PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107

Taksonomia 17 2010

Marek Walesiak, Andrzej Dudek
Uniwersytet Ekonomiczny we Wrocławiu

KLASYFIKACJA SPEKTRALNA Z WYKORZYSTANIEM ODLEGŁOŚCI GDM[1]

Streszczenie: W artykule zaproponowano modyfikację metody klasyfikacji spektralnej. W tym celu w procedurze tej metody (zob. [Ng, Jordan, Weiss 2002]) przy wyznaczaniu macierzy podobieństwa (*affinity matrix*) w konstrukcji estymatora jądrowego zastosowano odległość GDM1 przy klasyfikacji danych metrycznych oraz GDM2 przy klasyfikacji danych porządkowych. Ponadto przetestowano przydatność metod klasyfikacji spektralnej (w tym metody z odległością GDM) w porównaniu z klasycznymi metodami analizy skupień dla wygenerowanych danych o znanej strukturze klas, wykorzystując do oceny zgodności wyników klasyfikacji skorygowany indeks Randa (zob. [Hubert, Arabie 1985]).

1. Wstęp

Od końca XX wieku w literaturze poświęconej analizie danych rozwija się analiza skupień bazująca na dekompozycji spektralnej (*spectral clustering*). W artykule scharakteryzowane zostaną różne warianty klasyfikacji spektralnej. Zastosowana zostanie odległość GDM w konstrukcji estymatora jądrowego służącego do obliczenia macierzy podobieństw w klasyfikacji spektralnej. Pozwoli to wykorzystać tę metodę przy klasyfikacji danych zarówno metrycznych (GDM1), jak i porządkowych (GDM2). Przetestowana zostanie przydatność metod klasyfikacji spektralnej (w tym z wykorzystaniem odległości GDM) oraz klasycznych metod analizy skupień dla wygenerowanych danych o znanej strukturze klas. Analizę porównawczą metod klasyfikacji dla danych o znanej strukturze klas przeprowadzono dla trzech typów danych.

2. Miara odległości GDM

W pracy Walesiaka [2002] zaproponowano uogólnioną miarę odległości GDM (*The Generalised Distance Measure*), w konstrukcji której wykorzystano ideę uogólnionego współczynnika korelacji obejmującego współczynnik korelacji liniowej Pearsona i współczynnik tau Kendalla:

[1] Artykuł powstał w ramach działalności statutowej Katedry Ekonometrii i Informatyki (Marek Walesiak) oraz w ramach projektu badawczego MNiSW pt. „Obiekty symboliczne w wielowymiarowej analizie statystycznej" nr NN111 105234 (Andrzej Dudek).

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{2\left[\sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2\right]^{\frac{1}{2}}}, \; d_{ik} \in [0;1], \tag{1}$$

gdzie: d_{ik} – miara odległości GDM1 dla danych metrycznych i GDM2 dla danych porządkowych,
$i, k, l = 1, \ldots, n$ – numery obiektów,
$j = 1, \ldots, m$ – numer zmiennej.

Dla zmiennych mierzonych na skali ilorazowej i(lub) przedziałowej w formule (1) stosowane jest podstawienie (odległość `GDM1`):

$$\begin{array}{ll} a_{ipj} = x_{ij} - x_{pj} & \text{dla} \quad p = k,l \\ b_{krj} = x_{kj} - x_{rj} & \text{dla} \quad r = i,l \end{array}, \tag{2}$$

gdzie: x_{ij} (x_{kj}, x_{lj}) – i-ta (k-ta, l-ta) obserwacja na j-tej zmiennej.

