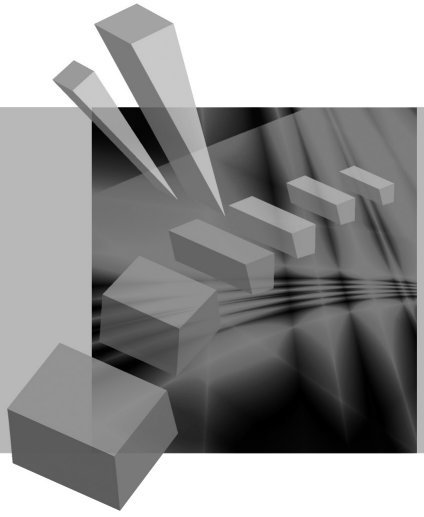
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

278

Taksonomia 20

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 278 • 2013

Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Marek Walesiak
Uniwersytet Ekonomiczny we Wrocławiu

ZAGADNIENIE DOBORU LICZBY KLAS W KLASYFIKACJI SPEKTRALNEJ

Streszczenie: W artykule przetestowano przydatność pięciu indeksów oceny jakości klasyfikacji w zagadnieniu doboru liczby klas w klasyfikacji spektralnej uwzględniającej cztery typy odległości (kwadrat odległości euklidesowej, odległość euklidesowa, odległość miejska, odległość GDM1). W eksperymentach wykorzystano klasyczne dane metryczne o znanej strukturze klas obiektów wygenerowane z wykorzystaniem z funkcji `cluster.Gen` pakietu `clusterSim` oraz nieklasyczne zbiory danych utworzone z wykorzystaniem funkcji pakietu `mlbench`, `geozoo` oraz zbiorów własnych. Dla modeli w każdym eksperymencie wygenerowano 40 zbiorów danych, przeprowadzono klasyfikację spektralną z zastosowaniem odpowiedniego indeksu i otrzymane rezultaty klasyfikacji porównano ze znaną strukturą klas za pomocą skorygowanego indeksu Randa.

Słowa kluczowe: analiza skupień, klasyfikacja spektralna, liczba klas.

1. Wstęp

Zagadnienie doboru liczby klas należy do najważniejszych kroków w każdej procedurze klasyfikacyjnej.

W artykule przetestowano przydatność pięciu indeksów oceny jakości klasyfikacji w zagadnieniu doboru liczby klas w klasyfikacji spektralnej uwzględniającej cztery typy odległości. W eksperymentach wykorzystano klasyczne dane metryczne o znanej strukturze klas obiektów wygenerowane z wykorzystaniem z funkcji `cluster.Gen` pakietu `clusterSim` oraz nieklasyczne zbiory danych utworzone z wykorzystaniem funkcji pakietu `mlbench`, `geozoo` oraz zbiorów własnych.

2. Klasyfikacja spektralna

W jednym z podstawowych kroków klasyfikacji spektralnej wyznacza się spektrum (widmo) macierzy Laplace'a. W matematyce zbiór wartości własnych macierzy nazywa się spektrum (widmem) macierzy (zob. np. [Kolupa 1976, s. 182]). Podstawowy algorytm klasyfikacji spektralnej dla danych metrycznych zaproponowano

w pracy Ng, Jordan i Weiss [2002]. Inne algorytmy klasyfikacji spektralnej scharakteryzowano m.in. w pracach Shortreed [2006] oraz Verma i Meila [2003].

Procedura klasyfikacji spektralnej obejmuje następujące kroki[1]:

1. Ustalenie zbioru obiektów i zmiennych. Po zgromadzeniu danych konstruuje się macierz danych $\mathbf{X} = \left[x_{ij} \right]_{n \times m}$ (i – numer obiektu, j – numer zmiennej), a w przypadku danych metrycznych znormalizowaną macierz danych $\mathbf{Z} = \left[z_{ij} \right]_{n \times m}$.

2. Dobór zmiennych.

Szczegółową charakterystykę etapów 1-2 zaprezentowano m.in. w pracach Walesiaka [2005; 2009].

