PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012

MAREK WALESIAK

KLASYFIKACJA SPEKTRALNA A SKALE POMIARU ZMIENNYCH[1]

1. WPROWADZENIE

Analiza skupień bazująca na dekompozycji spektralnej (*spectral clustering*) rozwija się w literaturze poświęconej wielowymiarowej analizie danych od końca XX w. Nazwa metody „klasyfikacja spektralna" wywodzi się stąd, że w jednym z jej podstawowych kroków wyznacza się spektrum (widmo) macierzy Laplace'a. W matematyce zbiór wartości własnych macierzy nazywa się spektrum (widmem) macierzy (zob. np. [7], s. 182). Podstawowy algorytm klasyfikacji spektralnej dla danych metrycznych zaproponowano w pracy [8]. Inne algorytmy klasyfikacji spektralnej scharakteryzowano m.in. w pracach [10] i [14].

W artykule scharakteryzowano metodę klasyfikacji spektralnej z punktu widzenia skal pomiaru zmiennych. Rozpatrzono jej zastosowanie w klasyfikacji danych nominalnych, porządkowych, przedziałowych oraz ilorazowych. W tym celu w procedurze tej metody przy wyznaczaniu macierzy podobieństwa (*affinity matrix*) zastosowano funkcję (1) z miarami odległości właściwymi dla danych mierzonych na różnych skalach pomiaru. Dzięki takiemu podejściu dla danych niemetrycznych (nominalnych i porządkowych) możliwe jest pośrednie wzmocnienie skali pomiaru zmiennych.

Zaproponowana metoda klasyfikacji spektralnej może być z powodzeniem stosowana we wszystkich zagadnieniach klasyfikacyjnych, w tym dotyczących pomiaru, analizy i wizualizacji preferencji.

2. TYPY SKAL POMIAROWYCH I ICH CHARAKTERYSTYKA

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru, wprowadzone przez Stevensa w pracy [13]. Skale pomiaru są uporządkowane od najsłabszej do najmocniejszej: nominalna, porządkowa, przedziałowa, ilorazowa. Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową do niemetrycznych.

Podstawowe własności skal pomiaru przedstawia tab. 1.

[1] Praca naukowa finansowana ze środków na naukę w latach 2009-2012 jako projekt badawczy nr N N111 446037 nt. „Pomiar, analiza i wizualizacja preferencji ujawnionych i wyrażonych z wykorzystaniem metod wielowymiarowej analizy statystycznej i programu R".

Tabela 1.

Podstawowe własności skal pomiaru

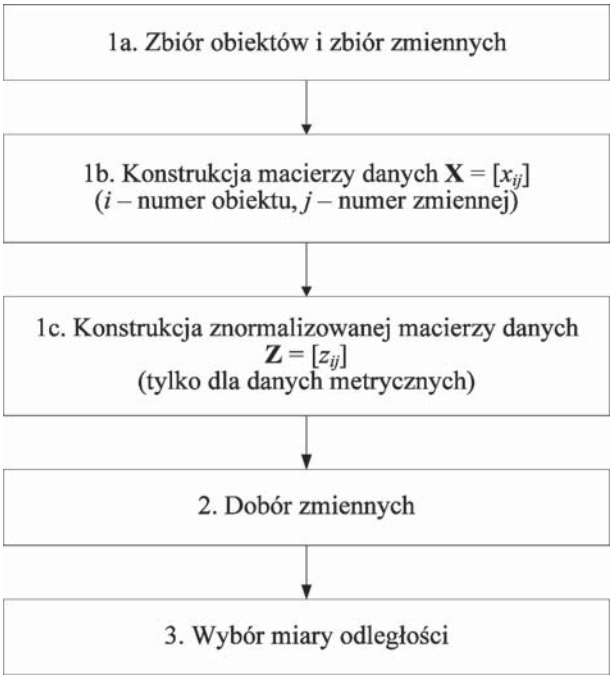
Typ skali	Dozwolone przekształcenia matematyczne	Dopuszczalne relacje	Dopuszczalne operacje arytmetyczne
Nominalna	$z = f(x), f(x)-$ dowolne przekształcenie wzajemnie jednoznaczne	równości ($x_A = x_B$), różności ($x_A \neq x_B$)	zliczanie zdarzeń (liczba relacji równości, różności)
Porządkowa	$z = f(x), f(x)-$ dowolna ściśle monotonicznie rosnąca funkcja	powyższe oraz większości ($x_A > x_B$) i mniejszości ($x_A < x_B$)	zliczanie zdarzeń (liczba relacji równości, różności, większości, mniejszości)
Przedziałowa	$z = bx + a \quad (b > 0), z \in R$ dla wszystkich x zawartych w R, wartość zerowa na tej skali jest zwykle przyjmowana arbitralnie lub na podstawie konwencji*	powyższe oraz równości różnic i przedziałów ($x_A - x_B = x_C - x_D$)	powyższe oraz dodawanie i odejmowanie
Ilorazowa	$z = bx \quad (b > 0), z \in R_+$ dla wszystkich x zawartych w R_+, naturalnym początkiem skali ilorazowej jest wartość zerowa (zero lewostronnie ogranicza zakres skali)	powyższe oraz równości ilorazów ($\frac{x_A}{x_B} = \frac{x_C}{x_D}$)	powyższe oraz mnożenie i dzielenie

* Por. [1], s. 240.
Źródło: [18], s. 15.

Jedna z podstawowych reguł teorii pomiaru mówi, że jedynie rezultaty pomiaru w skali mocniejszej mogą być transformowane na liczby należące do skali słabszej (por. np. [11], s. 17; [12], s. 19; [16], s. 40; [23]; [24]). Bezpośrednia transformacja skal pomiaru zmiennych polegająca na ich wzmacnianiu nie jest możliwa, ponieważ z mniejszej ilości informacji nie można uzyskać większej jej ilości. W klasyfikacji spektralnej możliwe jest pośrednie wzmocnienie skali pomiaru zmiennych. Pierwotna macierz danych, w której zmienne mierzone są na skali nominalnej lub porządkowej zostaje przekształcona w macierz danych, w której zmienne mierzone są na skali przedziałowej.

3. KLASYFIKACJA SPEKTRALNA DLA RÓŻNYCH SKAL POMIARU ZMIENNYCH

Rys. 1 przedstawia trzy pierwsze etapy klasyfikacji spektralnej (występujące także w klasycznej analizie skupień), obejmujące ustalenie zbioru obiektów i zmiennych (po zgromadzeniu danych konstruuje się macierz danych, a w przypadku danych metrycznych w następnym kroku znormalizowaną macierz danych), dobór zmiennych oraz wybór miary odległości. Szczegółową charakterystykę tych etapów zaprezentowano m.in. w pracach [17] i [19].



Rysunek 1. Trzy pierwsze etapy w klasyfikacji spektralnej oraz w klasycznej analizie skupień

Źródło: Opracowanie własne.

Dalsza procedura klasyfikacji spektralnej obejmuje następujące kroki[2]:

4. Obliczenie symetrycznej macierzy podobieństw $\mathbf{A} = [A_{ik}]_{nxn}$ (*affinity matrix*) między obiektami, dla której $A_{ii} = 0$ oraz

$$A_{ik} = \exp(-\sigma \cdot d_{ik}) \text{ dla } \quad i \neq k, \tag{1}$$

gdzie: σ – parametr skali,

d_{ik} – miary odległości dla różnych skal pomiaru ujęte w tab. 2,

$i, k = 1, \ldots, n$ – numery obiektów.