Zasób informacji skali porządkowej jest nieporównanie mniejszy. Jedyną dopuszczalną operacją empiryczną na skali porządkowej jest zliczanie zdarzeń (tzn. wyznaczanie liczby relacji większości, mniejszości i równości). W konstrukcji miernika odległości musi być wykorzystana informacja o relacjach, w jakich pozostają porównywane obiekty w stosunku do pozostałych obiektów ze zbioru obiektów. Dla zmiennych mierzonych na skali porządkowej w formule (1) stosuje się podstawienie (odległość `GDM2` – zob. [Walesiak 1993, s. 44-45]):

$$a_{ipj}\left(b_{krj}\right) = \begin{cases} 1 & \text{dla} \quad x_{ij} > x_{pj} \; \left(x_{kj} > x_{rj}\right) \\ 0 & \text{dla} \quad x_{ij} = x_{pj} \; \left(x_{kj} = x_{rj}\right) \\ -1 & \text{dla} \quad x_{ij} < x_{pj} \; \left(x_{kj} < x_{rj}\right) \end{cases}, \text{ dla } p = k,\ l;\ r = i,\ l. \tag{3}$$

Własności oraz rezultaty badań symulacyjnych miary (1) zawiera m.in. praca Walesiaka [2006].

3. Procedura klasyfikacji spektralnej

W klasyfikacji spektralnej pierwotne dane z przestrzeni m-wymiarowej przekształcone zostają, przez wyznaczenie wektorów własnych macierzy Laplace'a, w zbiór danych o liczbie wymiarów odpowiadających liczbie klas u.

Procedura klasyfikacji spektralnej zaproponowana przez autorów, takich jak Ng, Jordan i Weiss [2002], obejmuje następujące etapy (zob. [Walesiak, Dudek 2009b]):

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. Macierz podobieństw $\mathbf{A} = [A_{ik}]$ ma następujące właściwości [Perona, Freeman 1998, s. 3]: $\underset{i,k}{\forall} A_{ik} \in [0;1]$, $A_{ii} = 1$, $A_{ik} = A_{ki}$. W prezentowanym algorytmie elementy z głównej przekątnej macierzy $\mathbf{A} = [A_{ik}]$ zastąpiono zerami ($A_{ii} = 0$).

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: $\mathbf{I} - \mathbf{L}$. Własności tej macierzy przedstawiono m.in. w pracy von Luxburg [2006, s. 5]. W algorytmie dla uproszczenia analizy pomija się macierz jednostkową $\mathbf{I}$.

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. Przeprowadza się normalizację 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 (proponuje się tutaj wykorzystanie metody *k*-średnich).

Istnieją odmiany analizy spektralnej różniące się:

a. Typem estymatora jądrowego w etapie 2. Zwykle wykorzystuje się tutaj estymator gaussowski bazujący na kwadracie odległości euklidesowej (zob. [Karatzoglou 2006, s. 26]):

$$A_{ik} = \exp(-\sigma \cdot d_{ik}^2), \; i, k = 1, \ldots, n, \tag{4}$$

gdzie: d_{ik} – odległość euklidesowa między obiektami *i* oraz *k*, σ – parametr skali (szerokość pasma – *kernel width*).

Inne estymatory jądrowe stosowane w klasyfikacji spektralnej zawarte są w pracach Karatzoglou [2006, s. 13-14] oraz Polanda i Zeugmanna [2006] i obejmują m.in.: jądro wielomianowe, jądro liniowe, jądro w postaci tangensa hiperbolicznego, jądro Bessela, jądro Laplace'a, jądro ANOVA, jądro łańcuchowe (dla danych tekstowych).

b. Formułą konstrukcji macierzy Laplace'a w etapie 3 (zob. np. [Verma, Meila 2003; von Luxburg 2006]):

– nienormalizowana macierz Laplace'a:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{5}$$

– znormalizowana macierz Laplace'a:

$$\mathbf{L} = \mathbf{D}^{-1}\mathbf{A}. \tag{6}$$

Dla tych konstrukcji macierzy Laplace'a procedura klasyfikacji spektralnej jest też inna (zob. [Shortreed 2006, s. 41-47]).

Zasadnicze znaczenie w klasyfikacji spektralnej mają dwa parametry: σ – oznaczający szerokość pasma (*kernel width*) oraz u – oznaczający liczbę skupień.