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

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

gdzie: σ – parametr skali,
d_{ik} – miary odległości dla różnych skal pomiaru (zob. Walesiak [2012]),
$i, k = 1, \ldots, n$ – numery obiektów.

W artykule przetestowano zastosowanie we wzorze (1) miar odległości d_{ik} dla danych metrycznych ujętych w tab. 1.

Tabela 1. Miary odległości d_{ik} dla danych metrycznych

Nr	Nazwa miary odległości	Formuła	Funkcja (pakiet) programu R
1	kwadrat odległości euklidesowej	$d_{ik} = \sum_{j=1}^{m} (z_{ij} - z_{kj})^2$	`dist (stats)`
2	euklidesowa	$d_{ik} = \sqrt{\sum_{j=1}^{m} (z_{ij} - z_{kj})^2}$	`dist (stats)`
3	miejska (Manhattan)	$d_{ik} = \sum_{j=1}^{m} \lvert z_{ij} - z_{kj} \rvert$	`dist (stats)`
4	GDM1	$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{2 \left[\sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2 \right]^{\frac{1}{2}}}$, $a_{ikj} = z_{ij} - z_{kj}$, $a_{ilj} = z_{ij} - z_{lj}$, $b_{kij} = z_{kj} - z_{ij}$, $b_{klj} = z_{kj} - z_{lj}$	`dist.GDM (clusterSim)`

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

Źródło: opracowanie własne.

[1] Jest to algorytm zaproponowany w pracy Ng, Jordan i Weiss [2002] (por. Walesiak i Dudek [2009; 2010]). W artykule Walesiaka [2012] dokonano jego modyfikacji w kroku 3 przy obliczaniu macierzy podobieństw (*affinity matrix*).

W kroku tym można zastosować do obliczenia elementów macierzy podobieństw A_{ik} ($i \neq k$) estymatory jądrowe (zob. Karatzoglou [2006], s. 13-14; Poland i Zeugmann [2006]): jądro gaussowskie (z odległością (1) z tab. 1), jądro wielomianowe, jądro liniowe, jądro w postaci tangensa hiperbolicznego, jądro Bessela, jądro Laplace'a (z odległością (2) z tab. 1), jądro ANOVA, jądro łańcuchowe (dla danych tekstowych).

4. 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 $\mathbf{A} = [A_{ik}]$). W rzeczywistości znormalizowana macierz Laplace'a przyjmuje postać: $\mathbf{I} - \mathbf{L}$. W algorytmie dla uproszczenia analizy pomija się macierz jednostkową $\mathbf{I}$ (zob. Ng, Jordan i Weiss [2002]). Własności tej macierzy przedstawiono m.in. w pracy von Luxburg [2007], s. 5-6.

5. Obliczenie wartości własnych i odpowiadających im wektorów własnych dla macierzy $\mathbf{L}$, a następnie uporządkowanie wektorów własnych według malejących wartości własnych. Pierwsze u wektorów własnych (u – liczba klas) tworzy macierz $\mathbf{E} = [e_{ij}]$ o wymiarach $n \times u$.

6. 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 $\mathbf{Y} = [y_{ij}]$ jest równa jeden.

7. Macierz $\mathbf{Y}$ stanowi punkt wyjścia zastosowania klasycznych metod analizy skupień (proponuje się tutaj wykorzystanie metody *k*-średnich).

Podejście spektralne ujęte w krokach 3-6 nie jest nową metodą klasyfikacji. W wyniku zastosowania tego podejścia dokonuje się takiego rozmieszczenia obiektów w przestrzeni o liczbie wymiarów równej liczbie klas, aby klasy obiektów były wyraźnie separowalne. Klasyfikację obiektów przeprowadza się w podejściu spektralnym, wykorzystując w tym celu jedną z klasycznych metod analizy skupień (w prezentowanym algorytmie zastosowano metodę *k*-średnich).

