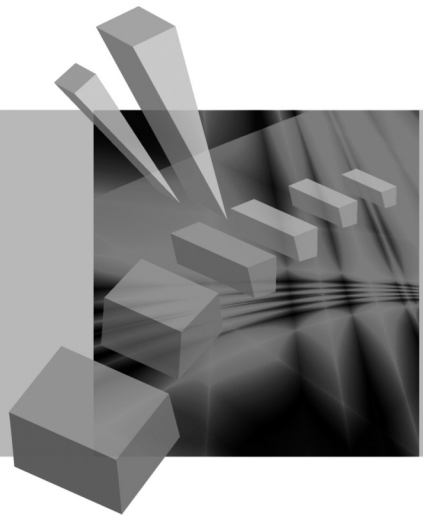
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Andrzej Dudek
Uniwersytet Ekonomiczny we Wrocławiu

SMS – PROPOZYCJA NOWEGO ALGORYTMU ANALIZY SKUPIEŃ

Streszczenie: Klasyfikacja spektralna (*spectral clustering* [Ng, Jordan, Weiss 2002; von Luxburg 2006]) i klasyfikacja za pomocą średniej przesunięcia okna w kierunku wektora średniej (*mean shift clustering* [Wang, Xiu, Zamar 2007]) to dwa stosunkowo nowe podejścia w analizie skupień, dające, zwłaszcza dla skupień o nietypowych kształtach, lepsze rezultaty niż klasyczne metody *k*-średnich, *k*-medoidów czy hierarchiczne metody aglomeracyjne. Artykuł zawiera propozycję algorytmu o roboczej nazwie SMS (*Spectral-Mean Shift*) łączącego cechy obu podejść i wyróżniającego się wśród innych algorytmów analizy skupień m.in.:
– możliwością analizy skupień o nietypowych kształtach;
– możliwością automatycznego rozpoznawania liczby skupień;
– lepszą odpornością na zmienne zakłócające.

Słowa kluczowe: analiza skupień, klasyfikacja spektralna, SMS.

1. Wstęp

Klasyfikacja spektralna (*spectral clustering* [Ng, Jordan, Weiss 2002; von Luxburg 2006]) i klasyfikacja za pomocą przesunięcia okna w kierunku wektora średniej (*mean shift clustering* [Wang, Xiu, Zamar 2007]) to dwa stosunkowo nowe podejścia w analizie skupień, dające, zwłaszcza dla skupień o nietypowych kształtach, lepsze rezultaty niż klasyczne metody *k-średnich*, *k-medoidów* czy hierarchiczne metody aglomeracyjne. W artykule zostanie zaproponowane połączenie obu podejść w algorytm o roboczej nazwie SMS (*Spectral – Mean Shift*).

2. Podejście spektralne

Klasyfikacja spektralna ([Ng, Jordan, Weiss 2002; Karatzoglou 2006; von Luxburg 2006]) to stosunkowo nowe i szybko rozwijające się podejście w analizie skupień. Nie jest ona ściśle nowym algorytmem, ale raczej sposobem przygotowania danych do znanych metod analizy skupień (np. *k-średnich*). Podejście to daje bardzo obiecujące rezultaty zwłaszcza dla nietypowych (tj. niepochodzących z rozkładu normalne-

go) kształtów. W większości badań technika ta jest stosowana w połączeniu z klasyczną metodą *k-średnich*, ale założenie to nie jest oparte na żadnej podstawie teoretycznej. Procedura klasyfikacji spektralnej ma wiele wariantów, dla wszystkich z nich można wyróżnić sześć najważniejszych etapów (według [Ng, Jordan, Weiss 2002, za Walesiak, Dudek 2009]):

1. Konstrukcja macierzy danych $\mathbf{X} = [x_{ij}]$ o wymiarach $n \times m$ ($i = 1, \ldots, n$ – numer obiektu, $j = 1, \ldots, m$ – numer zmiennej). Dla danych metrycznych należy przeprowadzić normalizację wartości zmiennych.