Parametr σ ma fundamentalne znaczenie w klasyfikacji spektralnej. W literaturze zaproponowano wiele heurystycznych sposobów wyznaczania wartości tego parametru (zob. np. prace: [Zelnik-Manor, Perona 2004; Fischer, Poland 2004; Poland, Zeugmann 2006]). W metodach heurystycznych wyznacza się wartość σ na podstawie pewnych statystyk opisowych macierzy odległości $[d_{ik}]$. Lepszy sposób wyznaczania parametru σ zaproponował Karatzoglou [2006]. Poszukuje się takiej wartości parametru σ, która minimalizuje wewnątrzklasową sumę kwadratów odległości przy zadanej liczbie klas u. Jest to heurystyczna metoda poszukiwania minimum lokalnego.

Zbliżony koncepcyjnie algorytm znajdowania optymalnego parametru σ zaproponowano w pracy Walesiaka i Dudka [2009b]. Z macierzy danych $\mathbf{X}$ (ze znormalizowanej macierzy danych – dla danych metrycznych) wybierana jest próba bootstrapowa $\mathbf{X}'$ składającą się z n' obiektów opisanych wszystkimi m zmiennymi. Wartość n' jest najczęściej tak dobierana, aby $\frac{1}{2}n \le n' \le \frac{3}{4}n$.

Początkowy przedział przeszukiwania optymalnej wartości parametru σ ustalany jest jako $S_0 = [0; D]$ (gdzie D oznacza sumę odległości d_{ik} w macierzy odległości). Dalsza procedura iteracyjna jest następująca:

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

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 suma odległości wewnątrzklasowych jest minimalna.

Krok 4. Jeśli dla wybranego σ_r^k zachodzi nierówność $p_r^k \le \varphi$ (domyślnie przyjęto $\varphi = 10^{-3}$), algorytm kończy działanie. W przeciwnym przypadku przechodzi się z wybranym przedziałem do kroku 1 i kontynuuje procedurę.

Podobnie jak w przypadku klasycznych metod klasyfikacji zachodzi potrzeba ustalenia optymalnej liczby klas. Algorytm wyznaczenia optymalnej liczby klas zaproponował Girolami [2002].

Macierz podobieństw (*affinity matrix*) $\mathbf{A} = [A_{ik}]$ 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 **K** jest porządkowany malejąco, a liczba jego dominujących elementów (wyznaczona np. przez kryterium osypiska) wyznacza optymalną liczbę skupień *u*, na którą algorytm klasyfikacji spektralnej powinien podzielić zbiór badanych obiektów.

4. Propozycja procedury klasyfikacji spektralnej z miarą odległości GDM

W artykule proponowana jest modyfikacja metody klasyfikacji spektralnej umożliwiająca jej zastosowanie w klasyfikacji danych metrycznych oraz porządkowych. W tym celu w kroku 2 procedury w konstrukcji estymatora jądrowego zastosowano odległość GDM:

$$A_{ik} = \exp(-\sigma \cdot d_{ik}), \tag{7}$$

gdzie: σ – parametr skali (szerokość pasma – *kernel width*),
d_{ik} – odległość GDM między obiektami *i* oraz *k* dla danych metrycznych o postaci (1) z podstawieniem (2) oraz dla danych porządkowych o postaci (1) z podstawieniem 3.

Zastosowanie odległości GDM o postaci (1) z podstawieniem (3) w konstrukcji estymatora jądrowego umożliwia analizę danych porządkowych w klasyfikacji spektralnej (zob. [Walesiak, Dudek 2009b]). Dane pierwotne $\mathbf{X} = [x_{ij}]$ mierzone są na skali porządkowej. W wyniku zastosowania estymatora jądrowego z odległością GDM2 podobieństwa w macierzy $\mathbf{A} = [A_{ik}]$ mierzone są na skali przedziało-

wej. Ostatecznie w kroku 5 otrzymuje się metryczną macierz danych $\mathbf{Y}$ o wymiarach $n \times u$. Pozwala to na zastosowanie w klasyfikacji dowolnych metod analizy skupień (w tym metod bazujących bezpośrednio na macierzy danych, np. metody *k*-średnich).