W kroku tym można zastosować do obliczenia elementów macierzy podobieństw A_{ik} $(i \neq k)$ estymatory jądrowe (zob. [6], s. 13-14; [9]): jądro gaussowskie, jądro wielomianowe, jądro liniowe, jądro w postaci tangensa hiperbolicznego, jądro Bessela, jądro Laplace'a, jądro ANOVA, jądro łańcuchowe (dla danych tekstowych).

W oryginalnym algorytmie klasyfikacji spektralnej dla danych metrycznych w pracy [8] zastosowano jądro gaussowskie:

$$A_{ik} = \exp(-\frac{d_{ik}^2}{2\sigma^2}) \text{ dla } \quad i \neq k, \tag{2}$$

[2] Jest to algorytm zaproponowany w pracy [8] (por. [20], [21]). W artykule dokonano jego modyfikacji w kroku 4 przy obliczaniu macierzy podobieństw (*affinity matrix*).

Tabela 2.

Miary odległości dla różnych skal pomiaru*

Nazwa miary odległości	Skala pomiaru zmiennych	Funkcja (pakiet) programu R
Minkowskiego: miejska (Manhattan); euklidesowa; Czebyszewa (maximum)	metryczne	`dist (stats)`
Canberra	ilorazowe	`dist (stats)`
Braya-Curtisa	ilorazowe	`dist.BC (clusterSim)`
GDM1	metryczne	`dist.GDM (clusterSim)`
GDM2	porządkowe	`dist.GDM (clusterSim)`
Sokala i Michenera	nominalne	`dist.SM (clusterSim)`

* Formuły odległości zawiera m.in. praca [18].
Źródło: opracowanie własne.

gdzie: $d_{ik} = \sqrt{\sum_{j=1}^{m} (z_{ij} - z_{kj})^2}$,

$z_{ij}, (z_{kj})$– znormalizowana wartość j-tej zmiennej dla i-tego (k-tego) obiektu.

5. 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: $\mathbf{I} - \mathbf{L}$. W algorytmie dla uproszczenia analizy pomija się macierz jednostkową $\mathbf{I}$ (zob. [8]). Własności tej macierzy przedstawiono m.in. w pracy [15], s. 5-6.

6. 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 $\mathbf{E} = \left[e_{ij}\right]$ o wymiarach $n \times u$.

Podobnie jak w przypadku klasycznym analizy skupień zachodzi potrzeba ustalenia optymalnej liczby klas. Odpowiedni algorytm zaproponował Girolami w pracy [4].

Macierz podobieństw (*affinity matrix*) $\mathbf{A} = [A_{ik}]$ (dla $\sigma = 1$) poddawana jest dekompozycji $\mathbf{A} = \mathbf{U}\ \mathbf{\Lambda}\ \mathbf{U}^T$, gdzie $\mathbf{U}$ jest macierzą wektorów własnych macierzy $\mathbf{A}$ składającą się z wektorów $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$, a $\mathbf{\Lambda}$ jest macierzą diagonalną zawierającą wartości własne $\lambda_1, \lambda_2, \ldots, \lambda_n$.

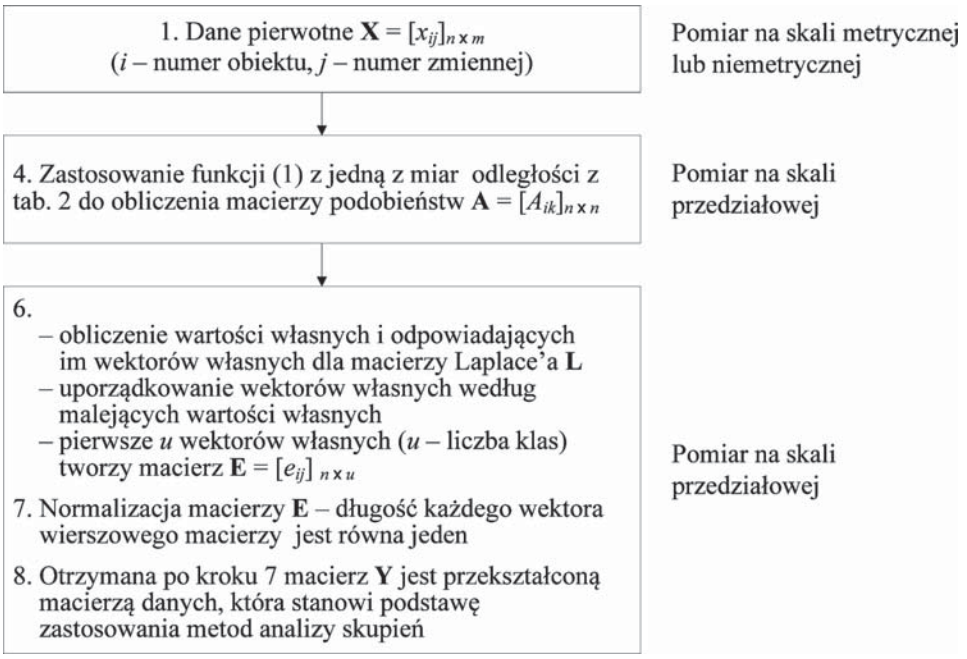
Obliczany jest wektor $\mathbf{K} = (k_1, k_2, \ldots, k_n)$, gdzie $k_i = \lambda_i \left\{\mathbf{1}_n^T \mathbf{u}_i\right\}^2$ ($\mathbf{1}_n^T$ – wektor o wymiarach $1 \times n$ zawierający wartości $1/n$). Wektor $\mathbf{K}$ jest porządkowany malejąco, a liczba jego dominujących elementów (wyznaczona np. poprzez kryterium osypiska) wyznacza optymalną liczbę skupień u, na którą algorytm klasyfikacji spektralnej powinien podzielić zbiór badanych obiektów.

7. Przeprowadza się normalizację macierzy $\mathbf{E}$ 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 $\mathbf{Y} = \left[y_{ij}\right]$ jest równa jeden.

8. Macierz **Y** stanowi punkt wyjścia zastosowania klasycznych metod analizy skupień (proponuje się tutaj wykorzystanie metody *k*-średnich).

Rys. 2 pokazuje wybrane kroki postępowania w klasyfikacji spektralnej i odpowiadające im skale pomiaru.



Rysunek 2. Wybrane kroki postępowania w klasyfikacji spektralnej i odpowiadające im skale pomiaru

Źródło: Opracowanie własne.

Jeśli dane pierwotne $\mathbf{X} = [x_{ij}]$ mierzone są na skali niemetrycznej (porządkowej, nominalnej) w wyniku zastosowania funkcji (1) z jedną z odległości właściwych dla tych skal pomiaru (zob. tab. 2) podobieństwa w macierzy $\mathbf{A} = [A_{ik}]$ mierzone są na skali przedziałowej. Ostatecznie w kroku 7 otrzymuje się metryczną macierz danych **Y** o wymiarach $n \times u$. Pozwala ona na zastosowanie dowolnych metod analizy skupień (w tym metod bazujących bezpośrednio na macierzy danych, np. metodę *k*-średnich).