2. Zastosowanie estymatora jądrowego do obliczenia macierzy podobieństw $\mathbf{A} = [A_{ik}]$ (*affinity matrix*) między obiektami.

3. Konstrukcja znormalizowanej macierzy Laplace'a $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ($\mathbf{D}$ – diagonalna macierz wag, w której na głównej przekątnej znajdują się sumy każdego wiersza z macierzy $\mathbf{A} = [A_{ik}]$, a poza główną przekątną są zera). W rzeczywistości znormalizowana macierz Laplace'a przyjmuje postać: $\mathbf{I} - \mathbf{L}$.

4. Obliczenie wartości własnych i odpowiadających im wektorów własnych (o długości równej jeden) dla macierzy $\mathbf{L}$. Uporządkowanie wektorów własnych według malejących wartości własnych. Pierwsze u wektorów własnych (u – liczba klas) tworzy macierz $\mathbf{E} = [e_{ij}]$ o wymiarach $n \times u$.

5. Przeprowadzenie normalizacji tej macierzy zgodnie ze wzorem $y_{ij} = e_{ij} \Big/ \sqrt{\sum_{j=1}^{u} e_{ij}^{\,2}}$ ($i = 1, \ldots, n$ – numer obiektu, $j = 1, \ldots, u$ – numer zmiennej, u – liczba klas). Dzięki tej normalizacji długość każdego wektora wierszowego macierzy $\mathbf{Y} = [y_{ij}]$ jest równa jeden.

6. Macierz $\mathbf{Y}$ stanowi punkt wyjścia zastosowania klasycznych metod analizy skupień.

Poszczególne warianty algorytmów w podejściu spektralnym mogą się różnić (por. [Walesiak, Dudek 2009]):

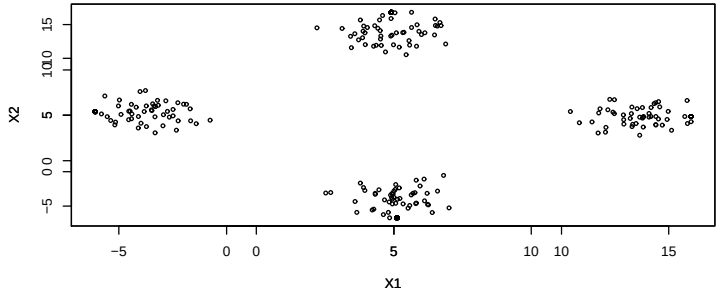
- typem estymatora jądrowego w kroku 2;
- formułą konstrukcji macierzy Laplace'a w kroku 4 (zob. np. [von Luxburg 2006]);
- określaniem wartości parametru σ – szerokości pasma (*kernel width*) liczby skupień.

W oryginalnej propozycji Ng, Jordan i Weiss [2002] postulują użycie algorytmu *k*-średnich jako ostatecznego algorytmu analizy skupień w kroku szóstym.

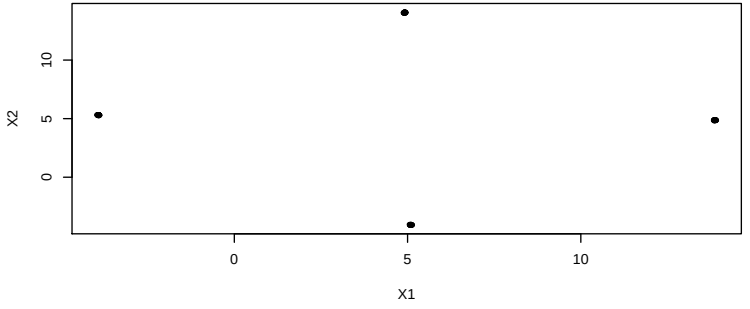
3. Klasyfikacja za pomocą przesunięcia okna w kierunku wektora średniej

Klasyfikacja za pomocą przesunięcia okna w kierunku wektora średniej to nieparametryczna technika analizy skupień iteracyjnie przesuwająca punkty *m*-wymiarowej przestrzeni euklidesowej w kierunku środków ciężkości skupień.