5. Analiza porównawcza metod klasyfikacji dla danych o znanej strukturze klas

Analizę porównawczą metod klasyfikacji dla danych o znanej strukturze klas przeprowadzono dla trzech typów danych.

W dwóch pierwszych eksperymentach wykorzystano dane metryczne oraz porządkowe o znanej strukturze klas obiektów wygenerowane z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` (zob. [Walesiak, Dudek 2009a]). Charakterystykę czterech modeli wykorzystanych w analizie symulacyjnej prezentuje tab. 1.

Tabela 1. Charakterystyka modeli w analizie symulacyjnej

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

* tylko dla danych porządkowych;

v – liczba zmiennych, *nk* – liczba kategorii (jedna liczba oznacza stałą liczbę kategorii); *cl* – liczba klas; *lo* – liczba obiektów w klasach (jedna liczba oznacza klasy równoliczne); *ks* – kształt skupień (1 – skupienia wydłużone, 2 – skupienia wydłużone i słabo separowalne, 3 – skupienia normalne, 4 – skupienia zróżnicowane dla klas).

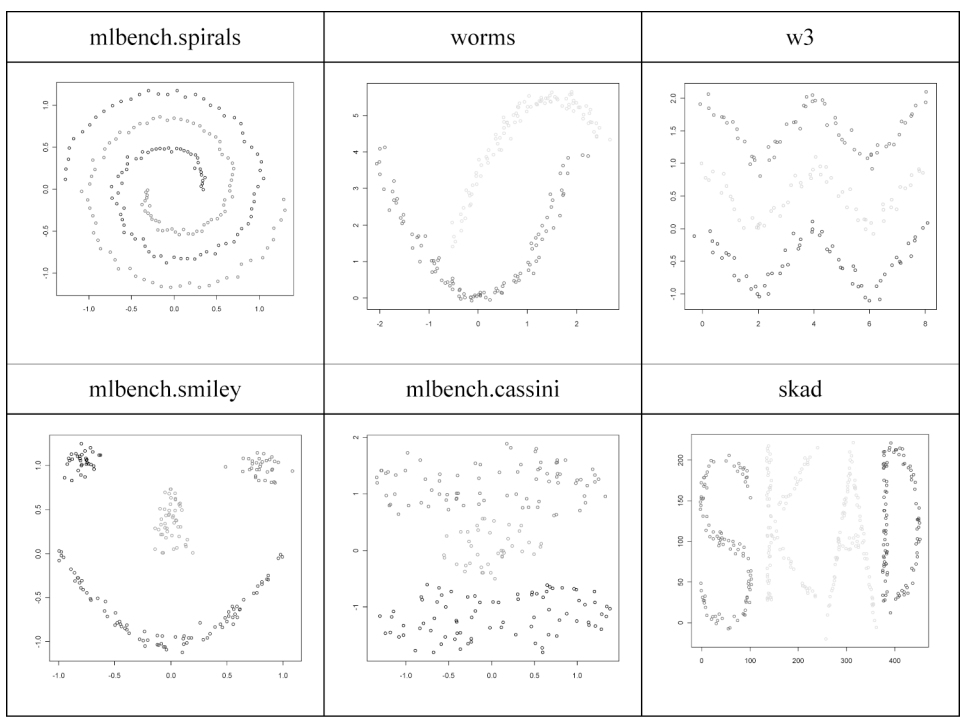
Źródło: opracowanie własne.

W eksperymencie trzecim zbiory danych (zob. rys. 1) utworzono z wykorzystaniem funkcji pakietu mlbench (spirals, smiley, cassini) oraz zbiorów własnych (worms, w3, skad).

Dla modeli w każdym eksperymencie wygenerowano 20 zbiorów danych, przeprowadzono procedurę klasyfikacyjną i porównano otrzymane rezultaty klasyfikacji ze znaną strukturą klas za pomocą skorygowanego indeksu Randa [Hubert, Arabie 1985].

