Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 37
Ekonometria 23 2009

Andrzej Dudek
Uniwersytet Ekonomiczny we Wrocławiu – Wydział w Jeleniej Górze

KLASYFIKACJA SPEKTRALNA A TRADYCYJNE METODY ANALIZY SKUPIEŃ

Streszczenie: Klasyfikacja spektralna to rozwijające się od końca poprzedniego wieku podejście w analizie skupień. Podejście to, mimo niekiedy niezbyt rozbudowanej podbudowy teoretycznej, daje bardzo dobre wyniki empiryczne zarówno na zbiorach testowych, jak i na rzeczywistych zbiorach danych. W artykule przedstawiono algorytm analizy spektralnej w postaci ogólnej oraz wyniki symulacji obliczeniowych porównujących wyniki klasyfikacji opartej na dekompozycji spektralnej z metodą *k*-średnich, metodą *k*-medoidów, metodą Warda i metodą kompletnego połączenia na zbiorach danych o znanej strukturze wygenerowanych z wielowymiarowego rozkładu normalnego, na zbiorach danych z zakłóceniami oraz na zbiorach danych otrzymanych z przetworzenia rzeczywistych obrazów.

Słowa kluczowe: analiza skupień, klasyfikacja spektralna, analiza symulacyjna.

1. Wstęp

Metody analizy skupień dokonują grupowania obiektów na względnie jednorodne klasy. Tradycyjnie spośród najważniejszych metod analizy skupień wyróżnia się metodę *k*-średnich, metodę *k*-medoidów i metody hierarchiczne aglomeracyjne: [Gordon 1999, s. 79, za Gatnar, Walesiak 2004] pojedynczego połączenia (*single-link*), kompletnego połączenia (*complete-link*), średniej klasowej (*group average-link*), ważonej średniej klasowej (*weighted average-link*), powiększonej sumy kwadratów odległości (*incremental sum of squares*), środka ciężkości (*centroid*), medianową (*median*), giętką (*flexible*).

Te metody są dobrze opisane w literaturze (np. [Everitt, Landau, Leese 2001; Gordon 1999; Hartigan 1975]) i znajdują szerokie zastosowanie w naukach ekonomicznych (por. [Walesiak 1993; Gatnar, Walesiak 2004]). Od końca poprzedniego wieku rozwijają się jednak dość szybko metody klasyfikacji „nowej generacji", wśród których największą popularność zyskują klasyfikacja oparta na modelu (*model-based clustering*) i klasyfikacja oparta na dekompozycji spektralnej (*spectral clustering*). To ostatnie podejście, wykorzystując tradycyjne metody analizy skupień do właściwego procesu klasyfikacji, opiera się na założeniu, że zamiast doko-

nywać podziału obiektów w przestrzeni R^n dokonywany jest ich podział w przestrzeni Hilberta wyznaczonej przez wektory własne macierzy Laplace'a.

W artykule przedstawiono podejście spektralne oraz dokonano porównania tego podejścia z najbardziej popularnymi metodami analizy skupień: metodą *k*-średnich, metodą *k*-medoidów, metodą Warda i metodą kompletnego połączenia.

W pierwszej części artykułu przedstawiono podstawy matematyczne klasyfikacji przy użyciu dekompozycji spektralnej w postaci ogólnej. Część druga opisuje najbardziej znane algorytmy analizy skupień oraz etapy procedury klasyfikacyjnej. Części trzecia do piątej zawierają opisy eksperymentów porównujących efekty działania klasyfikacji opartej na dekompozycji spektralnej z tradycyjnymi metodami analizy skupień dla zbiorów danych wygenerowanych z wielomianowego rozkładu normalnego o znanej strukturze, dla zbiorów danych ze zmiennymi zakłócającymi, a także zbiorów nietypowych otrzymanych z digitalizacji prostych obrazów.

2. Dekompozycja spektralna

Niech **X** oznacza macierz danych o *n* wierszach i *m* kolumnach, *u* – liczbę klas, na które procedura klasyfikacyjna ma podzielić obiekty macierzy **X**.

Niech **D** będzie macierzą podobieństwa obiektów **X**. Wartość d_{ij} oznacza stopień „bliskości" obiektów x_i i x_j – im jest mniejsza, tym obiekty leżą bliżej.

Macierz **D** może być otrzymana na wiele sposobów. Najczęściej **D** jest obliczana zgodnie z (1) (jądro gaussowskie).

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

gdzie: **D** – macierz podobieństwa,
X – macierz danych,
m – liczba kolumn w macierzy danych (liczba zmiennych),
σ – parametr skali (stała).