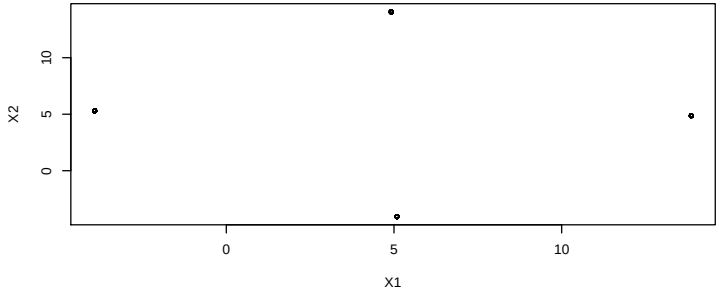
Dane pierwotne



Dane przesunięte w pierwszym kroku algorytmu



Dane przesunięte w drugim kroku algorytmu



Rys. 1. Dane w kolejnych rundach klasyfikacji za pomocą przesunięcia okna w kierunku wektora średniej

Źródło: opracowanie własne.

Niech $\{x_i\} i = 1, 2, \ldots, n$ będzie zbiorem danych w *m*-wymiarowej przestrzeni euklidesowej. Wektor przesunięcia średniej jest zdefiniowany (w ogólnej postaci) jako (1):

$$ms_h(x) = \frac{\sum_{i=1}^{n} -x_i k_x'\left(\left(\frac{x - x_i}{h}\right)^2\right)}{\sum_{i=1}^{n} -k_x'\left(\left(\frac{x - x_i}{h}\right)^2\right)} - x, \tag{1}$$

gdzie *k*' jest pochodną funkcji jądrowej (gaussowskiej, Bessela, Epanechnikova itd.). W szczególnym przypadku dla jądra Epanechnikova (por. [Comaniciu, Meer 1999; Korzeniewski 2005]) przyjmuje ona formę (2):

$$ms_h(x) = \sum_{k_x(x_i)>0} x_i - x, \tag{2}$$

gdzie:

$$k_x(x_i) = \begin{cases} \frac{3}{4}\left(1 - \sum_{j=j}^{m}\left(\frac{x_j - x_{ij}}{h}\right)^2\right) & d(x, x_i) < h \\ 0 & d(x, x_i) \geq h \end{cases}.$$

Wektor przesunięcia średniej zawsze skierowany jest w kierunku maksymalnego zwiększenia gęstości. W procedurze klasyfikacji z przesunięciem okna w kierunku wektora średniej w kolejnych krokach algorytmu najpierw obliczany jest wektor przesunięcia, a następnie współrzędne punktów są przesuwane w kierunku wyznaczonym przez ten wektor, aż do osiągnięcia konwergencji (punktów stacjonarnych). Schematyczny obraz procedury klasyfikacji z przesunięciem okna w kierunku wektora średniej przedstawia rys. 1.

4. Algorytm SMS

Algorytm *Spectral – Mean Shift clustering* łączy cechy obu podejść. Jego idea polega na tym, aby punkty wyznaczone przez znormalizowaną macierz wartości własnych macierzy Laplace'a iteracyjnie przesuwać w kierunku wyznaczonym przez wektor przesunięcia średniej. W postaci ogólnej algorytm ten można więc zapisać w siedmiu podstawowych krokach:

1. Konstrukcja macierzy danych.
2. Obliczenia macierzy podobieństw **A.**
3. Konstrukcja znormalizowanej macierzy Laplace'a.
4. Obliczenie wartości własnych i odpowiadających im wektorów własnych macierzy L. Utworzenie macierzy $\mathbf{E} = [e_{ij}]$.

5. Obliczenie znormalizowanej macierzy **Y.**

6. Iteracyjne przesuwanie punktów macierzy **Y** w kierunku wyznaczonym przez wektor przesunięcia średniej aż do osiągnięcia konwergencji (punktów stacjonarnych).