Dla danych metrycznych (eksperymenty 1 i 3) uwzględniono następujące metody klasyfikacji: 1. `specc1` – klasyfikacja spektralna z jądrem gaussowskim i σ z pakietu `kernlab`; 2. `specc2` – klasyfikacja spektralna z jądrem gaussowskim i σ z artykułu; 3. `speccGDM1` – klasyfikacja spektralna z odległością `GDM1` i σ z artykułu; 4. `kmeans` – metoda *k*-średnich; 5. `pam` – metoda *k*-medoidów; 6. `complete` – metoda kompletnego połączenia; 7. `average` – metoda średniej klasowej; 8. `ward` – metoda Warda; 9. `centroid` – metoda środka ciężkości; 10. `diana` – hierarchiczna metoda deglomeracyjna.

Dla metod o numerach 5-10 zastosowano odległość `GDM1` oraz kwadrat odległości euklidesowej. Dla danych porządkowych (eksperyment 2) uwzględniono w analizie metody klasyfikacji o numerach 5-10 z odległością `GDM2` oraz klasyfikację spektralną z odległością `GDM2` i σ z artykułu (`speccGDM2`).



Rys. 1. Przykładowe zbiory danych utworzone z wykorzystaniem funkcji pakietu mlbench (spirals, smiley, cassini) oraz zbiorów własnych (worms, w3, skad)

Źródło: opracowanie własne z wykorzystaniem programu R.

Tabela 2 prezentuje uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 20 symulacji dla danych metrycznych wygenerowanych w pakiecie clusterSim.

W przypadku zbiorów danych metrycznych bez zmiennych zakłócających metody klasyfikacji spektralnej, z pewnymi wyjątkami, dają gorsze rezultaty od klasycznych metod analizy skupień. Uwzględnienie zmiennych zakłócających (występujących zwykle w rzeczywistych problemach klasyfikacyjnych) pokazuje wyraźną przewagę metod klasyfikacji spektralnej w odkrywaniu rzeczywistej struktury klas. Proponowana metoda `speccGDM1` daje zbliżone, choć nieco gorsze, rezultaty do metody klasyfikacji spektralnej z jądrem gaussowskim.

Tabela 2. Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych metrycznych wygenerowanych w pakiecie clusterSim

Metoda	Średnia ($k7$+$k8$+$k9$)/3		Kształt skupień								Liczba zmiennych zakłócających					
			1		2		3		4		0		1		2	
1	*2*		*3*		*4*		*5*		*6*		*7*		*8*		*9*	
specc2	0,683	1	0,928	6/7	0,706	8	0,751	7	0,924	7/8	0,827	5/6	0,735	1	0,487	2
specc1	0,681	2	0,817	9	0,686	9	0,735	8	0,942	6/7	0,795	8/9	0,723	2	0,524	1
speccGDM1	0,661	3	0,906	7/8	0,711	7/6	0,717	9	0,866	8/10	0,800	7/8	0,716	3	0,467	3
average[a]	0,561	4	0,950	1	0,831	1	0,800	3	0,967	4	0,887	2	0,439	6	0,356	4
average[b]	0,568	4	0,950	1	0,832	2	0,800	3	0,979	1	0,890	1	0,465	5	0,350	4
pam[a]	0,558	5	0,950	1	0,830	2	0,800	1	0,942	5	0,881	4	0,472	4	0,322	5
pam[b]	0,558	5	0,950	1	0,830	3	0,800	1	0,942	6	0,881	3	0,472	4	0,322	5
ward[a]	0,550	6	0,950	1	0,828	3	0,800	4	0,973	2	0,888	1	0,443	5	0,320	6
ward[b]	0,551	6	0,950	1	0,836	1	0,800	4	0,972	3	0,889	2	0,443	6	0,319	6
centroid[a]	0,526	7	0,950	1	0,820	4	0,800	5	0,971	3	0,885	3	0,426	7	0,267	7
centroid[b]	0,385	10	0,950	1	0,782	4	0,800	5	0,968	4	0,875	4	0,269	9	0,012	10
diana[a]	0,463	8	0,930	5	0,646	10	0,798	6	0,575	10	0,737	10	0,391	8	0,260	8
diana[b]	0,512	7	0,950	1	0,644	10	0,796	6	0,901	9	0,823	7	0,418	7	0,296	7
kmeans	0,452	9/8	0,784	10	0,760	5	0,633	10	0,978	1/2	0,789	9/10	0,371	9/8	0,195	9/8
complete[a]	0,415	10	0,858	8	0,747	6	0,800	2	0,862	9	0,817	6	0,277	10	0,150	10
complete[b]	0,399	9	0,950	1	0,707	7	0,800	2	0,952	5	0,852	5	0,245	10	0,101	9

a – z odległością GDM1; b – z kwadratem odległości euklidesowej.