3. Indeksy oceny jakości klasyfikacji służące wyborowi liczby klas

Do rozwiązania zagadnienia wyboru optymalnej liczby klas można w klasyfikacji spektralnej wykorzystać:

a. Metody bazujące na dekompozycji spektralnej (np. metodę Girolamiego [2002] – zob. Walesiak [2012]),

b. Indeksy oceny jakości klasyfikacji stosowane w klasycznej analizie skupień (np. indeksy z pakietu `clusterSim`: Daviesa-Bouldina – `index.DB`, Calińskiego

i Harabasza – `index.G1`, Bakera i Huberta – `index.G2`, Huberta i Levina – `index.G3`, gap – `index.Gap`, Hartigana – `index.H`, Krzanowskiego i Lai – `index.KL`, Silhouette – `index.S`).

W części symulacyjnej artykułu zastosowano w klasyfikacji spektralnej pięć indeksów służących wyborowi liczby klas (zob. tab. 2).

Tabela 2. Wybrane indeksy oceny jakości klasyfikacji służące wyborowi liczby klas

Lp.	Nazwa indeksu	Symbol	Formuła	Kryterium wyboru liczby klas
1	Zmienność wewnątrzklasowa	WC	$WC(u) = tr\mathbf{W}_u$	$\hat{u} = \arg\min_u \{WC(u)\}$
2	Calińskiego i Harabasza	G1	$G1(u) = \frac{B_u/(u-1)}{W_u/(n-u)}$, $G1(u) \in R_+$	$\hat{u} = \arg\max_u \{G1(u)\}$
3	Krzanowskiego i Lai	KL	$KL(u) = \left\lvert \frac{DIFF_u}{DIFF_{u+1}} \right\rvert$, $KL(u) \in R_+$ $DIFF_u = (u-1)^{2/m} W_{u-1} - u^{2/m} W_u$	$\hat{u} = \arg\max_u \{KL(u)\}$
4	Daviesa-Bouldina	DB	$DB(u) = \frac{1}{u}\sum_{r=1}^{u} \max_{\substack{s \\ r \neq s}} \left(\frac{S_r + S_s}{d_{rs}} \right)$	$\hat{u} = \arg\min_u \{DB(u)\}$
5	Hartigana	H	$H(u) = \left(\frac{W_u}{W_{u+1}} - 1 \right)(n-u-1)$, $H(u) \in R_+$	najmniejsze u, dla którego $H(u) \leq 10$

$\mathbf{B}_u$ – macierz kowariancji międzyklasowej, $\mathbf{W}_u$ – macierz kowariancji wewnątrzklasowej, tr – ślad macierzy, $B_u(W_u) = tr(\mathbf{B}_u)\left(tr\mathbf{W}_u\right)$, $r, s = 1, ..., u$ – numer klasy, u – liczba klas, $i, k = 1, \ldots, n$ – numer obiektu, n – liczba obiektów, $j = 1, \ldots, m$ – numer zmiennej, m – liczba zmiennych, $d_{rs} = \sqrt{\sum_{j=1}^{m} \left| z_{\bullet j}^r - z_{\bullet j}^s \right|^2}$ – odległość Euklidesa między środkami ciężkości klas r i s; $z_{\bullet j}^r$ ($z_{\bullet j}^s$) – j-ta współrzędna środka ciężkości klasy r (s); $S_r = \sqrt{\frac{1}{n_r}\sum_{i \in P_r}\sum_{j=1}^{m} \left| z_{ij}^r - z_{\bullet j}^r \right|^2}$ – miara rozproszenia obiektów w klasie (odchylenie standardowe odległości obiektów w r-tej klasie od środka ciężkości klasy).

Źródło: opracowanie własne na podstawie prac: Walesiak [2011], s. 61; Everitt, Landau, Leese i Stahl [2011], s. 114-115.

4. Analiza porównawcza indeksów oceny jakości klasyfikacji służących wyborowi liczby klas w klasyfikacji spektralnej z czterema miarami odległości

Analizę porównawczą na podstawie dwóch typów danych metrycznych (klasycznych i nieklasycznych) przeprowadzono dla pięciu indeksów z tab. 2 oraz czterech miar odległości z tab. 1 zastosowanych w klasyfikacji spektralnej.