7. Właściwa klasyfikacja według reguły sekwencyjnej: *jeśli odległość obiektu od prototypu klasy jest mniejsza niż zadana wartość, obiekt zostaje dołączony do danego skupienia, jeżeli nie, tworzy prototyp nowej klasy.*

5. Rezultaty klasyfikacji za pomocą nowego algorytmu dla nietypowych kształtów skupień

W procedurze symulacyjnej analizy porównawczej porównano rezultaty klasyfikacji za pomocą nowego algorytmu z innymi algorytmami klasyfikacji dla podejścia spektralnego[1]. W eksperymencie wykorzystano sześć modeli zbiorów danych:

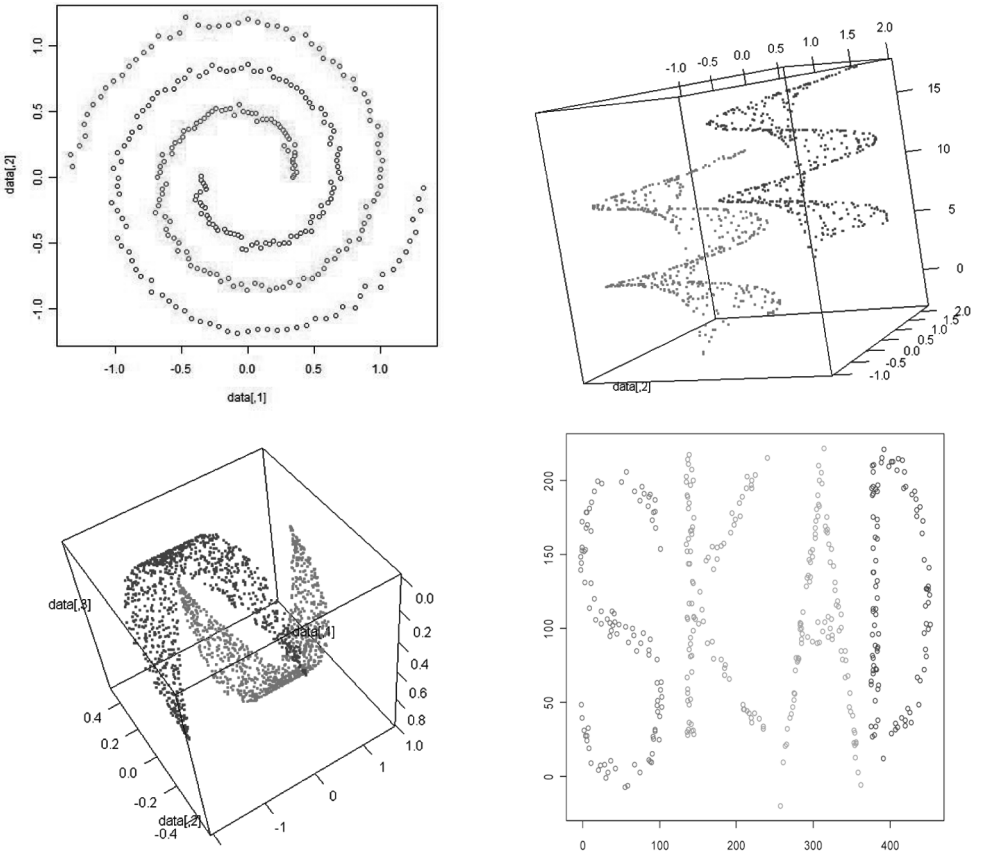
– zbiory *spirale* i *dini* pochodzą z pakietu `mlbench` środowiska **R**,
– zbiory *clusterSim1* i *clusterSim2* pochodzą z pakietu `clusterSim` środowiska **R**,
– zbiory *banany* i *SKAD* tworzone są przez autorskie funkcje programu **R**, niedołączone jeszcze do żadnej biblioteki tego środowiska.