4. PARAMETR σ W KLASYFIKACJI SPEKTRALNEJ

Parametr σ ma fundamentalne znaczenie w klasyfikacji spektralnej. W literaturze zaproponowano wiele heurystycznych sposobów wyznaczania wartości tego parametru (zob. np. [3]; [9]; [25]). W metodach heurystycznych wyznacza się wartość σ na podstawie pewnych statystyk opisowych macierzy odległości $[d_{ik}]$. Lepszy sposób wyznaczania parametru σ zaproponował Karatzoglou w pracy [6]. Poszukuje się takiej

wartości parametru σ, która minimalizuje zmienność wewnątrzklasową przy zadanej liczbie klas u. Jest to heurystyczna metoda poszukiwania minimum lokalnego. Zbliżony koncepcyjnie algorytm znajdowania optymalnego parametru σ zaproponowano w pracy [20]:

Krok 0. Wybierana jest próba bootstrapowa $\mathbf{X}'$ składającą się z n' obiektów opisanych wszystkimi m zmiennymi (wartość n' jest najczęściej tak dobierana, aby $\frac{1}{2}n \leq n' \leq \frac{3}{4}n$). Początkowy przedział przeszukiwania optymalnej wartości parametru σ ustalany jest jako $S_0 = [0;\ D]$ (gdzie D oznacza sumę odległości w dolnym trójkącie macierzy odległości a dla kwadratu odległości euklidesowej – pierwiastek z sumy odległości w dolnym trójkącie macierzy odległości).

Krok 1. Przedział S_k (gdzie k oznacza numer iteracji; na początku $S_k = S_0$) dzielony jest na przedziały jednakowej długości: $p_r^k = [\underline{p_r^k};\ \overline{p_r^k}]$, $r = 1, \ldots, R$ (R – liczba przedziałów w każdej iteracji; domyślnie $R = 10$).

Krok 2. Dla każdego przedziału p_r^k obliczamy jego środek: $\sigma_r^k = \frac{\underline{p_r^k} + \overline{p_r^k}}{2}$. Dla wszystkich wartości σ_r^k przeprowadzana jest klasyfikacja spektralna zbioru $\mathbf{X}'$ na ustaloną liczbę klas u.

Krok 3. Wybierane jest takie σ_r^k, dla którego zmienność wewnątrzklasowa jest minimalna.

Krok 4. Z przedziałem zawierającym wybraną wartość σ_r^k w kroku 3 przechodzi się do kroku 1 i kontynuuje procedurę do osiągnięcia zadanej liczby iteracji (domyślnie: 3).

5. OPROGRAMOWANIE W ŚRODOWISKU R

Klasyfikację spektralną zgodną z algorytmem zmodyfikowanym w artykule przeprowadza się z wykorzystaniem funkcji `speccl` pakietu `clusterSim` (zob. [22]):

```
speccl(data,nc,distance="GDM1",sigma="automatic",
sigma.interval="default",mod.sample=0.75,R=10,iterations=3)
```

Argumenty:

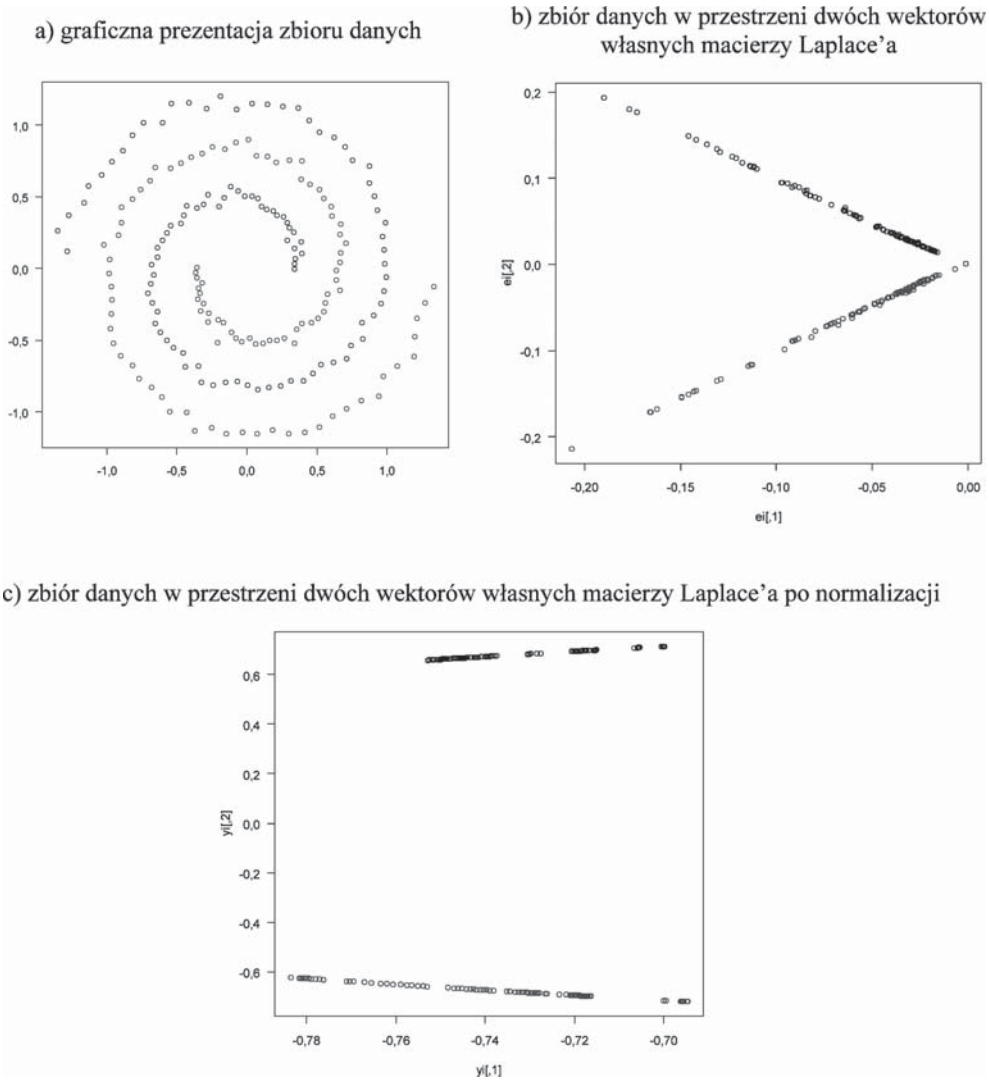
`data`	macierz danych
`nc`	liczba klas
`distance`	miary odległości z tabeli 2 (`"sEuclidean"` – kwadrat odległości euklidesowej, `"euclidean"` – odległość euklidesowa, `"manhattan"` – odległość miejska, `"maximum"` – odległość Czebyszewa, `"canberra"` – odległość Canberra, `"BC"` – odległość Braya-Curtisa, `"GDM1"` – odległość GDM dla danych metrycznych, `"GDM2"` – odległość GDM dla danych porządkowych, `"SM"` – odległość Sokala-Michenera dla danych nominalnych)

`sigma`	parametr skali: `sigma="automatic"` – parametr ustalany automatycznie zgodnie z algorytmem z punktu 4 `sigma=200` – parametr podany przez użytkownika, np. 200
`sigma.interval`	przedział przeszukiwania parametru sigma: `sigma.interval="default"` – przedział wartości od zera do sumy odległości w dolnym trójkącie macierzy odległości (dla kwadratu odległości euklidesowej – do pierwiastka z sumy odległości w dolnym trójkącie macierzy odległości) `sigma.interval=1000` – przedział wartości od zera do wartości podanej przez użytkownika, np. 1000
`mod.sample`	proporcja danych stosowanych do estymacji parametru sigma
`R`	liczba przedziałów w każdej iteracji
`iterations`	maksymalna liczba iteracji

Graficzną prezentację wybranych kroków klasyfikacji spektralnej dla danych metrycznych przedstawiających strukturę dwóch klas zobrazowano na rys. 3. Do wygenerowania zbioru danych metrycznych wykorzystano funkcję `mlbench.spirals` pakietu `mlbench` (zob. rys. 3a). Do klasyfikacji zbioru obiektów zastosowano metodę klasyfikacji spektralnej wyznaczając w kroku 4 macierz podobieństw zgodnie ze wzorem (1) z odległością GDM1. Rys. 3b i 3c prezentują odpowiednio obiekty z macierzy $\mathbf{E}$ o wymiarach 200×2 (krok 6) oraz obiekty ze znormalizowanej macierz $\mathbf{Y} = \left[y_{ij}\right]$ o wymiarach 200×2 (krok 7).