W eksperymencie pierwszym wykorzystano klasyczne dane metryczne o znanej strukturze klas obiektów wygenerowane z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` (zob. Walesiak i Dudek [2012]) na podstawie modeli zawartych w tab. 3.

Tabela 3. Charakterystyka modeli w analizie symulacyjnej

nm	*m*	*u*	*lo*	Środki ciężkości klas	Macierz kowariancji $\sum$
5	3	3	40	(1,5; 6, – 3), (3; 12; –6) (4,5; 18; –9)	$\sigma_{jj} = 1$ $(1 \le j \le 3)$, $\sigma_{12} = \sigma_{13} = -0{,}9$, $\sigma_{23} = 0{,}9$
6	2	5	40, 20, 25, 25, 20	(5; 5), (–3; 3), (3; –3), (0; 0), (–5; –5)	$\sigma_{jj} = 1$, $\sigma_{jl} = 0{,}9$
23	2	3	30, 60, 35	(0; 4), (4; 8), (8; 12)	$\Sigma_1 = \begin{bmatrix} 1 & -0{,}9 \\ -0{,}9 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1{,}5 & 0 \\ 0 & 1{,}5 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{bmatrix}$
24	2	4	30	(0; 5), (5; 14), (14; 5), (5; –4)	$\sigma_{jj} = 1$, $\sigma_{jl} = 0$

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).

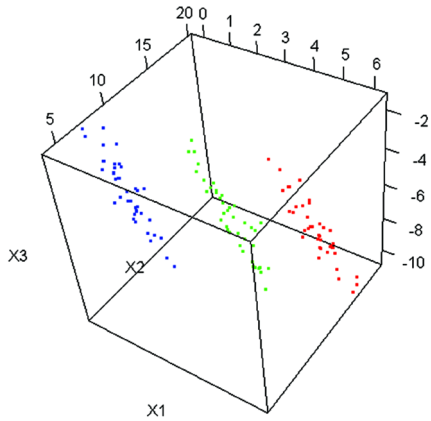
Źródło: opracowanie własne.

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

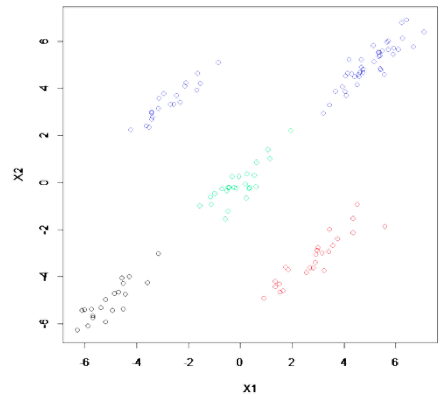
W eksperymencie drugim nieklasyczne zbiory danych zawierające 360 obiektów (zob. rys. 2) wygenerowano z wykorzystaniem pakietów `mlbench` (funkcja `mlbench.spirals`), `geozoo` (funkcja `dini.surface`) oraz zbiorów `worms` (Walesiak i Dudek [2009]) i `circles`.

Dla modeli w każdym eksperymencie wygenerowano 40 zbiorów danych, przeprowadzono klasyfikację spektralną z czterema odległościami i odpowiednimi indeksami wyboru liczby klas (rozważano podziały od 2 do 10 klas), a następnie porównano otrzymane rezultaty klasyfikacji ze znaną strukturą klas za pomocą skorygowanego indeksu Randa (zob. Hubert i Arabie [1985]).

