfikacji danych symbolicznych. Niemniej jednak, podobnie jak w przypadku danych klasycznych, istotnym zagadnieniem jest tutaj dobór wielkości parametru σ (szerzej o zagadnieniu doboru tego parametru piszą m.in. Walesiak [2013], Karatzoglu [2006]). Drugim ważnym zagadnieniem w przypadku danych symbolicznych jest wybór odpowiedniej miary odległości. Miary odległości dla danych symbolicznych omówione są m.in. w pracach Gatnara i Walesiaka [2011], Bocka i Didaya [2000].

Szerzej problematykę dotyczącą zagadnienia doboru miary odległości w klasyfikacji spektralnej danych symbolicznych interwałowych prezentuje praca Pełki [2014].

Podejście spektralne, bazujące na dekompozycji tablicy danych symbolicznych, może znaleźć zastosowanie na różnych etapach klasyfikacji.

Najlepsze wyniki klasyfikacji otrzymano łącząc podejście spektralne z adaptacją metody *bagging* Leischa. Najgorsze wyniki otrzymano dla podejścia spektralnego zastosowanego do macierzy współwystąpień zbudowanej na podstawie tablicy danych symbolicznych.

Literatura

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, Chichester.

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., Walesiak M. (red.), 2011, *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa.

Huang S., Wang H., Yang D., Li T., 2015, *Spectral co-clustering ensemble*, Knowledge-Based Systems, vol. 84, s. 46-55.

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

Jia J., Liu B., Jiao L., 2011, *Soft spectral clustering ensemble applied to image segmentation*, Frontiers of Computer Science in China, vol. 5(1), s. 66-78.

Jia J., Xiao X., Liu B., Jiao L., 2011, *Bagging-based spectral clustering ensemble selection*, Pattern Recognition Letters, no. 32, s. 1456-1467.

Karatzoglu A., 2006, *Kernel Methods. Software, Algorithms and Applications*. Rozprawa doktorska. Uniwersytet Techniczny w Wiedniu.

Kuncheva L., 2014, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley, Chichester.

Liu H., Liu T., Wu J., Tao D., Fu Y., 2015, *Spectral ensemble clustering*. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, s. 715-724.

Ng A., Jordan M., Weiss Y., 2002, *On Spectral Clustering: Analysis and an Algorithm*, [w:] Diettrich T., Becker S., Ghahramani Z. (red.), *Advances in Neural Information Processing Systems*, MIT Press, London, s. 849-856.

Noirhomme-Fraiture M., Brito P., 2011, *Far beyond the classical data models: Symbolic data analysis*, Statistical Analysis and Data Mining, vol. 4, issue 2, s. 157-170.

Pełka M., 2014, *Problematyka doboru miary odległości w klasyfikacji spektralnej danych symbolicznych*, Studia Ekonomiczne, nr 195/14, s. 140-150.

Shorteed S., 2006, *Learning in spectral clustering*. Rozprawa doktorska, Univeristy of Washington.

Von Luxburg U., Bousquet O., Belkin M., 2005, *Limits of spectral clustering*, [w:] Saul L., Weiss Y., Bottou L. (red.)., *Advances in Neural Information Processing Systems (NIPS) 17*, MIT Press Cambridge, MA, s. 857-864.

Walesiak M., 2013, *Zagadnienie doboru liczby klas w klasyfikacji spektralnej*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 278, s. 33-34.

Walesiak M., Dudek A., 2009, *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 84, s. 9-19.

Walesiak M., Dudek A., 2015, *The clusterSim package*, URL: http://www.R-project.org.