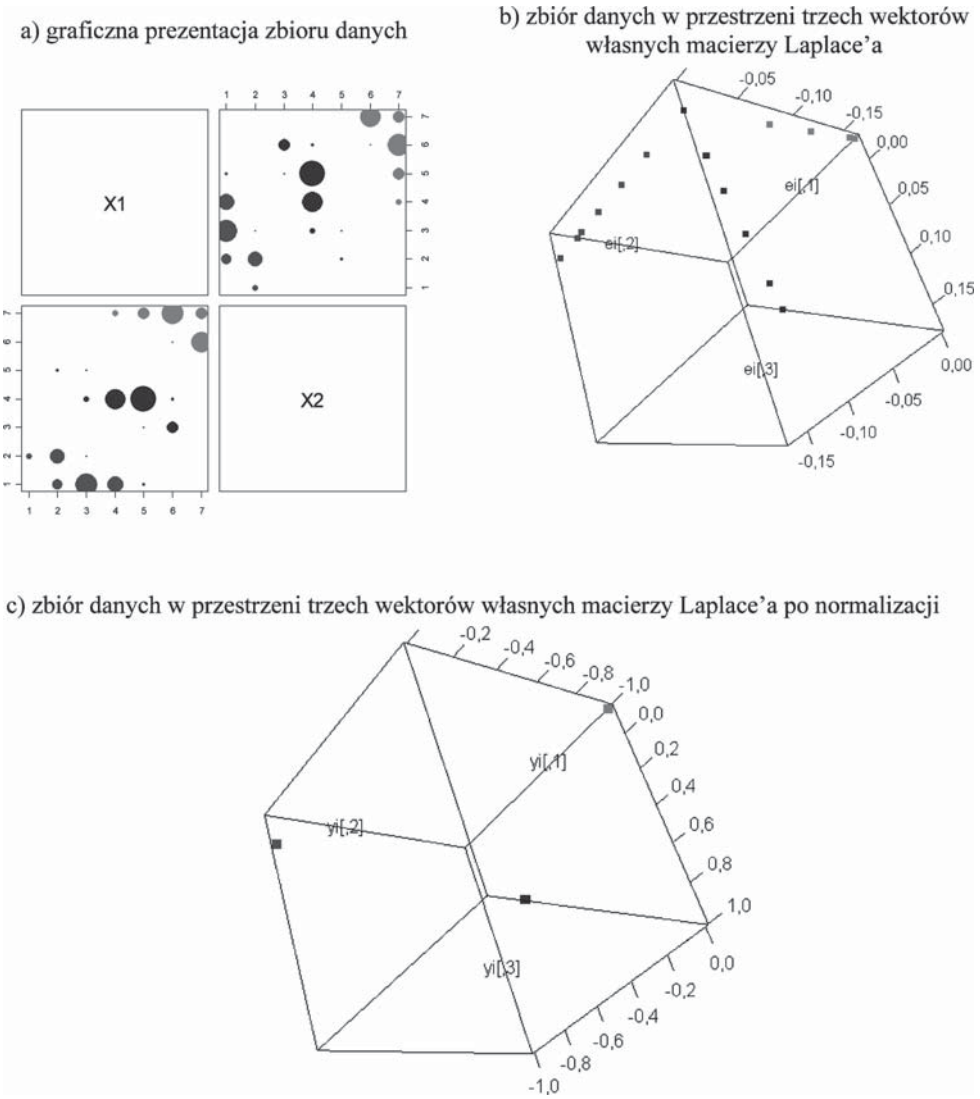
Graficzną prezentację wybranych kroków klasyfikacji spektralnej dla danych porządkowych przedstawiających strukturę trzech klas zobrazowano na rys. 4. Do wygenerowania zbioru danych porządkowych[3] wykorzystano funkcję `cluster.Gen` pakietu `clusterSim` (zob. rys. 4a). Do klasyfikacji zbioru obiektów zastosowano metodę klasyfikacji spektralnej wyznaczając w kroku 4 macierz podobieństw zgodnie ze wzorem (1) z odległością `GDM2`. Rys. 4b i 4c prezentują odpowiednio obiekty z macierzy $\mathbf{E}$ o wymiarach 150×3 (krok 6) oraz obiekty ze znormalizowanej macierz $\mathbf{Y} = \left[y_{ij}\right]$ o wymiarach 150×3 (krok 7).

[3] Przy tworzeniu wykresu rozrzutu dla danych porządkowych trzeba wziąć pod uwagę częstość występowania identycznych par kategorii. W funkcji `plotCategorial` znajduje to wyraz w długości promienia koła.



Rysunek 3. Wybrane etapy klasyfikacji spektralnej dla przykładowego zbioru danych metrycznych wygenerowanego z wykorzystaniem funkcji `mlbench.spirals` pakietu `mlbench`

Źródło: Opracowanie własne.



Rysunek 4. Wybrane etapy klasyfikacji spektralnej dla przykładowego zbioru danych porządkowych wygenerowanego z wykorzystaniem funkcji `clusterGen` pakietu `clusterSim`

Źródło: Opracowanie własne.

6. ANALIZA PORÓWNAWCZA METOD KLASYFIKACJI SPEKTRALNEJ Z METODAMI ANALIZY SKUPIEŃ DLA DANYCH O ZNANEJ STRUKTURZE KLAS

Analizę porównawczą metod klasyfikacji spektralnej z metodami analizy skupień, z uwzględnieniem różnych miar odległości, dla danych o znanej strukturze klas przeprowadzono dla trzech typów danych.

W eksperymencie pierwszym i trzecim wykorzystano odpowiednio dane metryczne oraz porządkowe o znanej strukturze klas obiektów wygenerowane z wykorzystaniem funkcji `cluster.Gen` pakietu clusterSim na podstawie modeli zawartych w tab. 3.

Tabela 3.