Macierz **D** może być obliczana również z wykorzystaniem innych funkcji jądra:
- jądro wielomianowe,
- jądro liniowe,
- jądro hiperboliczne,
- jądro Bessela,
- jądro łańcuchowe (dla danych tekstowych).

Istnieją również sposoby konstrukcji macierzy **D** bezpośrednio z macierzy odległości euklidesowych albo poprzez usunięcie z niej wszystkich elementów większych niż wartość progowa ε (metoda ε-sąsiedztwa) lub poprzez pozostawienie

w niej dla każdego obiektu tylko odległości od określonej liczby najbliższych sąsiadów (metoda *k*-najbliższych sąsiadów).

Dla macierzy **D** konstruowana jest macierz wag **W** zgodnie z (2):

$$w_{ij} = \begin{cases} \sum_{k=1}^{n} d_{kj} & \text{gdy } x = j \\ 0 & \text{gdy } x \neq j \end{cases}, \tag{2}$$

gdzie: **W** – macierz wag,
D – macierz podobieństwa,
n – liczba wierszy w macierzy danych (liczba obiektów),

oraz macierz **L** zgodnie z (3)

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

gdzie: **W** – macierz wag,
D – macierz podobieństwa.

Macierz **L** nazywana jest macierzą Laplace'a[1]. Dla macierzy **L** obliczane są wektory własne. Pierwsze *u* z nich (*u* zgodnie z założeniami wstępnymi oznacza liczbę klas) tworzy macierz **E**, przy czym każdy wektor własny jest traktowany jako kolumna macierzy **E** (macierz **E** ma więc dokładnie *n* wierszy).

Macierz **E** jest normalizowana zgodnie z (4)

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

Otrzymana macierz **E′** to macierz wejściowa dla procedury klasyfikacyjnej. Do podziału tej macierzy na klasy można użyć dowolnej „tradycyjnej" metody klasyfikacji, przy czym najczęściej wykorzystywana jest metoda *k*-średnich, która została zaproponowana w oryginalnej wersji algorytmu [Ng, Jordan, Weiss 2001].

3. „Tradycyjna" procedura klasyfikacyjna

W literaturze przedmiotu w typowej procedurze klasyfikacyjnej wyodrębnia się zazwyczaj osiem etapów (por. [Milligan 1996, s. 342-343; Walesiak 2004]):

[1] W literaturze przedmiotu (np. [von Luxburg 2006]) macierz **L** jest nazywana *laplasjanem*, wydaje się jednak, że w języku polskim właściwszym określeniem będzie macierz Laplace'a, gdyż sformułowanie *laplasjan* jest zarezerwowane dla rachunku różniczkowego.

1) wybór obiektów do klasyfikacji,
2) wybór zmiennych charakteryzujących obiekty,
3) wybór formuły normalizacji wartości zmiennych,
4) wybór miary odległości,
5) wybór metody klasyfikacji,
6) ustalenie liczby klas,
7) walidację wyników klasyfikacji,
8) opis (interpretację) i profilowanie klas).

Spośród metod normalizacji wyróżnić można:

- standaryzację klasyczną,
- unitaryzację,
- unitaryzację zerowaną,
- normalizację w przedziale [–1; 1],
- przekształcenia ilorazowe.

Spośród najważniejszych takich metod klasyfikacji wymienić należy:

- metodę *k*-średnich,
- metody hierarchiczne, aglomeracyjne (por. [Gordon 1999, s. 79]):
  - pojedynczego połączenia,
  - kompletnego połączenia,
  - średniej klasowej,
  - ważonej średniej klasowej (Mcquitty),
  - powiększonej suma kwadratów odległości (Ward),
  - środka ciężkości,
  - medianową,
  - giętką;
- metody optymalizacyjne:
  - metodę *k*-medoidów [Kaufman, Rousseeuw 1990].

Do ustalenia liczby klas w etapie 6 wykorzystuje się mierniki jakości klasyfikacji. W literaturze przedmiotu występuje ich ok. 100, jednak w praktyce najczęściej używa się trzech najlepszych mierników globalnych z eksperymentu Milligana i Coopera [1985]:

- indeks Calińskiego i Harabasza [1974],
- indeks Bakera i Huberta [Hubert 1974; Baker, Hubert 1975],
- indeks Huberta i Levine [1976]

oraz indeksy nieuwzględnione w eksperymencie Miligana i Coopera, a dość powszechnie spotykane w literaturze:

- indeks sylwetkowy (*silhouette*) [Kaufman, Rousseeuw 1990],
- indeks Krzanowskiego i Lai [1985],
- indeks GAP (Tibshirani, Walther, Hastie [2001]),
- indeks Daviesa i Bouldina [1979].