6/7 – pozycja metody, gdy dla klasycznych metod analizy skupień stosujemy odległość GDM1/kwadrat odległości euklidesowej.

Źródło: obliczenia własne z wykorzystaniem programu R.

Uwzględnienie dla klasycznych metod analizy skupień odległości `GDM1` oraz kwadratu odległości euklidesowej daje zbliżone rezultaty, jeśli chodzi o stopień odkrywania rzeczywistej struktury klas.

Tabela 3 prezentuje uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 20 symulacji dla danych porządkowych wygenerowanych w pakiecie clusterSim.

W przypadku zbiorów danych porządkowych bez zmiennych zakłócających najlepsza jest metoda Warda. Metoda klasyfikacji spektralnej `speccGDM2` daje gorsze rezultaty od klasycznych metod analizy skupień. Należy jednak pamiętać, że zbiory tego typu bardzo rzadko występują w rzeczywistych problemach klasyfi-

kacyjnych. Uwzględnienie zmiennych zakłócających pokazuje wyraźną przewagę metody klasyfikacji spektralnej `speccGDM2`.

Tabela 3. Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych porządkowych wygenerowanych w pakiecie clusterSim

Metoda	Średnia (*k7*+*k8*+*k9*)/3		Kształt skupień								Liczba zmiennych zakłócających					
			1		2		3		4		0		1		2	
1	*2*		*3*		*4*		*5*		*6*		*7*		*8*		*9*	
speccGDM2	0,644	1	0,901	7	0,754	7	0,793	7	0,612	6	0,765	7	0,659	1	0,510	1
average	0,599	2	1,000	1	0,974	1	1,000	1	0,947	2	0,980	2	0,477	3	0,339	2
pam	0,591	3	1,000	1	0,969	3	1,000	1	0,933	4	0,975	4	0,480	2	0,318	3
ward	0,591	4	1,000	1	0,967	4	1,000	1	0,963	1	0,982	1	0,473	4	0,317	4
centroid	0,562	5	1,000	1	0,973	2	1,000	1	0,946	3	0,980	2	0,431	5	0,274	5
diana	0,496	6	0,956	5	0,770	6	0,998	6	0,565	7	0,822	6	0,418	6	0,249	6
complete	0,461	7	0,924	6	0,893	5	1,000	1	0,909	5	0,931	5	0,296	7	0,155	7

Źródło: obliczenia własne z wykorzystaniem programu R.

Tabela 4 prezentuje uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 20 symulacji dla danych metrycznych z pakietu mlbench i danych własnych.

Tabela 4. Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych metrycznych z pakietu mlbench i danych własnych