Charakterystyka modeli w analizie symulacyjnej

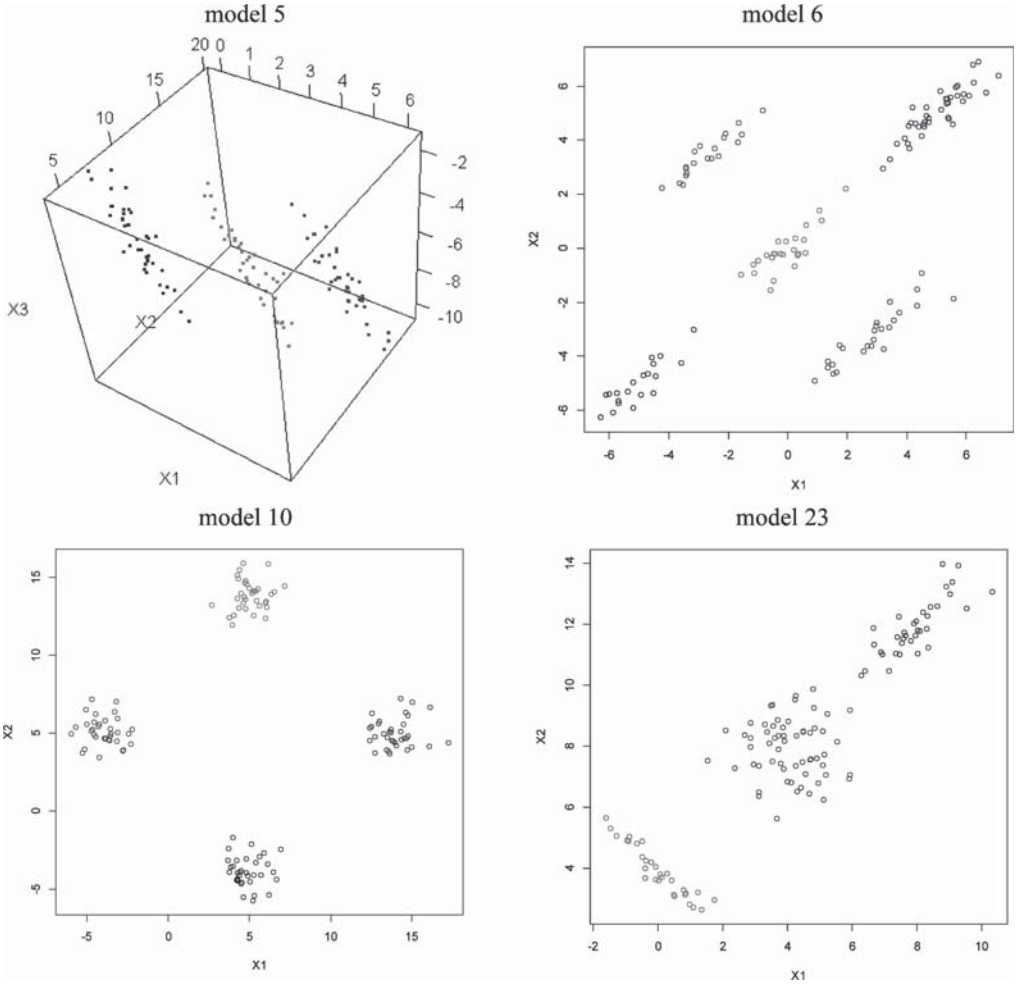
Nr modelu	*m*	*nk**	*u*	*lo*	środki ciężkości klas	Macierz kowariancji	*ks*
5	3	7	3	40	(1,5; 6, – 3), (3; 12; –6) (4,5; 18; –9)	$\sigma_{jj} = 1\ (1 \le j \le 3)$ $\sigma_{12} = \sigma_{13} = -0,9, \quad \sigma_{23} = 0,9$	1
6	2	5, 7	5	40, 20, 25, 25, 20	(5; 5), (–3; 3), (3; –3), (0; 0), (–5; –5)	$\sigma_{jj} = 1, \quad \sigma_{jl} = 0,9$	2
10	2	6, 8	4	35	(–4; 5), (5; 14), (14; 5), (5; –4)	$\sigma_{jj} = 1, \quad \sigma_{jl} = 0$	3
23	2	5	3	30, 60, 35	(0; 4), (4; 8), (8; 12)	$\Sigma_1 = \begin{bmatrix} 1 & -0,9 \\ -0,9 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 1 & 0,9 \\ 0,9 & 1 \end{bmatrix}$	4

* tylko dla danych porządkowych;
m – liczba zmiennych, *nk* – liczba kategorii (jedna liczba oznacza stałą liczbę kategorii); *u* – liczba klas; *lo* – liczba obiektów w klasach (jedna liczba oznacza klasy równoliczne); *ks* – kształt skupień: a) skupienia dobrze separowalne (1 – skupienia wydłużone, 3 – skupienia normalne), skupienia słabo separowalne (2 – skupienia wydłużone, 4 – skupienia zróżnicowane dla klas).
Źródło: [21].

Na rys. 5 i 6 przedstawiono graficzną prezentację przykładowych zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` dla danych metrycznych (rys. 5) i danych porządkowych (rys. 6).

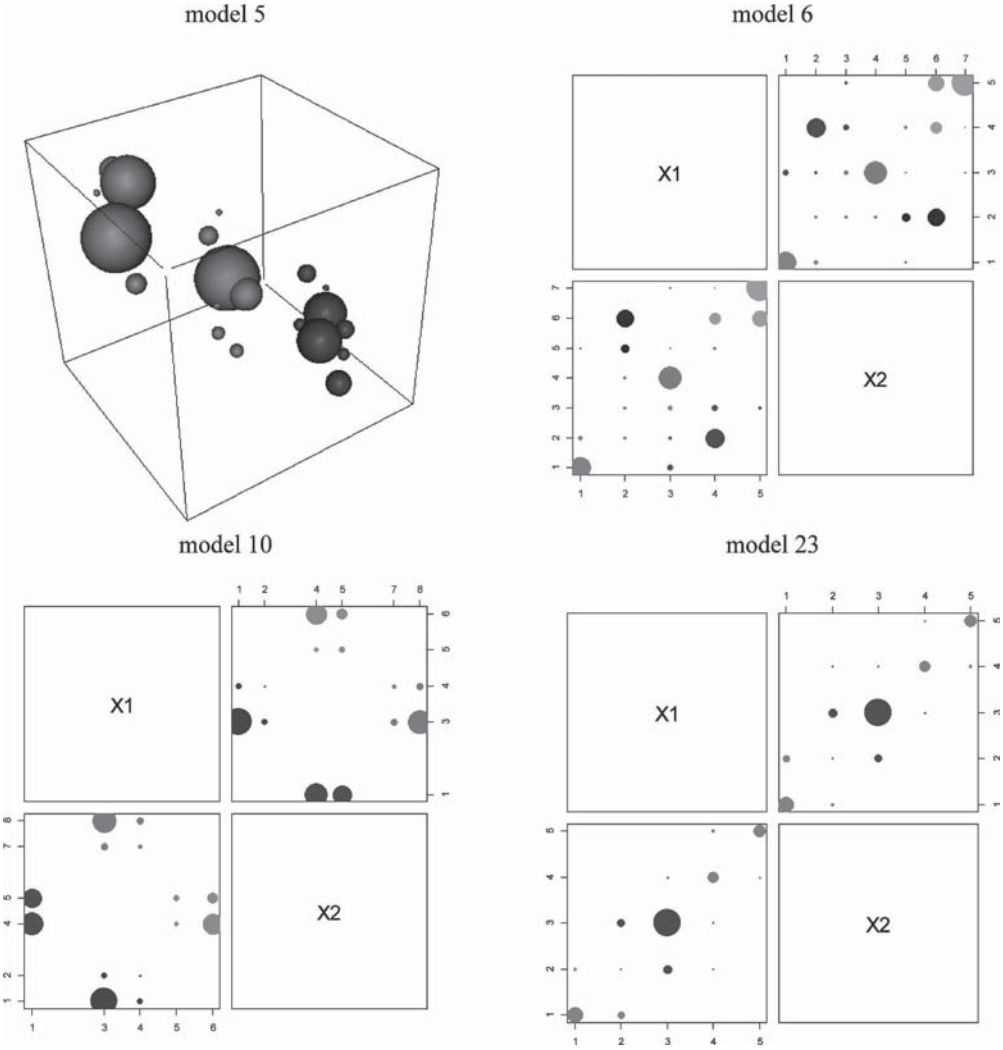
W eksperymencie drugim zbiory danych zawierające 360 obiektów (zob. rys. 7) wygenerowano z wykorzystaniem funkcji pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` [20] i `banana` [2].