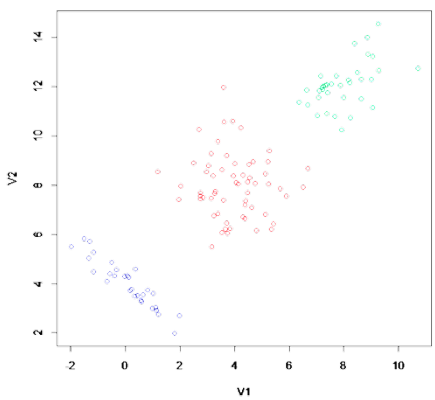
model 5 (skupienia dobrze separowalne – skupienia wydłużone)



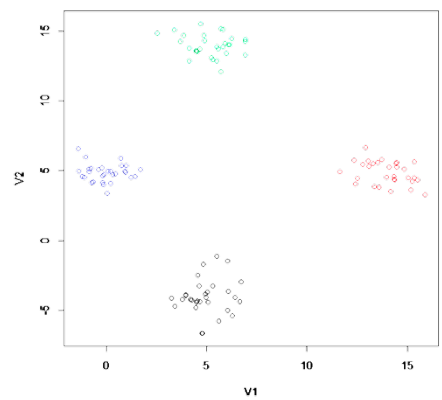
model 6 (skupienia słabo separowalne – skupienia wydłużone)



model 23 (skupienia słabo separowalne – skupienia zróżnicowane dla klas)



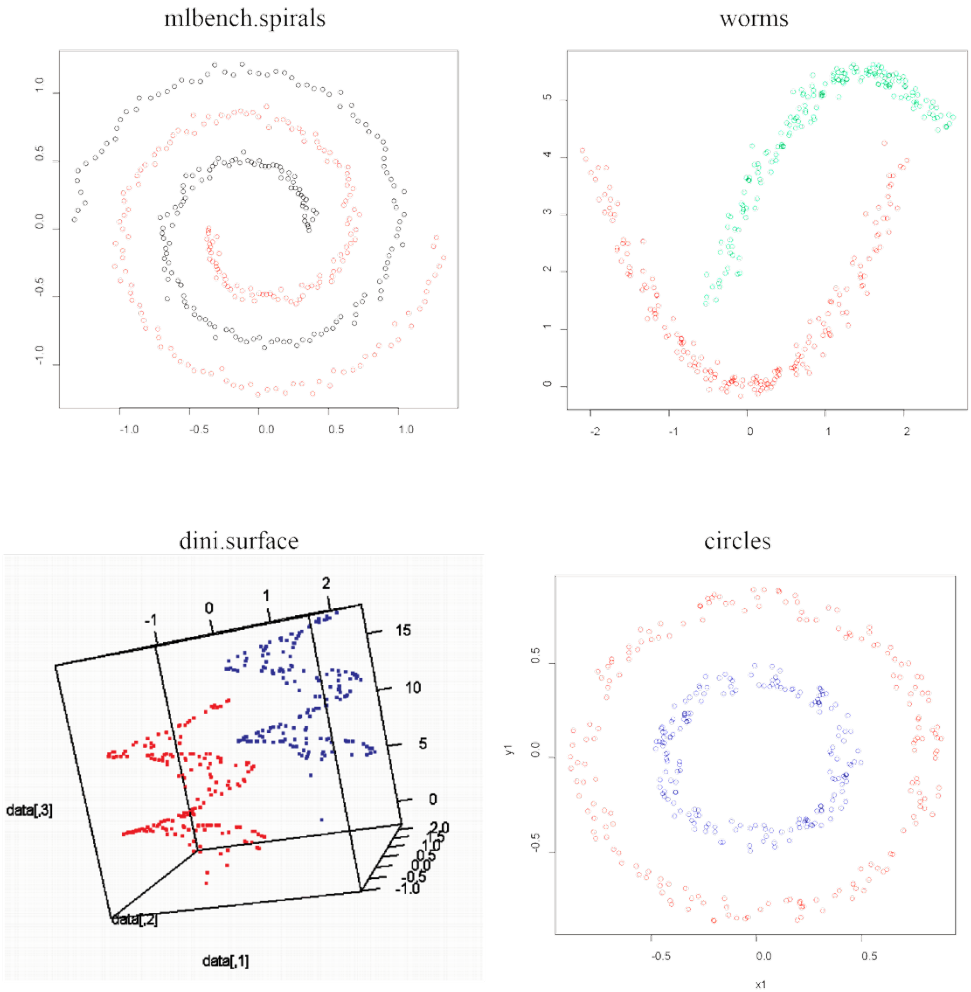
model 24 (skupienia dobrze separowalne – skupienia normalne)



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

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

Tabela 4 prezentuje uporządkowanie analizowanych metod klasyfikacji spektralnej (z 4 odległościami) zastosowanych z odpowiednimi indeksami wyboru liczby klas według średnich wartości skorygowanego indeksu Randa policzonego z 40 symulacji dla klasycznych danych metrycznych wygenerowanych w pakiecie `clusterSim`.