4. Porównanie klasyfikacji opartej na dekompozycji spektralnej z tradycyjnymi metodami analizy skupień dla danych o znanej strukturze bez zmiennych zakłócających

W pierwszym eksperymencie porównano wyniki klasyfikacji opartej na dekompozycji spektralnej z wynikami otrzymanymi przy wykorzystaniu metody *k*-średnich, metody *k*-medoidów, klasyfikacji hierarchicznej Warda i klasyfikacji hierarchicznej metodą kompletnego połączenia dla danych wygenerowanych z wielowymiarowego rozkładu normalnego o znanej strukturze klas. Charakterystykę wykorzystanych zbiorów danych przedstawia tab. 1. Zbiory zostały wygenerowane funkcją cluster.Gen pakietu cluster.Sim autorstwa Walesiaka i Dudka [2008] w środowisku **R**.

Tabela 1. Modele wykorzystane w symulacji porównawczej

	Liczba obiektów	Liczba zmiennych	Liczba klas	Środki klas	Macierz wariancji/kowariancji
I	250	2	5	(0, 0), (0, 10), (5, 5), (10, 0), (10, 10)	$\sigma_{jj} = 1\ (1 \le j \le 2)$, $\sigma_{jl} = 0{,}9\ (1 \le j \ne l \le 2)$
II	250	3	5	(0, 0, 0), (10, 10, 10), (–10, –10, –10), (10, –10, 10), (–10, 10, 10)	$\sigma_{jj} = 3\ (1 \le j \le 3)$, $\sigma_{jl} = 2\ (1 \le j \ne l \le 3)$
III	240	2	4	(–4, 5), (5, 14), (14, 5), (5, –4)	$\sigma_{jj} = 1\ (1 \le j \le 2)$, $\sigma_{jl} = 0\ (1 \le j \ne l \le 2)$
IV	160	3	4	(–4, 5, –4), (5, 14, 5), (14, 5, 14), (5, –4, 5)	$\sigma_{jj} = 1\ (1 \le j \le 3$, $\sigma_{jl} = 0\ (1 \le j \ne l \le 3)$
V	180	2	3	(0, 0), (1,5, 7), (3, 14)	Różne macierze wariancji/kowariancji dla różnych skupień

Źródło: opracowanie własne na podstawie [Walesiak, Dudek 2008].

Dla każdego modelu wygenerowano 50 zbiorów danych, przeprowadzono procedurę klasyfikacyjną i porównano otrzymane rezultaty klasyfikacji z rzeczywistą strukturą klas za pomocą skorygowanego indeksu Randa [Hubert, Arabie 1985]. Średnie wartości tego miernika dla każdego modelu i dla każdej metody klasyfikacji przedstawia tab. 2.

Tabela 2. Porównanie wyników klasyfikacji dla zbiorów bez zmiennych zakłócających

Model	Metoda klasyfikacji				
	k-średnich	*k*-medoidów	Warda	kompletnego połączenia	podejście spektralne
I	0,89732	0,998485	0,996977	0,996474	0,953093
II	0,862309	0,996754	0,996725	0,996012	0,910186
III	0,880942	1	1	1	0,937404
IV	0,830127	1	1	1	0,965459
V	0,984844	0,994477	0,997502	0,99211	0,936914

Źródło: opracowanie własne z wykorzystaniem pakietów Kernlab i ClusterSim w środowisku **R**.

Warto zauważyć, że zdecydowanie najlepiej w przypadku takich „czystych" zbiorów zachowują się metody hierarchiczne i metoda klasyfikacji wokół medoidów. Podejście spektralne daje nieznacznie gorsze rezultaty. Podejście to nie jest więc alternatywą dla metod dotychczas znanych. Należy jednak pamiętać, że zbiory tego typu bardzo rzadko występują w rzeczywistych problemach klasyfikacyjnych. W następnych krokach należy więc porównać działanie algorytmów dla danych zawierających szumy oraz dla zbiorów nietypowych nieotrzymanych z rozkładu normalnego.