Dla poszczególnych modeli w każdym eksperymencie wygenerowano 50 zbiorów danych, przeprowadzono procedurę klasyfikacyjną i porównano otrzymane rezultaty klasyfikacji ze znaną strukturą klas za pomocą skorygowanego indeksu Randa [Hubert, Arabie 1985]).

Uwzględniono następujące metody klasyfikacji: 1. `spec-kmeans` – klasyfikacja spektralna z metodą *k-średnich* użytą jako algorytm analizy skupień w kroku 6.; 2. `spec-ward` – klasyfikacja spektralna z hierarchiczną aglomeracyjną metodą powiększonej sumy kwadratów odległości użytą jako algorytm analizy skupień w kroku 6.; 3. `spec-mcquitty` – klasyfikacja spektralna z hierarchiczną aglomeracyjną metodą ważonej średniej klasowej użytą jako algorytm analizy skupień w kroku 6.; 4. `spec-pam`– klasyfikacja spektralna z metodą *k-medoidów* użytą jako algorytm analizy skupień w kroku 6.; 5. `spec-kmeans` – klasyfikacja spektralna z metodą *k-średnich* użytą jako algorytm analizy skupień w kroku 6.; 6. `SMS` – nowy algorytm.

Tabela 1 prezentuje uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 50 symulacji. W pięciu na sześć przypadkach nowy algorytm dał najlepsze wyniki. W szóstym przypadku wszystkie algorytmy podejścia spektralnego dały ten sam rezultat.

[1] Porównanie rezultatów klasyfikacji dla nietypowych skupień pomiędzy algorytmami podejścia spektralnego a „klasycznymi" algorytmami analizy skupień można znaleźć np. w pracach [Dudek 2009;Walesiak, Dudek 2010].



Rys. 2. Przykładowe zbiory danych utworzone z wykorzystaniem funkcji pakietu mlbench (*spirale*, *dini*) oraz zbiorów własnych (*banany*, *SKAD*)

Źródło: opracowanie własne z wykorzystaniem programu **R**.

Tabela 1. Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa

Metoda	Średnia wartość indeksu Randa/ kolejność		Zbiory danych											
			cluster-Sim1		cluster-Sim2		spirale		banany		dini		SKAD	
spec-kmeans	0,878	4	1	–	1	–	0,926	–	0,839	2-4	0,568	4	0,936	2-4
spec-ward	0,867	5	1	–	1	–	0,926	–	0,838	5	0,591	5	0,847	6
spec-mcquitty	0,883	2	1	–	1	–	0,926	–	0,838	6	0,6	6	0,936	2-4
spec-pam	0,877	3	1	–	1	–	0,926	–	0,839	2-4	0,564	2	0,936	2-4
spec-diana	0,866	6	1	–	1	–	0,926	–	0,839	2-4	0,567	3	0,868	5
SMS	0,893	**1**	**1**	–	1	–	0,926	–	**0,842**	**1**	**0,632**	**1**	**0,945**	**1**

Źródło: obliczenia własne z wykorzystaniem programu R.

Należy również nadmienić, iż nowy algorytm jako jedna z metod podejścia spektralnego daje lepsze rezultaty niż tradycyjne metody klasyfikacji w przypadku zbiorów zawierających zmienne zakłócające (por. np. [Dudek 2009]).

6. Wykrywanie liczby skupień

Algorytm SMS można zaproponować również w wersji znajdującej optymalną liczbę skupień dla danego zbioru danych. W tym celu należy nieznacznie zmodyfikować punkty 4-7 algorytmu i zastąpić je punktami 4'-8':

Obliczenie wartości własnych i odpowiadających im wektorów własnych dla macierzy **L**. Utworzenie macierzy **E** z pierwszych u wektorów własnych. Powtarzanie kroków 5-7 dla $u = 2,3, \ldots, n/2$.

