PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 84

Ekonometria 27 2009

Marek Walesiak, Andrzej Dudek
Uniwersytet Ekonomiczny we Wrocławiu

ODLEGŁOŚĆ GDM DLA DANYCH PORZĄDKOWYCH A KLASYFIKACJA SPEKTRALNA

Streszczenie: W artykule zaproponowano modyfikację metody klasyfikacji spektralnej umożliwiającą jej zastosowanie w klasyfikacji danych porządkowych. W tym celu w procedurze tej metody przy wyznaczaniu macierzy podobieństwa (*affinity matrix*) w konstrukcji estymatora jądrowego zastosowano odległość GDM dla danych porządkowych (odległość GDM2). Ponadto zaproponowano metodę ustalania parametru σ (szerokość pasma – *kernel width*) mającego zasadnicze znaczenie w klasyfikacji spektralnej. W części empirycznej artykułu pokazano zastosowanie klasyfikacji spektralnej w odniesieniu do wybranych struktur danych porządkowych.

Słowa kluczowe: klasyfikacja spektralna, dane porządkowe, odległość GDM2.

1. Wstęp

Od końca XX w. w literaturze poświęconej analizie danych rozwija się analiza skupień bazująca na dekompozycji spektralnej (*spectral clustering*). W artykule zaproponowano modyfikację metody klasyfikacji spektralnej umożliwiającą jej zastosowanie w klasyfikacji danych porządkowych. W tym celu w procedurze tej metody przy wyznaczaniu macierzy podobieństwa (*affinity matrix*) w konstrukcji estymatora jądrowego zastosowano odległość GDM dla danych porządkowych (odległość GDM2). Zaproponowano również metodę ustalania parametru σ (szerokość pasma – *kernel width*) mającego duże znaczenie w klasyfikacji spektralnej. W części empirycznej tekstu ukazano zastosowanie klasyfikacji spektralnej wybranych struktur danych porządkowych.

2. Analiza skupień dla danych porządkowych

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru, tj. nominalną, porządkową (rangową), przedziałową (interwałową), ilorazową (stosunkową). Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową − do niemetrycznych. Skale pomiaru są uporządkowane od najsłabszej (nominalna) do najmocniejszej (ilorazowa). Z typem skali wiąże się grupa prze-

kształceń, ze względu na które skala zachowuje swe własności. Na skali porządkowej dozwolonym przekształceniem matematycznym dla obserwacji jest dowolna ściśle monotonicznie rosnąca funkcja, która nie zmienia dopuszczalnych relacji, tj. równości, różności, większości i mniejszości.

Zasób informacji skali porządkowej jest nieporównanie mniejszy niż skal metrycznych. Jedyną dopuszczalną operacją empiryczną na skali porządkowej jest zliczanie zdarzeń (tzn. wyznaczanie liczby relacji większości, mniejszości i równości). Szczegółową charakterystykę skal pomiaru zawierają m.in. prace: [Walesiak 1996, s. 19-24; Walesiak 2006, s. 12-15].

Typowa procedura analizy skupień w odniesieniu do danych porządkowych obejmuje następujące etapy (por. [Milligan 1996, s. 342-343; Walesiak 2005; Walesiak 2009]): wybór obiektów i zmiennych, wybór miary odległości, wybór metody klasyfikacji, ustalenie liczby klas, ocenę wyników klasyfikacji, opis (interpretację) i profilowanie klas.