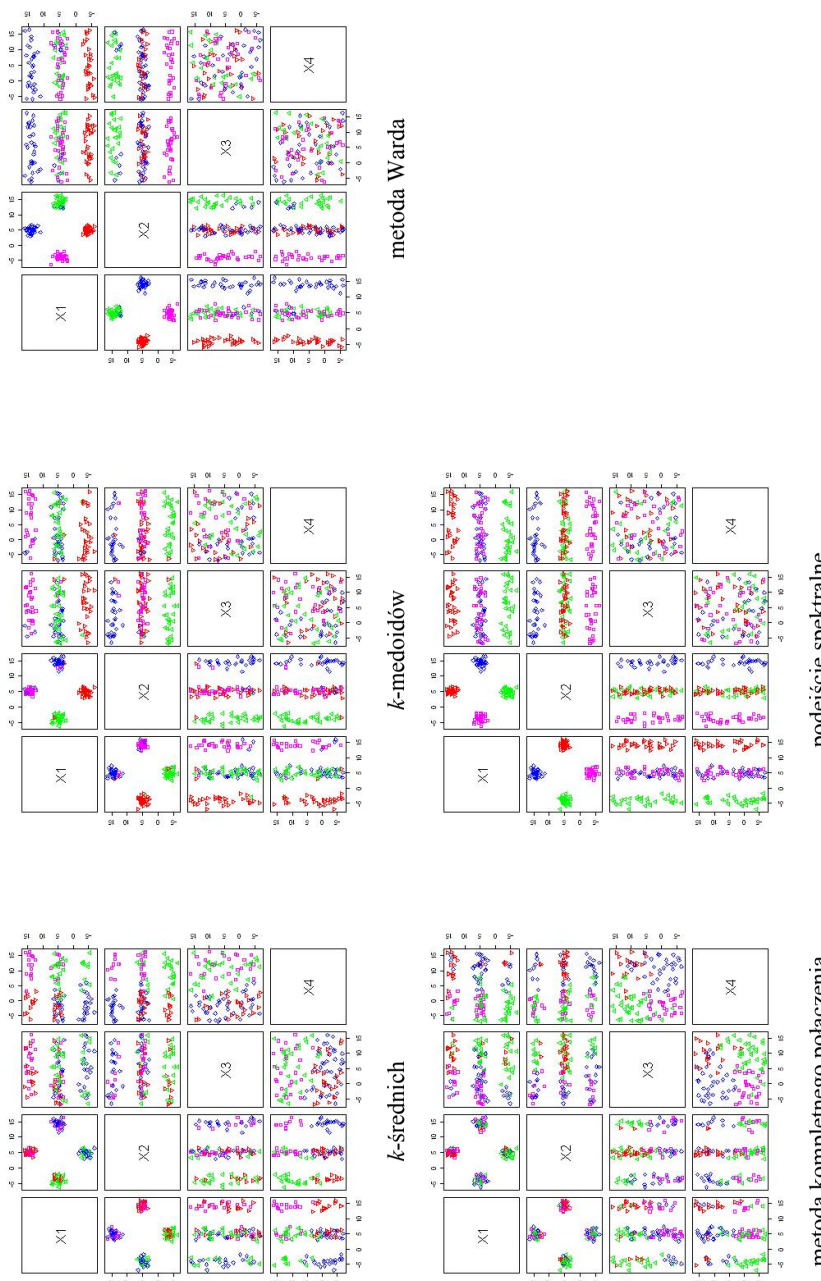
5. Porównanie klasyfikacji opartej na dekompozycji spektralnej z tradycyjnymi metodami analizy skupień dla danych o znanej strukturze ze zmiennymi zakłócającymi

W eksperymencie drugim porównano wyniki klasyfikacji opartej na dekompozycji spektralnej z wynikami otrzymanymi przy wykorzystaniu metody *k*-średnich, metody *k*-medoidów, klasyfikacji hierarchicznej Warda i klasyfikacji hierarchicznej metodą kompletnego połączenia dla danych wygenerowanych z wielowymiarowego rozkładu normalnego o znanej strukturze klas z 1, 2, 3 lub 4 zmiennymi zakłócającymi. Rezultaty przykładowej symulacji w jednym etapie symulacji dla modelu 3 przedstawia rys 1.

Do symulacji użyto modeli opisanych w tab. 1, poszerzonych o zmienne zakłócające. Dla każdego modelu wygenerowano po 50 zbiorów danych z 1, 2, 3 lub 4 zmiennymi zakłócającymi, przeprowadzono procedurę klasyfikacyjną i porównano otrzymane rezultaty klasyfikacji z rzeczywistą strukturą klas za pomocą skorygowanego indeksu Randa [Hubert, Arabie 1985]. Średnie wartości tego miernika dla każdego modelu i dla każdej metody klasyfikacji przedstawia tab. 3.