Rys. 2. Przykładowe zbiory danych utworzone z wykorzystaniem funkcji pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `circles`

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

W przypadku typowych zbiorów danych metrycznych najlepiej strukturę klas odkrywały metody klasyfikacji spektralnej z kwadratem odległości euklidesowej (z indeksami odpowiednio: WC, DB, G1, KL). Nieco gorsze rezultaty otrzymuje się z wykorzystaniem klasyfikacji spektralnej z odległością GDM1 z tymi samymi indeksami (poz. 4, 5, 6, 7 w zestawieniu). Najgorsze rezultaty otrzymuje się dla indeksu Hartigana.

Tabela 4. Uporządkowanie analizowanych metod klasyfikacji spektralnej z wybraną miarą odległości oraz indeksem oceny jakości 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
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	specc(1)_WC	0,754	0,977	0,612	0,539	0,935	0,766	0,742
2	specc(1)_DB	0,754	0,977	0,612	0,539	0,935	0,766	0,742
3	specc(1)_G1	0,751	0,977	0,612	0,539	0,935	0,766	0,737
4	specc(1)_KL	0,738	0,918	0,606	0,859	0,689	0,768	0,708
5	specc(4)_WC	0,732	0,946	0,536	0,628	0,969	0,770	0,694
6	specc(4)_DB	0,732	0,946	0,536	0,628	0,969	0,770	0,694
7	specc(4)_G1	0,728	0,946	0,536	0,625	0,952	0,765	0,691
8	specc(4)_KL	0,721	0,975	0,614	0,844	0,701	0,784	0,658
9	specc(3)_WC	0,691	0,780	0,865	0,747	0,763	0,789	0,592
10	specc(3)_DB	0,691	0,780	0,865	0,747	0,763	0,789	0,592
11	specc(3)_G1	0,660	0,843	0,874	0,730	0,735	0,796	0,525
12	specc(3)_KL	0,587	0,790	0,843	0,842	0,712	0,797	0,378
13	specc(2)_WC	0,577	0,908	0,885	0,555	0,833	0,795	0,359
14	specc(2)_DB	0,577	0,908	0,885	0,555	0,833	0,795	0,359
15	specc(2)_KL	0,560	0,965	0,749	0,919	0,722	0,839	0,281
16	specc(2)_G1	0,496	0,912	0,886	0,555	0,820	0,793	0,199
17	specc(1)_H	0,484	0,440	0,267	0,777	0,587	0,518	0,451
18	specc(4)_H	0,482	0,536	0,231	0,708	0,557	0,508	0,455
19	specc(2)_H	0,304	0,456	0,266	0,762	0,431	0,479	0,129
20	specc(3)_H	0,296	0,348	0,273	0,608	0,290	0,380	0,211

* $(k8 + k9)/2$, gdzie $k8 = (k4 + k5 + k6 + k7)/4$

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

Symbole indeksów wyjaśniono w tab. 2.

Źródło: obliczenia własne z wykorzystaniem programu R.

Tabela 5 prezentuje uporządkowanie analizowanych metod klasyfikacji (z 4 odległościami) zastosowanych z odpowiednimi indeksami wyboru liczby klas 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 `circles`.