W stosunku do procedury analizy skupień dla danych metrycznych nie występuje tutaj etap normalizacji wartości zmiennych. Normalizacja polega na pozbawieniu wartości zmiennych mian i ujednoliceniu rzędów wielkości w celu doprowadzenia ich do porównywalności. W odniesieniu do danych porządkowych nie zachodzi potrzeba normalizacji, ponieważ są one niemianowane. Ponadto rząd wielkości obserwacji zmiennej porządkowej nie ma znaczenia, gdyż między obserwacjami wyznacza się tylko relacje równości, różności, większości i mniejszości. Miara odległości (zob. etap 2 procedury analizy skupień) dla obiektów opisanych zmiennymi porządkowymi może wykorzystywać w swojej konstrukcji tylko wymienione uprzednio relacje. To ograniczenie powoduje, że musi być ona miarą kontekstową, która wykorzystuje informacje o relacjach, w jakich pozostają porównywane obiekty w stosunku do pozostałych obiektów z badanego zbioru obiektów. Taką miarą odległości dla danych porządkowych jest miara GDM zaproponowana przez Walesiaka [Walesiak 1993, s. 44-45]:

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

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

– i-ta (k-ta, l-ta) obserwacja na j-tej zmiennej, $i, k, l = 1, \ldots, n$ – numery obiektów, $j = 1, \ldots, m$ – numer zmiennej.

W literaturze z zakresu statystycznej analizy wielowymiarowej nie zaproponowano dotychczas innych miar odległości dla zmiennych porządkowych. Miara odległości Kendalla [Kendall 1966, s. 181], odległość Gordona [Gordon1999, s. 19] czy odległość Podaniego [Podani 1999] nie są typowymi miarami dla zmiennych porządkowych, ponieważ przy ich stosowaniu zakłada się, że odległości między sąsiednimi obserwacjami na skali porządkowej są sobie równe (na skali porządkowej odległości między dowolnymi dwiema obserwacjami nie są znane). Zastosowanie tych miar odległości wymaga uprzedniego porangowania obserwacji. Przyjmuje się wtedy upraszczające założenie, że rangi są mierzone co najmniej na skali przedziałowej (wtedy dopuszcza się wyznaczanie różnic między wartościami skali).

W przedstawionej procedurze analizy skupień dla danych porządkowych wśród metod klasyfikacji można wykorzystać tylko te bazujące na macierzy odległości: metodę *k*-medoidów, metody klasyfikacji hierarchicznej (pojedynczego połączenia, kompletnego połączenia, średniej klasowej, ważonej średniej klasowej, Warda, środka ciężkości, medianowa), hierarchiczną metodę deglomeracyjną Macnaughtona-Smitha i in. [Macnaughton-Smith i in. 1964]. Nie jest możliwe zastosowanie tutaj metod bezpośrednio bazujących na macierzy danych ze względu na ich porządkowy charakter.

3. Procedura klasyfikacji spektralnej

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

Typowa procedura klasyfikacji spektralnej obejmuje takie kroki, jak (por. [Ng, Jordan, Weiss 2002]):

1. Konstrukcja macierzy danych $\mathbf{X} = [x_{ij}]$ o wymiarach $n \times m$ ($i = 1, \ldots, n$ – numer obiektu, $j = 1, ..., m$ – 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. Obliczenie diagonalnej macierzy wag **D** (na głównej przekątnej znajdują się sumy każdego wiersza z macierzy $\mathbf{A} = [A_{ik}]$, a poza główną przekątną są zera).

4. Konstrukcja znormalizowanej macierzy Laplace'a $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. 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 w celu uproszczenia analizy pominięto macierz jednostkową **I**.

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

6. 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.

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

Istnieją odmiany analizy spektralnej różniące się:

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

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

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 m.in. w pracach: [Karatzoglou 2006, s. 13-14; Poland, Zeugmann 2006].

- Formułą konstrukcji macierzy Laplace'a w kroku 4 (zob. np. [Verma, Meila 2003; von Luxburg 2006]). W tym przypadku procedura klasyfikacji spektralnej jest też inna (zob. [Shortreed 2006, s. 41-47]).

Zasadnicze znaczenie w klasyfikacji spektralnej mają dwa parametry: σ – 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 [Karatzoglou 2006]. Polega ona na poszukiwaniu 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 proponowany jest w niniejszym artykule. Z macierzy danych **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 przeprowadzania 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 $p_r^k \le \phi$ (domyślnie przyjęto $\phi = 10^{-3}$), to algorytm kończy działanie. W przeciwnym razie przechodzi się z wybranym przedziałem do kroku 1 i kontynuuje się procedurę.