Wprowadzenie do zbiorów danych zmiennych zakłócających powoduje, że tradycyjne metody klasyfikacji zachowują się zdecydowanie gorzej, przy czym już dla jednej zmiennej zakłócającej wartość średnia skorygowanego indeksu RAND maleje zdecydowanie wolniej niż dla pozostałych metod, a przy czterech zmiennych zakłócających podejście spektralne daje średnio najlepsze wyniki w 4 na 5 modelach (przy czym dla piątego modelu metoda ta „przegrywa" tylko z metodą Warda).

Należy jednak zauważyć, iż w procedurze klasyfikacyjnej pominięto etap identyfikacji zmiennych zakłócających i ich usunięcia. Kolejnym etapem wskazującym użyteczność metod klasyfikacji spektralnej powinno być więc porównanie metod klasyfikacji dla zbiorów pochodzących z digitalizacji obrazów zawierających rzeczywiste struktury danych, niekoniecznie mających rozkład normalny.



Rys. 1. Wyniki procedury klasyfikacyjnej dla poszczególnych metod dla modelu III dla 2 zmiennych zakłócających

Źródło: opracowanie własne z wykorzystaniem pakietów Kernlab i ClusterSim w środowisku **R**.

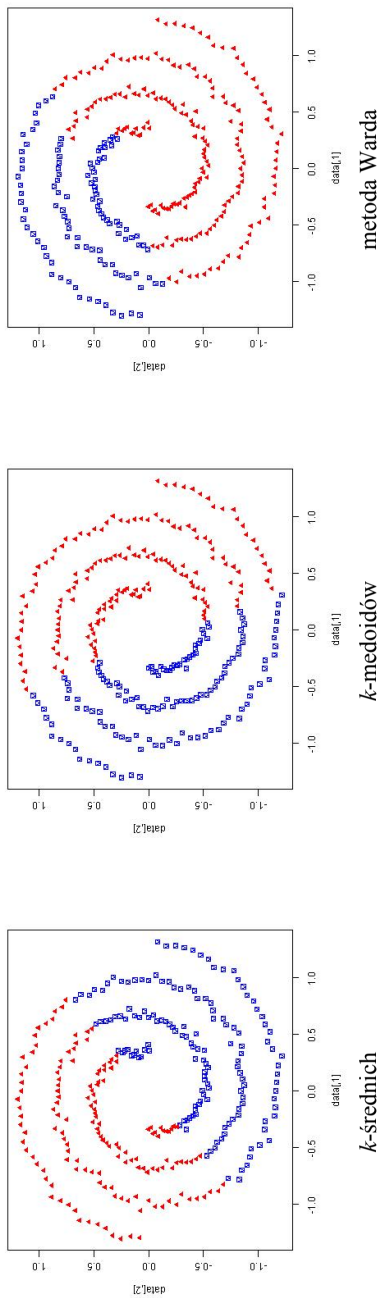
Tabela 3. Porównanie wyników klasyfikacji dla zbiorów ze zmiennymi zakłócającymi

Model	Zakłócenia	Metoda klasyfikacji				
		k-średnich	*k*-medoidów	Warda	kompletnego połączenia	podejście spektralne
I	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
II	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
III	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
IV	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
V	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

Źródło: opracowanie własne z wykorzystaniem pakietów Kernlab i ClusterSim w środowisku **R**.

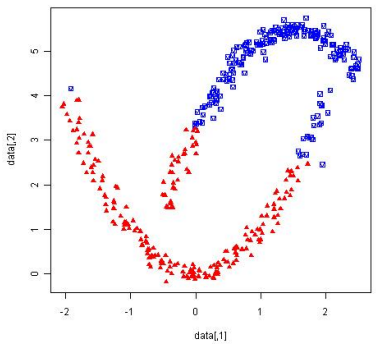
6. Porównanie klasyfikacji opartej na dekompozycji spektralnej z tradycyjnymi metodami analizy skupień dla zbiorów „nietypowych”

W eksperymencie trzecim porównano wyniki klasyfikacji opartej na dekompozycji spektralnej z wynikami innych klasyfikacji dla zbiorów danych otrzymanych z digitalizacji obrazów z czytelną strukturą klas o kształcie innym niż koła czy wielokąty wypukłe. Zbiory te zostały roboczo nazwane kolejnymi literami alfabetu. Do ich utworzenia zostały wykorzystane funkcje pakietu mlbench i procedury własne w języku **R**. Dla każdego typu zbioru wygenerowano 10 jego realizacji. Rysunki 2-5 pokazują wyniki klasyfikacji dla poszczególnych zbiorów, a tab. 4 przedstawia średnie wartości skorygowanego indeksu RAND dla poszczególnych metod klasyfikacji. Można zauważyć, że dla tego typu zbiorów klasyfikacja spektralna zachowuje się zdecydowanie najlepiej spośród wszystkich analizowanych metod, za każdym razem trafnie znajdując prawidłową strukturę klas, co spośród pozostałych metod udaje się tylko raz metodzie Warda w przypadku zbioru **D**. Można więc przypuszczać, że podejście spektralne dużo lepiej nadaje się do podziału na klasy