Metoda	Średnia		Zbiory danych											
			spirals		worms		w3		smiley		cassini		skad	
specc1	0,796	1	0,830	3	0,795	2	0,840	1	0,837	2/3	0,759	6/5	0,715	3
specc2	0,792	2	0,866	2	0,847	1	0,720	2	0,797	3/5	0,754	7/6	0,767	1
speccGDM1	0,715	3	0,957	1	0,537	3	0,406	3	0,870	1/2	0,796	5/3	0,722	2
ward[a]	0,397	6	0,042	6	0,411	8	0,003	7	0,646	5	0,935	2	0,348	9
ward[b]	0,467	4	0,028	8	0,361	10	0,006	4	0,950	1	0,844	2	0,611	5
pam[a]	0,424	4	0,011	10	0,448	6	–0,005	8	0,794	4	0,919	3	0,374	8
pam[b]	0,424	5	0,011	10	0,448	7	–0,005	8	0,794	6	0,919	1	0,374	10
average[a]	0,411	5	0,026	9	0,393	10	0,003	6	0,605	8	0,981	1	0,455	7
average[b]	0,393	6	0,029	7	0,432	8	–0,003	7	0,637	7	0,783	4	0,482	8
centroid[a]	0,396	7	0,044	5	0,423	7	0,005	5	0,611	7	0,827	4	0,466	6
centroid[b]	0,389	7	0,016	9	0,466	5	–0,002	6	0,825	4	0,556	9	0,473	9
diana[a]	0,305	10	0,037	7	0,452	5	–0,006	9	0,486	10	0,522	10	0,341	10
diana[b]	0,386	8	0,040	4	0,467	4	–0,009	9	0,627	8	0,539	10	0,651	4
kmeans	0,369	9	0,031	8/6	0,455	4/6	–0,009	10	0,623	6/9	0,595	9/7	0,519	4/6
complete[a]	0,370	8	0,045	4	0,400	9	0,010	4	0,568	9	0,720	8	0,475	5
complete[b]	0,353	10	0,037	5	0,424	9	0,002	5	0,587	10	0,564	8	0,505	7

a – z odległością GDM1; b – z kwadratem odległości euklidesowej.

8/6 – pozycja metody, gdy dla klasycznych metod analizy skupień stosujemy odległość GDM1/kwadrat odległości euklidesowej.

Źródło: obliczenia własne z wykorzystaniem programu R.

Dla nietypowych zbiorów danych metody klasyfikacji spektralnej zdecydowanie lepiej od klasycznych metod analizy skupień odkrywają prawidłową strukturę klas. Proponowana metoda `speccGDM1` daje zbliżone rezultaty do metody klasyfikacji spektralnej z jądrem gaussowskim.

Literatura

Fischer I., Poland J., *New Methods for Spectral Clustering*, Technical Report No. IDSIA-12-04, Dalle Molle Institute for Artificial Intelligence, Manno-Lugano, Switzerland 2004.

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

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

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

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, 849-856.

Perona P., Freeman W.T., *A Factorization Approach to Grouping*, Lecture Notes In Computer Science, vol. 1406, Proceedings of the 5th European Conference on Computer Vision, volume I, 1998, 655-670.

Poland J., Zeugmann T., *Clustering the Google Distance with Eigenvectors and Semidefinite Programming. Knowledge Media Technologies, First International Core-To-Core Workshop,* Dagstuhl, July 23-27, Germany 2006 (Klaus P., Jantke & Gunther Kreuzberger (red.), *Diskussionsbeiträge*, Institut für Medien und Kommunikationswisschaft, Technische Universität Ilmenau, no 21, July 2006).

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

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

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

Walesiak M., Dudek A., *`clusterSim` package*, URL http://www.R-project.org, 2009a.

Walesiak M., Dudek A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, [w:] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 84 (w druku), Wrocław 2009b.

Walesiak M., *Propozycja uogólnionej miary odległości w statystycznej analizie wielowymiarowej*, [w:] J. Paradysz (red.), *Statystyka regionalna w służbie samorządu lokalnego i biznesu*, Internetowa Oficyna Wydawnicza, Centrum Statystyki Regionalnej, Akademia Ekonomiczna w Poznaniu, Poznań 2002, s. 115-121.

Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654, Seria: Monografie i Opracowania nr 101, AE, Wrocław 1993.

Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydanie drugie rozszerzone, AE, Wrocław 2006.

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

SPECTRAL CLUSTERING WITH THE USE OF GDM DISTANCE

Summary: In the article, the proposal of spectral clustering method, based on procedure of Ng, Jordan and Weiss [2002], is presented. In construction of affinity matrix we implement kernel function with GDM1 distance for classification of metric data and GDM2 distance for classification of ordinal data. The article evaluates, based on three types of data simulated, ten clustering methods (three spectral clustering methods, seven classical clustering methods). Each clustering result was compared with known cluster structure from models applying Hubert and Arabie's [1985] corrected Rand index.