W przypadku nietypowych zbiorów danych metrycznych najlepiej strukturę klas odkrywały metody klasyfikacji spektralnej z odległością GDM1 (z indeksami odpowiednio G1, WC, DB). Nieco gorsze rezultaty otrzymuje się z wykorzystaniem klasyfikacji spektralnej z kwadratem odległości euklidesowej (z indeksami odpowiednio: G1, WC, DB). Gorzej z poszczególnymi indeksami prezentowały się metody klasyfikacji spektralnej z odległościami odpowiednio euklidesową i miejską.

Tabela 5. Uporządkowanie analizowanych metod klasyfikacji spektralnej z wybraną miarą odległości oraz indeksem oceny jakości klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych metrycznych otrzymanych z pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `circles`

Poz.	Metoda	Średnia*	Zbiory danych			
			spirals	worms	dini	circles
1	*2*	*3*	*4*	*5*	*6*	*7*
1	specc(4)_G1	0,915	0,980	0,837	0,849	0,994
2	specc(4)_WC	0,914	0,980	0,835	0,849	0,994
3	specc(4)_DB	0,914	0,980	0,835	0,849	0,994
4	specc(1)_G1	0,886	0,994	0,962	0,590	1,000
5	specc(1)_WC	0,879	0,994	0,961	0,563	1,000
6	specc(1)_DB	0,879	0,994	0,961	0,563	1,000
7	specc(4)_KL	0,724	0,659	0,818	0,694	0,724
8	specc(1)_KL	0,718	0,731	0,755	0,662	0,724
9	specc(2)_G1	0,714	0,896	0,979	0,022	0,960
10	specc(2)_WC	0,708	0,858	0,965	0,053	0,956
11	specc(2)_DB	0,708	0,858	0,965	0,053	0,956
12	specc(3)_WC	0,682	0,877	0,759	0,149	0,943
13	specc(3)_DB	0,682	0,877	0,759	0,149	0,943
14	specc(3)_G1	0,681	0,889	0,770	0,122	0,943
15	specc(4)_H	0,654	0,547	0,754	0,668	0,648
16	specc(1)_H	0,648	0,649	0,844	0,383	0,715
17	specc(3)_KL	0,534	0,533	0,788	0,113	0,703
18	specc(2)_KL	0,514	0,536	0,797	0,050	0,674
19	specc(2)_H	0,462	0,417	0,792	0,024	0,615
20	specc(3)_H	0,440	0,370	0,647	0,066	0,675

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

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

Symbole indeksów wyjaśniono w tab. 2.

Źródło: obliczenia własne z wykorzystaniem programu R.

Skrypty do analiz symulacyjnych z punktu 4 są autorstwa dra Andrzeja Dudka. W analizach symulacyjnych wykorzystano funkcję `specc1` pakietu `clusterSim` w wersji 0.41-5, przyjmując w domyśle parametry służące wyszukiwaniu parametru skali σ. Parametr σ (zob. wzór (1)) ma fundamentalne znaczenie w klasyfikacji spektralnej. Poszukuje się takiej wartości parametru σ, która minimalizuje zmienność wewnątrzklasową przy zadanej liczbie klas u. Jest to heurystyczna metoda poszukiwania minimum lokalnego. W klasyfikacji spektralnej z odległościami: euklidesowa, kwadrat euklidesowej, miejska, otrzymane rezultaty klasyfikacji uzależnione są od górnej granicy przedziału przeszukiwania parametru sigma oraz od przyjętej liczby przedziałów w każdej iteracji (domyślnie: 10). W klasyfikacji spektralnej z odległością GDM1 górna granica nie ma wpływu na wyniki klasyfikacji. Górna

granica parametru sigma w zasadzie niewiele się zmienia dla danej liczby obiektów ze względu na unormowanie odległości GDM1 w przedziale [0; 1].