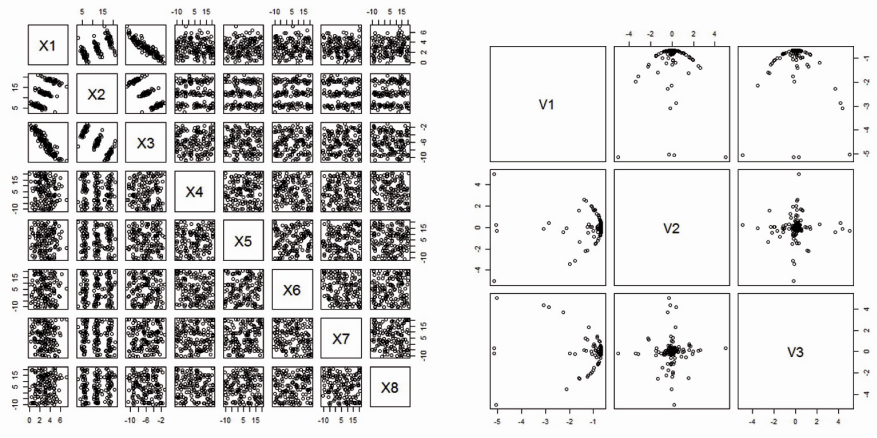
Obliczenie znormalizowanej macierzy **Y.**

Iteracyjne przesuwanie współrzędnych punktów z macierzy **Y** w kierunku wyznaczonym przez wektor przesunięcia średniej aż do osiągnięcia konwergencji (punktów stacjonarnych).

Właściwa klasyfikacja według reguły sekwencyjnej: *jeśli odległość obiektu od prototypu klasy jest mniejsza niż zadana wartość, obiekt zostaje dołączony do danego skupienia, jeżeli nie, tworzy prototyp nowej klasy.*

Wybór u, dla którego zadana liczba skupień (p. 4) jest równa rzeczywistej liczbie skupień (p. 7).

W przypadku gdy w kroku ósmym algorytmu nie można określić takiego u, że zadana liczba skupień jest równa rzeczywistej liczbie skupień, lub można znaleźć więcej niż jedną taką wartość u, do wyznaczenia optymalnej liczby klas należy wykorzystać jeden z indeksów jakości klasyfikacji i wskazać tę liczbę klas, dla której wartość indeksu jest optymalna (por. np. [Walesiak, Dudek 2007; Walesiak 2009, s. 418]).



Rys. 3. Przykładowy zbiór danych ze zmiennymi zakłócającymi przed transformacją spektralną i po niej

Źródło: opracowanie własne z wykorzystaniem programu **R**.

Należy jednak zaznaczyć, iż sytuacja taka występuje najczęściej, gdy dane pierwotne nie mają czytelnej struktury klas lub gdy występują w nich zmienne zakłócające. Przykładową taką sytuację przedstawia rys. 3.

Tabela 2. Wartości indeksu wskaźnikowego (S) dla liczby skupień u = 2,..,10 dla danych z rys. 3

u	2	3	4	5	6	7	8	9	10
S	0,209	0,223	–0,068	–0,084	–0,101	–0,122	–0,122	–0.107	–0,107

Źródło: obliczenia własne z wykorzystaniem programu R.

Do wyznaczenia liczby klas dla danych z rys. 3. wykorzystany został indeks sylwetkowy. Tabela 3 przedstawia wartości tego indeksu dla liczby skupień u = 2, 3, ..,10. Wartość optymalna (maksymalna) indeksu została osiągnięta dla 3 skupień, jednak jego niska wartość świadczy o słabej stabilności otrzymanych skupień.

7. Złożoność czasowa algorytmu

Nowy algorytm, jak wszystkie algorytmy podejścia spektralnego, wymaga obliczenia wartości własnych macierzy $\mathbf{L}$. Jest to najbardziej złożona obliczeniowo część algorytmu. Metody obliczenia wartości własnych często w praktyce wykorzystują dekompozycję *svd* [Anderson i in. 1999, s. 573], czas tej czynności jest wprost proporcjonalny do rozmiaru macierzy $\mathbf{L}$ $(n \times n)$. Tabela 3 przedstawia przybliżone średnie czasy tej czynności dla n = 10, 100 , ..., 1 000 000 w środowisku **R**.