Dla modeli w każdym eksperymencie wygenerowano 40 zbiorów danych, przeprowadzono procedurę klasyfikacyjną i porównano otrzymane rezultaty klasyfikacji ze znaną strukturą klas przy pomocy skorygowanego indeksu Randa (zob. [5]).



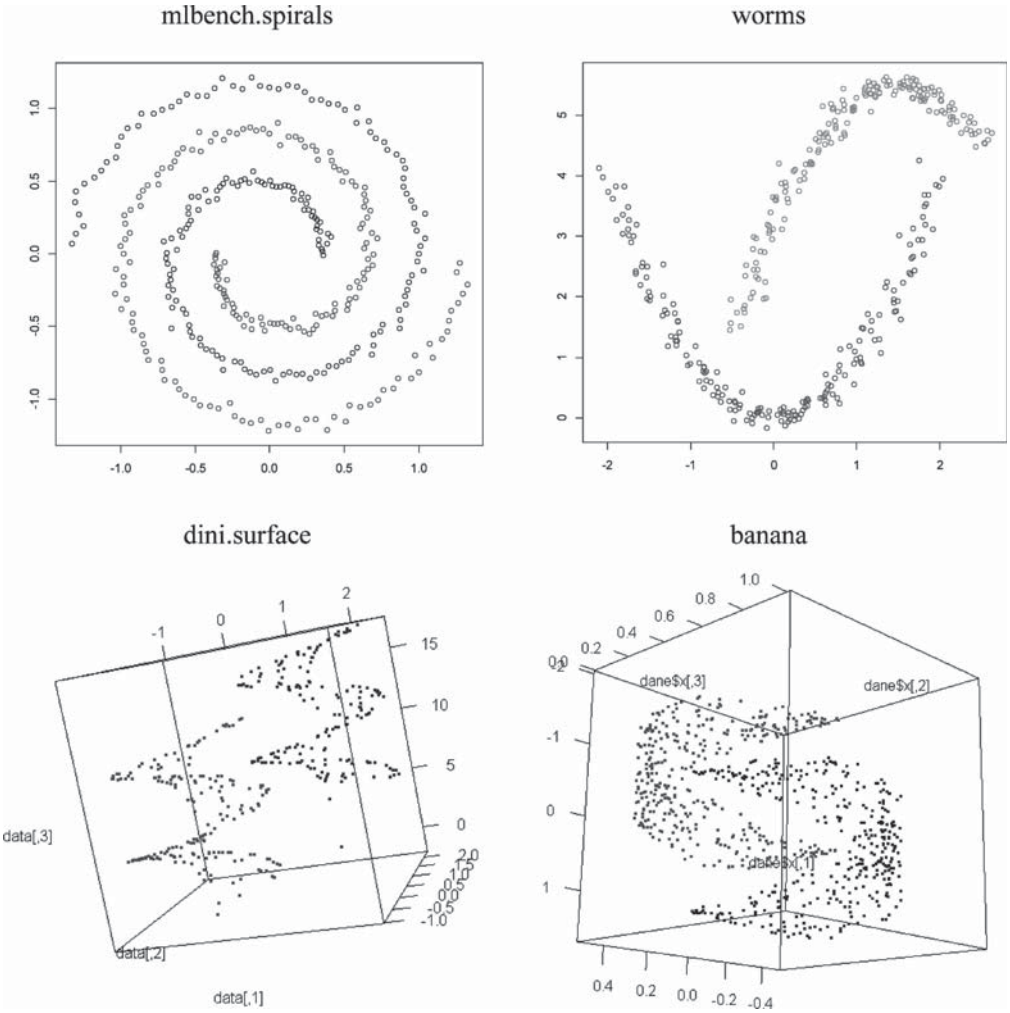
Rysunek 5. Graficzna prezentacja przykładowych zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` (dane metryczne)

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



Rysunek 6. Graficzna prezentacja przykładowych zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` (dane porządkowe)

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



Rysunek 7. Przykładowe zbiory danych utworzone z wykorzystaniem funkcji pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `banana`

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

Dla danych metrycznych (eksperyment 1 i 2) uwzględniono następujące metody klasyfikacji (z odległościami: (1) – kwadrat odległości euklidesowej, (2) – odległość euklidesowa, (3) – odległość miejska, (4) – odległość GDM1): 1. speccl – klasyfikacja spektralna; 2. pam – metoda *k*-medoidów; 3. complete – metoda kompletnego połączenia; 4. average – metoda średniej klasowej; 5. ward – metoda Warda; 6. centroid – metoda środka ciężkości; 7. diana – hierarchiczna metoda deglomeracyjna. Ponadto dla celów porównawczych zastosowano dodatkowo metodę *k*-średnich, która bazuje bezpośrednio na macierzy danych.

Dla danych porządkowych (eksperyment 3) uwzględniono w analizie metody klasyfikacji o numerach 1-7 z odległością GDM2.

Tab. 4 prezentuje uporządkowanie analizowanych metod klasyfikacji (z 4 odległościami) według średnich wartości skorygowanego indeksu Randa policzonego z 40 symulacji dla danych metrycznych wygenerowanych w pakiecie clusterSim.

Tabela 4.

Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych metrycznych wygenerowanych w pakiecie clusterSim

Poz.	Metoda	średnia*	Kształt skupień				Liczba zmiennych zakłócających		
			1	2	3	4	0	1	2
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	ward(3)	0,753	1,000	0,983	1,000	0,960	0,985	0,701	0,574
2	average(3)	0,746	0,989	0,974	1,000	0,954	0,979	0,695	0,563
3	pam(3)	0,736	0,999	0,992	1,000	0,967	0,989	0,701	0,517
4	speccl(2)	0,723	1,000	0,953	1,000	0,919	0,968	0,780	0,420
5	diana(3)	0,676	0,934	0,758	0,997	0,791	0,870	0,646	0,513
6	speccl(4)	0,670	0,899	0,851	0,898	0,867	0,879	0,695	0,435
7	speccl(1)	0,658	0,900	0,945	0,925	0,848	0,904	0,767	0,303
8	speccl(3)	0,643	0,950	0,948	0,975	0,884	0,939	0,692	0,298
9	complete(3)	0,632	0,751	0,733	1,000	0,921	0,851	0,573	0,471
10	average(4)	0,601	0,989	0,982	1,000	0,943	0,978	0,496	0,328
11	average(2)	0,599	1,000	0,975	1,000	0,957	0,983	0,495	0,320
12	ward(4)	0,594	1,000	0,978	1,000	0,960	0,984	0,478	0,320
13	pam(4)	0,591	1,000	0,975	1,000	0,924	0,974	0,482	0,317
14	ward(2)	0,590	1,000	0,979	1,000	0,960	0,985	0,471	0,316
15	pam(2)	0,582	1,000	0,992	1,000	0,966	0,989	0,455	0,302
16	centroid(3)	0,581	0,878	0,906	1,000	0,923	0,927	0,570	0,245
16	pam(1)	0,581	1,000	0,992	1,000	0,967	0,989	0,442	0,312
18	average(1)	0,577	1,000	0,973	1,000	0,962	0,984	0,463	0,285
19	ward(1)	0,575	1,000	0,979	1,000	0,964	0,986	0,446	0,293
20	centroid(4)	0,559	0,989	0,980	1,000	0,946	0,979	0,442	0,256

cd. Tabela 4.