Wang [2010] przeprowadził m.in. analizę symulacyjną przydatności sześciu klasycznych indeksów oceny jakości klasyfikacji (Calińskiego i Harabasza, Hartigana, Krzanowskiego i Lai, gap, jump, Silhouette) oraz dwóch własnych propozycji dla metody klasyfikacji spektralnej zgodnie z algorytmem Ng, Jordan i Weiss [2002]. Analizę symulacyjną przeprowadzono dla dwóch zbiorów danych nieklasycznych. Zaskakująco słabe wyniki w odkrywaniu struktury klas odnotowano dla indeksu G1 Calińskiego i Harabasza. Przeprowadzony eksperyment symulacyjny w prezentowanym artykule dla danych nieklasycznych pokazuje odmienny rezultat. Prawdopodobnie indeksy oceny jakości klasyfikacji obliczono w artykule Wanga [2010] na podstawie pierwotnej macierzy danych (krok 1 algorytmu), a powinno się je obliczyć na podstawie przekształconej macierzy danych $\mathbf{Y} = \left[y_{ij} \right]$ otrzymanej w kroku 6 algorytmu.

5. Podsumowanie

W artykule przetestowano przydatność pięciu indeksów oceny jakości klasyfikacji w zagadnieniu doboru liczby klas w klasyfikacji spektralnej uwzględniającej cztery typy odległości. W eksperymentach wykorzystano klasyczne oraz nieklasyczne dane metryczne o znanej strukturze klas obiektów.

W eksperymencie I najlepiej strukturę klas odkrywała klasyfikacja spektralna z kwadratem odległości euklidesowej oraz indeksami WC, DB, G1, KL, w eksperymencie II zaś klasyfikacja spektralna z odległością GDM1 oraz indeksami G1, WC, DB.

Przeprowadzone eksperymenty wykazały wysoką skuteczność indeksów oceny jakości klasyfikacji stosowanych w klasycznej analizie skupień w zastosowaniu do odkrywania liczby klas w klasyfikacji spektralnej.

Literatura

Everitt B.S., Landau S., Leese M., Stahl D. (2011), *Cluster Analysis*, Wiley, Chichester.

Girolami M. (2002), *Mercer kernel-based clustering in feature space*, „IEEE Transactions on Neural Networks", vol. 13, no. 3, pp. 780-784.

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

Karatzoglou A. (2006), *Kernel Methods. Software, Algorithms and Applications*, Rozprawa doktorska, Uniwersytet Techniczny we Wiedniu.

Kolupa M. (1976), *Elementarny wykład algebry liniowej dla ekonomistów*, Państwowe Wydawnictwo Naukowe, Warszawa.

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*, MIT Press, Cambridge, pp. 849-856.

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.

Shortreed S. (2006), *Learning in Spectral Clustering*, Rozprawa doktorska, University of Washington.

Verma D., Meila M. (2003), *A Comparison of Spectral Clustering Algorithms*, Technical report UW-CSE-03-05-01, University of Washington.

von Luxburg U. (2007), *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.

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, s. 185-203.

Walesiak M. (2009), *Analiza skupień*, [w:] M. Walesiak, E. Gatnar (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa, s. 407-433.

Walesiak M. (2011), *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R,* Wydawnictwo UE, Wrocław.

Walesiak M. (2012), *Klasyfikacja spektralna a skale pomiaru zmiennych,* „Przegląd Statystyczny" z. 1, s. 13-31.

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

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, s. 161-171.

Walesiak M., Dudek A. (2012), *clusterSim package*, URL http://www.R-project.org.

Wang J. (2010), *Consistent selection of the number of clusters via crossvalidation*, „Biometrika" vol. 97, issue 4, pp. 893-904.

AUTOMATIC DETERMINATION OF THE NUMBER OF CLUSTERS USING SPECTRAL CLUSTERING

Summary: The paper tested the usefulness of five indices assessing the quality of classification (within-group dispersion, Davies-Bouldin index, Caliński & Harabasz index, Hartigan index, Krzanowski & Lai index) in the issue of selection of the number of clusters in the spectral clustering taking into account four types of distance (squared Euclidean distance, Euclidean distance, Manhattan distance, GDM1 distance). The article evaluates twenty clustering procedures (four spectral clustering methods and five indices) based on two types of simulated data (classic and non-classic). Each clustering result was compared with the known cluster structure applying corrected Rand index.

Keywords: cluster analysis, spectral clustering, number of clusters.