Tabela 3. Średni przybliżony czas znalezienia wartości własnych macierzy $\mathbf{L}$

n	10	100	1000	10 000	100 000	1 000 000
Średni czas (w min.)*	0,028	2,83	283	28 333	283 333 333	28 333 333 333

* Dla n > 1000 estymowane wartości przybliżone.

Źródło: obliczenia własne z wykorzystaniem programu R.

Z przedstawionej symulacji wynika, iż praktyczne stosowanie algorytmu należy ograniczyć do zbiorów rzędu nie większego niż 1000 obiektów.

8. Wnioski i problemy otwarte

W artykule przedstawiona została propozycja nowego algorytmu analizy skupień łączącego cechy podejścia spektralnego i klasyfikacji za pomocą przesunięcia okna w kierunku wektora średniej. Prezentowany algorytm cechuje się:

- Możliwością osiągania lepszych rezultatów analizy skupień niż tradycyjne algorytmy klasyfikacyjne, zwłaszcza w przypadku nietypowych kształtów skupień.
- Możliwością automatycznego określenie liczby skupień.

- Dobrą odpornością na zmienne zakłócające.

Problemem otwartym jest klasyfikacja danych z dużych (>1000 obiektów) zbiorów danych.

Literatura

Anderson E., Bai Z., Bischof C., Blackford S., Demmel J., Dongarra J., Du Croz J., Greenbaum A., Hammarling A., McKenney A., Sorensen D. , *LAPACK User's Guide*, SIAM, Philadelphia 1999.

Comaniciu D., Meer P., *Mean Shift Analysis and Applications*, IEEE Int. Conf. Computer Vision (ICCV'99), Kerkyra, Greece 1999.

Dudek A., *Klasyfikacja spektralna a tradycyjne metody analizy skupień*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 37, Wydawnictwo UE, Wrocław 2009.

Hubert L.J., Arabie P., *Comparing partitions*, "Journal of Classification" 1985, no 1.

Karatzoglou A., *Kernel methods. Software, algorithms and applications*, Dissertation, Technical University, Wien 2006.

Korzeniewski J., *Propozycja nowego algorytmu wyznaczającego liczbę skupień*, [w:] Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, Taksonomia 12, Wydawnictwo AE, Wrocław 2005.

Ng A., Jordan M., Weiss Y., *On Spectral Clustering: Analysis and an Algorithm*, [w:] *Advances in Neural Information Processing Systems 14*, T. Dietterich, S. Becker, Z. Ghahramani (red.), MIT Press, 2002.

von Luxburg U., *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149, 2006.

Walesiak M., *Analiza skupień*, [w:] *Statystyczna analiza danych z wykorzystaniem programu R*, M. Walesiak, E. Gatnar (red.), Wydawnictwo Naukowe PWN, Warszawa 2009.

Walesiak M., Dudek A., *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – charakterystyka problemu*, Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 450, Szczecin 2007.

Walesiak M., Dudek A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 84, Wydawnictwo UE, Wrocław 2009.

Wang S., Qiu W., Zamar R. H., CLUES: *A non-parametric clustering method based on local shrinking*, „Computational Statistics & Data Analysis" 2007, vol. 52, issue 1.

SMS – PROPOSAL OF NEW CLUSTERING ALGORITHM

Summary: Spectral clustering [Ng, Jordan, Weiss 2002; von Luxburg 2006] and mean shift clustering {Wang, Xiu, Damar 2007] are two relatively new approaches in cluster analysis, giving, especially for clusters of unusual shapes, better results than classical methods such as *k-means*, *k-medoids* or hierarchical agglomerative methods. The article contains a proposal for algorithm with the working name – SMS (*Spectral-Mean Shift*) that combines features of both approaches, distinguishing among other cluster analysis algorithms with:
– the possibility of cluster analysis of unusual shapes,
– the ability to automatically identify the number of clusters,
– better resistance to interference (noisy) variables.

Keywords: cluster analysis, spectral clustering, SMS.