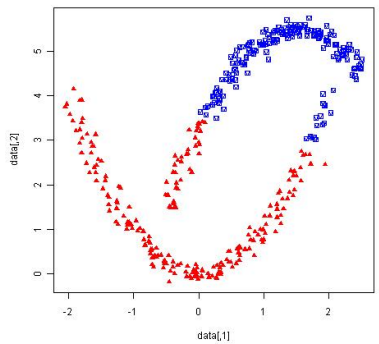


Rys. 2. Wyniki procedury klasyfikacyjnej dla poszczególnych metod dla zbioru „A"

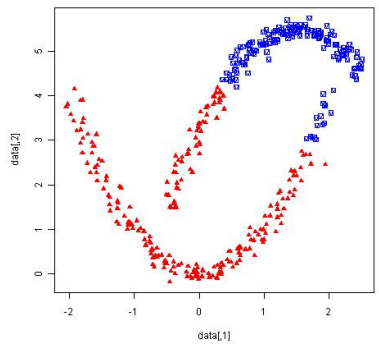
Źródło: opracowanie własne z wykorzystaniem pakietów Kernlab i mlbench w środowisku **R**.



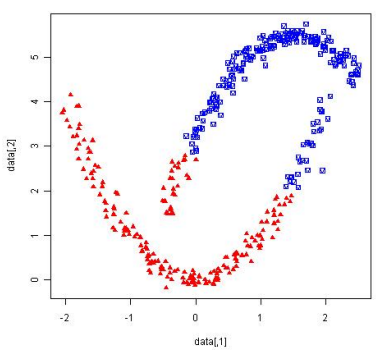
k-średnich



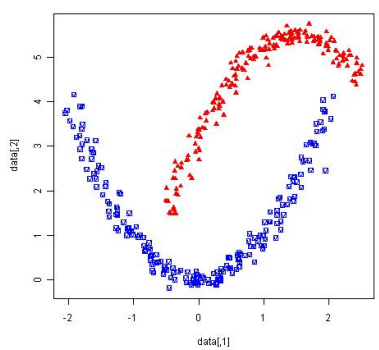
k-medoidów



metoda Warda



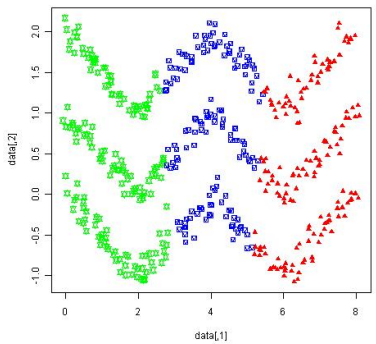
metoda kompletnego połączenia



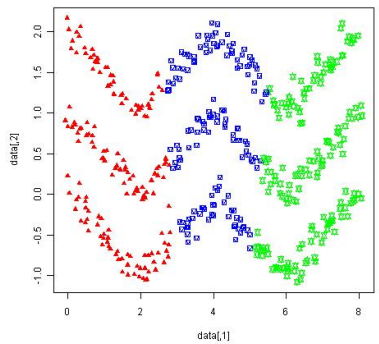
podejście spektralne

Rys. 3. Wyniki procedury klasyfikacyjnej dla poszczególnych metod dla zbioru „B”

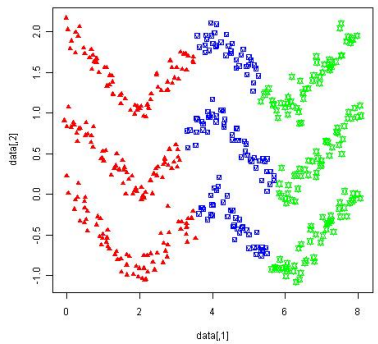
Źródło: opracowanie własne z wykorzystaniem pakietu Kernlab i procedur własnych w środowisku **R**.