21	diana(2)	0,534	0,988	0,757	0,983	0,846	0,894	0,438	0,270
22	centroid(1)	0,513	0,989	0,964	1,000	0,959	0,978	0,371	0,190
23	kmeans	0,502	0,819	0,839	0,898	0,967	0,881	0,406	0,219
24	diana(4)	0,491	0,966	0,762	0,992	0,582	0,826	0,404	0,242
25	diana(1)	0,457	0,988	0,735	0,992	0,678	0,848	0,345	0,179
26	complete(4)	0,447	0,894	0,912	1,000	0,886	0,923	0,281	0,135
27	complete(2)	0,437	0,960	0,869	1,000	0,931	0,940	0,253	0,119
27	complete(1)	0,437	0,960	0,869	1,000	0,931	0,940	0,253	0,119
29	centroid(2)	0,427	0,989	0,942	1,000	0,926	0,964	0,298	0,019

* (k8+k9+k10)/3, gdzie k8 = (k4+k5+k6+k7)/4
Liczba w nawiasie przy nazwach metod klasyfikacji: (1) – kwadrat odległości euklidesowej, (2) – odległość euklidesowa, (3) – odległość miejska, (4) – odległość GDM1.

Źródło: obliczenia własne z wykorzystaniem programu R[4].

W przypadku typowych zbiorów danych metrycznych metody klasyfikacji spektralnej sprawdzają się dobrze w odkrywaniu rzeczywistej struktury klas (pozycje: 4, 6, 7 i 8 w zestawieniu). W przeprowadzonym eksperymencie najlepiej strukturę klas odkrywały metody klasyczne (ward, average i pam) z odległością miejską. Wśród metod klasyfikacji spektralnej dominuje klasyfikacja spektralna z odległością euklidesową. Nieco gorsze rezultaty otrzymuje się z wykorzystaniem klasyfikacji spektralnej z odległością GDM1 (poz. 6 w zestawieniu).

Tab. 5 prezentuje uporządkowanie analizowanych metod klasyfikacji (z 4 odległościami) według średnich wartości skorygowanego indeksu Randa policzonego z 40 symulacji dla nietypowych danych metrycznych wygenerowanych z wykorzystaniem pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `banana`.

Dla nietypowych zbiorów danych metody klasyfikacji spektralnej zdecydowanie lepiej od klasycznych metod analizy skupień odkrywają prawidłową strukturę klas. Klasyfikacja spektralna z odległością GDM1 daje rezultaty lepsze od metod klasyfikacji spektralnej z pozostałymi odległościami.

Tab. 6 prezentuje uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 40 symulacji dla danych porządkowych wygenerowanych w pakiecie `clusterSim`.

W przypadku zbiorów danych porządkowych bez zmiennych zakłócających najlepsza jest metoda Warda. Metoda klasyfikacji spektralnej z odległością GDM2 daje gorsze rezultaty od klasycznych metod analizy skupień (za wyjątkiem metody diana). Należy jednak pamiętać, że zbiory tego typu bardzo rzadko występują w rzeczywistych problemach klasyfikacyjnych. Uwzględnienie zmiennych zakłócających pokazuje wyraźną przewagę metody klasyfikacji spektralnej z odległością GDM2.

[4] Skrypty do analiz symulacyjnych z punktu 6 są autorstwa dra Andrzeja Dudka.

Tabela 5.

Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych metrycznych otrzymanych z pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `banana`

Poz.	Metoda	średnia*	Zbiory danych			
			spirals	worms	dini	banana
1	*2*	*3*	*4*	*5*	*6*	*7*
1	speccl(4)	0,837	0,961	0,929	0,916	0,544
2	speccl(3)	0,822	0,901	0,959	0,694	0,736
3	speccl(2)	0,779	0,938	0,985	0,563	0,631
4	speccl(1)	0,741	1,000	0,889	0,407	0,671
5	pam(3)	0,259	0,006	0,438	0,274	0,316
6	average(3)	0,251	0,030	0,468	0,221	0,284
7	ward(3)	0,249	0,034	0,474	0,214	0,276
8	pam(2)	0,220	0,025	0,503	0,184	0,169
9	complete(3)	0,216	0,022	0,440	0,206	0,195
10	pam(4)	0,214	0,016	0,519	0,172	0,147
10	pam(1)	0,214	0,026	0,517	0,175	0,138
12	ward(1)	0,213	0,036	0,499	0,170	0,148
13	average(2)	0,211	0,039	0,517	0,152	0,135
13	diana(2)	0,211	0,040	0,528	0,155	0,120
15	diana(4)	0,210	0,037	0,516	0,158	0,129
16	diana(3)	0,209	-0,001	0,486	0,181	0,172
16	complete(2)	0,209	0,033	0,488	0,141	0,172
16	complete(1)	0,209	0,033	0,488	0,141	0,172
19	ward(2)	0,208	0,053	0,471	0,144	0,165
20	average(4)	0,205	0,029	0,471	0,143	0,177
20	kmeans	0,205	0,032	0,519	0,159	0,111
22	diana(1)	0,204	0,032	0,515	0,159	0,112
22	average(1)	0,204	0,034	0,503	0,150	0,130
24	ward(4)	0,202	0,046	0,487	0,142	0,132
25	centroid(1)	0,197	0,022	0,520	0,141	0,107
26	centroid(4)	0,194	0,038	0,478	0,145	0,116
27	complete(4)	0,193	0,041	0,464	0,140	0,126
28	centroid(3)	0,170	0,006	0,460	0,134	0,079
29	centroid(2)	0,167	0,020	0,487	0,083	0,078

* (k4+k5+k6+k7)/4

Liczba w nawiasie przy nazwach metod klasyfikacji: (1) – kwadrat odległości euklidesowej, (2) – odległość euklidesowa, (3) – odległość miejska, (4) – odległość GDM1.

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

Tabela 6.
Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych porządkowych wygenerowanych w pakiecie `clusterSim`

Poz.	Metoda	średnia*	Kształt skupień				Liczba zmiennych zakłócających		
			1	2	3	4	0	1	2
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	speccl(5)	0,696	0,998	0,951	0,798	0,777	0,881	0,709	0,497
2	average(5)	0,602	1,000	0,968	1,000	0,962	0,982	0,495	0,327
3	pam(5)	0,593	1,000	0,971	1,000	0,934	0,976	0,483	0,321
4	ward(5)	0,591	1,000	0,971	1,000	0,973	0,986	0,471	0,317
5	centroid(5)	0,560	1,000	0,962	1,000	0,965	0,982	0,451	0,248
6	diana(5)	0,493	0,959	0,753	0,998	0,595	0,826	0,388	0,266
7	complete(5)	0,444	0,882	0,885	1,000	0,851	0,904	0,279	0,149

* (k8+k9+k10)/3, gdzie: k8 = (k4+k5+k6+k7)/4
Liczba (5) w nawiasie przy nazwach metod klasyfikacji oznacza odległość GDM2.
Źródło: obliczenia własne z wykorzystaniem programu R.