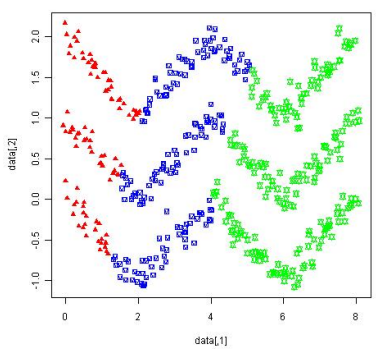
k-średnich



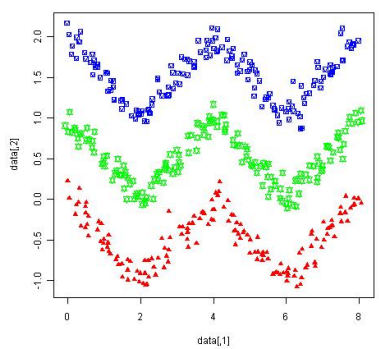
k-medoidów



metoda Warda



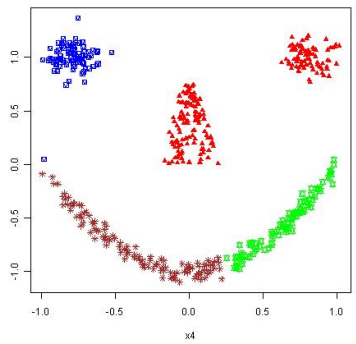
metoda kompletnego połączenia



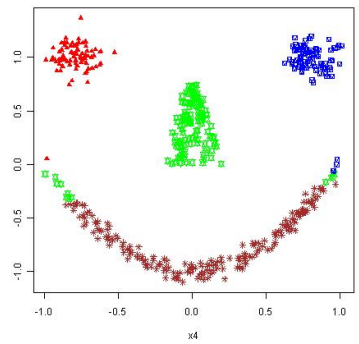
podejście spektralne

Rys. 4. Wyniki procedury klasyfikacyjnej dla poszczególnych metod dla zbioru „C"

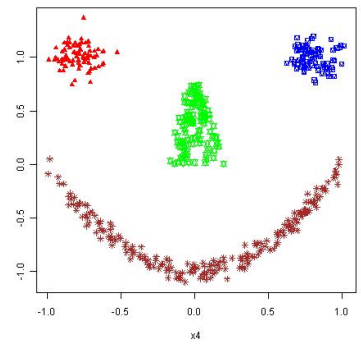
Źródło: opracowanie własne z wykorzystaniem pakietu Kernlab i procedur własnych w środowisku **R**.



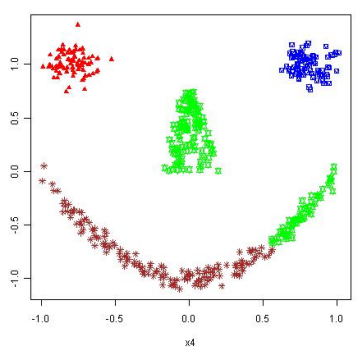
k-średnich



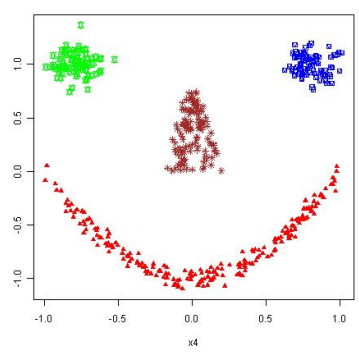
k-medoidów



metoda Warda



metoda kompletnego połączenia



podejście spektralne

Rys. 5. Wyniki procedury klasyfikacyjnej dla poszczególnych metod dla zbioru „D”

Źródło: opracowanie własne z wykorzystaniem pakietów Kernlab i mlbench w środowisku **R**.

Tabela 4. Porównanie wyników klasyfikacji dla poszczególnych metod dla zbiorów nietypowych

Model	Metoda klasyfikacji				
	k-średnich	*k*-medoidów	Warda	kompletnego połączenia	podejście spektralne
A	0,02679	0,004206	0,068263	0,036945	1
B	0,517194	0,524449	0,395452	0,51719	1
C	–0,00204	–0,0012	–0,001	0,003202	1
D	0,585176	0,896948	1	0,676797	1

Źródło: opracowanie własne z wykorzystaniem pakietów Kernlab, ClusterSim i mlbench w środowisku **R**.

tego typu zbiorów (które dużo częściej występują w rzeczywistych problemach klasyfikacyjnych niż zbiory wygenerowane z rozkładu normalnego).

7. Podsumowanie

Przedstawione wyniki symulacji obliczeniowych są bardzo zachęcające. Pokazują one, że zarówno dla danych testowych wygenerowanych sztucznie, jak i dla danych otrzymanych przez przetwarzanie obrazów podejście spektralne daje nie gorsze rezultaty niż tradycyjne metody klasyfikacji, a często zachowuje się zdecydowanie lepiej.

Mimo że niektóre etapy procedury nie mają jeszcze odpowiednio rozbudowanej podbudowy formalnej, podejście to daje bardzo dobre wyniki empiryczne i można z nim wiązać spore nadzieje odnośnie do rozwoju analizy skupień.

Należy się spodziewać, iż prace nad analizą spektralną będą przebiegać dwutorowo: uzupełnienie „luk" teorii i opracowywanie miar, wskaźników i metod (por. [Climescu-Haulica 2006]) oraz wskazywanie zastosowań tej metody w zagadnieniach klasyfikacyjnych, które do tej pory były omijane w badaniach ze względu na słabe wyniki otrzymywane przy użyciu tradycyjnych metod.

Literatura

Baker F.B., Hubert L.J., *Measuring the power of hierarchical cluster analysis*, „Journal of the American Statistical Association" 1975, no. 70, s. 31-38.

Caliński R.B., Harabasz J., *A dendrite method for cluster analysis*, „Communications in Statistics" 1974, vol. 3, s. 1-27.

Climescu-Haulica A., *How to choose the number of clusters. The Cramer Multiplicity Solution*, [w:] H.H.-J. Lenz, R. Decker (red.), *Advances in Data Analysis*, Berlin 2006, s. 15-23.

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

Davies D.L., Bouldin D.W., *A cluster separation measure*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" 1979, vol. 1, no. 2, s. 224-227.

Everitt B.S., Landau S., Leese M., *Cluster Analysis*, Edward Arnold, London 2001.

Gatnar E., Walesiak M. (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław 2004.
Gordon A.D., *Classification*, Chapman and Hall/CRC, London 1999.
Hartigan J.A., *Clustering Algorithms*, Wiley, New York, London, Sydney, Toronto 1975.
Hubert L.J., *Approximate evaluation technique for the single-link and complete-link hierarchical clustering procedures*, „Journal of the American Statistical Association" 1974, vol. 69, no. 347, s. 698-704.
Hubert L.J., Arabie P., *Comparing partitions*, „Journal of Classification" 1985, no. 1, s. 193-218.
Hubert L.J., Levine J.R., *Evaluating object set partitions: free sort analysis and some generalizations*, „Journal of Verbal Learning and Verbal Behaviour" 1976, vol. 15, s. 549-570.
Kaufman L., Rousseeuw P.J., *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley, New York 1990.
Krzanowski W.J., Lai Y.T., *A criterion of determining the number of groups in a data set using sum of squares clustering*, „Biometrics" 1985, vol. 44, s. 23-34.
von Luxburg U., *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149, 2006.
Milligan G.W., *Clustering Validation: Results and Implications for Applied Analyses*, [w:] P. Arabie, L.J. Hubert, G. de Soete (red.), *Clustering and Classification*, World Scientific, Singapore 1996, s. 341-375.
Milligan G.W., Cooper M.C., An examination of procedures for determining the number of clusters in a data set, „Psychometrika" 1985, no. 2, s. 159-179.
Ng A., Jordan I., Weiss Y., *On Spectral Clustering: Analysis and an Algorithm*, Neural Information Processing Symposium, (http://www.nips.cc/NIPS2001/papers/psgz/AA35.ps.gz), 2001.
Rousseeuw P.J., *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*, „Journal of Computational and Applied Mathematics" 1987, no. 20, s. 53-65.
Tibshirani R., Walther G., Hastie T., *Estimating the number of clusters in a data set via the gap statistic*, „Journal of the Royal Statistical Society" 2001, ser. B, vol. 63, part 2, s. 411-423.
Tibshirani R., Walther G., *Cluster Validation by Prediction Strength*, „Journal of Computational and Graphical Statistics" 2005, September, no. 3.
Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, [w:] Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654, Seria: Monografie i Opracowania nr 101, AE, Wrocław 1993.
Walesiak M., *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, [w:] Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1010, AE, Wrocław 2004, s. 52-71.
Walesiak M., Dudek A., *ClusterSim,* pakiet środowiska statystycznego **R**, http://wgrit.ae.jgora.pl/keii/clusterSim oraz http://cran.rproject.org/web/packages/clusterSim/index.html, 2008.

SPECTRAL CLUSTERING VS TRADITIONAL CLUSTERING METHODS

Summary: Spectral clustering has been known since the end of the 20th century and is developing quite fast. Despite the lack of a strong theoretical basis, this method gives very good empirical results on artificial and real data. In this paper, algorithm (in general form) of spectral clustering has been described along with the results of empirical simulations comparing spectral clustering with k-means, partition around medoids, Ward and complete link „traditional" methods. The simulations have been made on datasets with known cluster structure generated from multivariate normal distribution, on datasets with noisy variables and on processed real images data.