7. PODSUMOWANIE

W artykule zaproponowano modyfikację metody klasyfikacji spektralnej umożliwiającą jej zastosowanie w klasyfikacji danych prezentowanych na różnych skalach pomiaru. W procedurze klasyfikacji spektralnej, zaproponowanej przez autorów Ng, Jordan i Weiss [8], wprowadzono modyfikację polegającą na zastosowaniu funkcji (1) z miarami odległości właściwymi dla danych mierzonych na różnych skalach pomiaru. Dodatkowo dzięki takiemu podejściu pośrednio wzmacnia się skale pomiaru zmiennych. Dane niemetryczne zostają przekształcone w dane przedziałowe. Umożliwia to zastosowanie w klasyfikacji zbioru obiektów m.in. metody *k*-średnich. Scharakteryzowano funkcję `speccl` pakietu `clusterSim` umożliwiającą klasyfikację spektralną zgodną z algorytmem zmodyfikowanym w artykule.

W tym miejscu wskazać trzeba na ograniczenia związane z klasyfikacją spektralną. Efektywne wykorzystanie metod klasyfikacji spektralnej jest uzależnione od prawidłowego doboru parametru skali σ. W części 4 zaprezentowano heurystyczną metodę poszukiwania minimum lokalnego.

W części 6 poświęconej analizie porównawczej metod klasyfikacji spektralnej z metodami analizy skupień dla danych o znanej strukturze klas przeprowadzono analizy symulacyjne dla danych metrycznych oraz porządkowych. Nie uwzględniono badań symulacyjnych dotyczących takiego porównania dla danych nominalnych. Wynikało to z braku metod generowania danych nominalnych o znanej strukturze klas.

Uniwersytet Ekonomiczny we Wrocławiu

LITERATURA

[1] Ackoff R.L. (1969), *Decyzje optymalne w badaniach stosowanych*, PWN, Warszawa.
[2] Dudek A. (2012), *A comparison of the performance of clustering methods using spectral approach*, W: J. Pociecha, R. Decker (red.), *Data analysis methods and its applications*, Wydawnictwo C.H. Beck, Warszawa, 143-156.
[3] Fischer I., Poland J. (2004), *New methods for spectral clustering*, Technical Report No. IDSIA-12-04, Dalle Molle Institute for Artificial Intelligence, Manno-Lugano, Switzerland.
[4] Girolami M. (2002), *Mercer kernel-based clustering in feature space*, „IEEE Transactions on Neural Networks", vol. 13, no. 3, 780-784.
[5] Hubert L., Arabie P. (1985), *Comparing partitions*, „Journal of Classification", no. 1, 193-218.
[6] Karatzoglou A. (2006), *Kernel methods. Software, algorithms and applications*, Rozprawa doktorska, Uniwersytet Techniczny we Wiedniu.
[7] Kolupa M. (1976), *Elementarny wykład algebry liniowej dla ekonomistów*, Państwowe Wydawnictwo Naukowe, Warszawa.
[8] Ng A., Jordan M., Weiss Y. (2002), *On spectral clustering: analysis and an algorithm*, W: T. Dietterich, S. Becker, Z. Ghahramani (red.), Advances in Neural Information Processing Systems 14, Cambridge, MIT Press, 849-856.
[9] Poland J., Zeugmann T. (2006), *Clustering the Google distance with eigenvectors and semidefinite programming*, Knowledge Media Technologies, First International Core-to-Core Workshop, Dagstuhl, July 23-27, Germany.
[10] Shortreed S. (2006), *Learning in spectral clustering*, Rozprawa doktorska, University of Washington.
[11] Steczkowski J., Zeliaś A. (1981), *Statystyczne metody analizy cech jakościowych*, PWE, Warszawa.
[12] Steczkowski J., Zeliaś A. (1997), *Metody statystyczne w badaniach cech jakościowych*, Wydawnictwo AE, Kraków.
[13] Stevens S.S. (1946), *On the theory of scales of measurement*, „Science", Vol. 103, No. 2684, 677-680.
[14] Verma D., Meila M. (2003), *A comparison of spectral clustering algorithms*. Technical report UW-CSE-03-05-01, University of Washington.
[15] von Luxburg U. (2007), *A tutorial on spectral clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.
[16] Walesiak M. (1990), *Syntetyczne badania porównawcze w świetle teorii pomiaru*, „Przegląd Statystyczny", z. 1-2, 37-46.
[17] Walesiak M. (2005), *Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów*, W: A. Zeliaś (red.), „Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych", Wydawnictwo AE, Kraków, 185-203.
[18] Walesiak M. (2006), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*. Wydanie drugie rozszerzone. Wydawnictwo AE, Wrocław.
[19] Walesiak M. (2009), *Analiza skupień*, W: M. Walesiak, E. Gatnar (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa, 407-433.
[20] Walesiak M., Dudek A. (2009), *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe UE we Wrocławiu nr 84, 9-19.
[21] Walesiak M., Dudek A. (2010), *Klasyfikacja spektralna z wykorzystaniem odległości GDM*, W: K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 17, Prace Naukowe UE we Wrocławiu nr 107, 161-171.
[22] Walesiak M., Dudek A. (2011), *clusterSim package*, URL http://www.R-project.org.
[23] Wiśniewski J.W. (1986), *Korelacja i regresja w badaniach zjawisk jakościowych na tle teorii pomiaru*, „Przegląd Statystyczny", z. 3, 239-248.
[24] Wiśniewski J.W. (1987), *Teoria pomiaru a teoria błędów w badaniach statystycznych*, „Wiadomości Statystyczne", nr 11, 18-20.

[25] Zelnik-Manor L., Perona P. (2004), *Self-tuning spectral clustering*, W: Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04), http://books.nips.cc/nips17.html.

KLASYFIKACJA SPEKTRALNA A SKALE POMIARU ZMIENNYCH

Streszczenie

W artykule zaproponowano modyfikację metody klasyfikacji spektralnej (zob. Ng, Jordan i Weiss [2002]) umożliwiającą jej zastosowanie w klasyfikacji danych nominalnych, porządkowych, przedziałowych oraz ilorazowych. W tym celu w procedurze tej metody przy wyznaczaniu macierzy podobieństwa (*affinity matrix*) zastosowano funkcję $A_{ik} = \exp(-\sigma \cdot d_{ik})$ (σ – parametr skali) z miarami odległości d_{ik} właściwymi dla danych mierzonych na różnych skalach pomiaru. Takie podejście umożliwia ponadto pośrednie wzmocnienie skali pomiaru zmiennych dla danych niemetrycznych.

Zaproponowana metoda klasyfikacji spektralnej może być z powodzeniem stosowana we wszystkich zagadnieniach klasyfikacyjnych, w tym dotyczących pomiaru, analizy i wizualizacji preferencji.

Słowa kluczowe: klasyfikacja spektralna, miary odległości, skale pomiaru

SPECTRAL CLUSTERING AND MEASUREMENT SCALES OF VARIABLES

Abstract

In article the proposal of modification of spectral clustering method for nominal, ordinal, interval and ratio data, based on procedure of Ng, Jordan and Weiss [2002], is presented. In construction of affinity matrix we implement function $A_{ik} = \exp(-\sigma \cdot d_{ik})$ (σ – scale parameter) with distance measures d_{ik} appropriate for different scales of measurement. This approach gives possibility of conversion nonmetric data (nominal, ordinal) into interval data.

The proposed method of spectral clustering can be successfully used in all classification problems, including the measurement, analysis and visualization of preferences.

Key words: spectral clustering, distance measures, scales of variables