Podobnie jak w przypadku klasycznych metod klasyfikacji zachodzi potrzeba ustalenia optymalnej liczby klas. Algorytm wyznaczenia optymalnej liczby klas zaproponował Girolami [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 $k_i = \lambda_i \{\mathbf{1}_n^T \mathbf{u}_i\}^2$ ($\mathbf{1}_n^T$ – 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.

4. Propozycja procedury klasyfikacji spektralnej dla danych porządkowych

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

$$A_{ik} = \exp(-\sigma \cdot d_{ik}), \tag{3}$$

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

Zaletą takiego podejścia jest to, że w wyniku zastosowania odległości GDM w konstrukcji estymatora jądrowego możliwe jest zastosowanie klasyfikacji spektralnej w analizie danych porządkowych. Dane pierwotne $\mathbf{X} = [x_{ij}]$ mierzone są na skali porządkowej. W wyniku zastosowania estymatora jądrowego o postaci (3) podobieństwa w macierzy $\mathbf{A} = [A_{ik}]$ mierzone są na skali przedziałowej. Ostatecznie w kroku 6 otrzymuje się metryczną macierz danych **Y** o wymiarach $n \times u$. Pozwala to na zastosowanie w klasyfikacji dowolnych metod analizy skupień (w tym metody *k*-średnich).

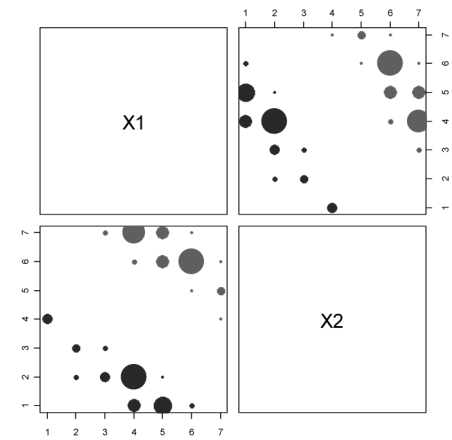
5. Przykłady klasyfikacji spektralnej wybranych struktur danych porządkowych

Przykład 1. Do wygenerowania danych porządkowych wykorzystano model 3 funkcji `cluster.Gen` z pakietu `clusterSim` (zob. [Walesiak, Dudek 2009]). Za pomocą dwuwymiarowej zmiennej losowej o rozkładzie normalnym wygenerowano po 40 obserwacji dla dwóch skupień o wydłużonym kształcie. Przyjęto następujące wektory wartości oczekiwanych dla skupień (0; 0), (1; 5) oraz identyczne macierze kowariancji Σ ($\sigma_{jj} = 1$, $\sigma_{jh} = -0,9$). Następnie przeprowadzono proces dyskretyzacji. Po zastosowaniu tej procedury otrzymuje się graficzną prezentację wygenerowanych danych porządkowych w układzie dwóch klas w przestrzeni dwuwymiarowej (zob. rys. 1a).

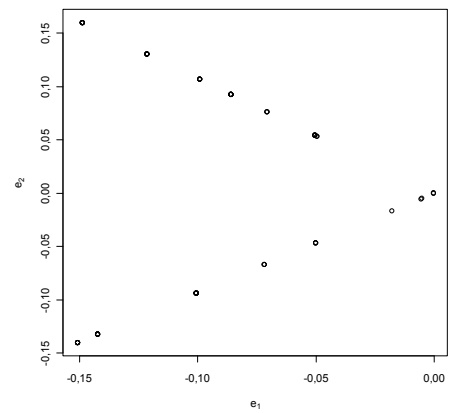
Do klasyfikacji zbioru 80 obiektów opisanych dwiema zmiennymi mierzonymi na skali porządkowej zastosowano metodę klasyfikacji spektralnej dla danych porządkowych, wyznaczając w kroku 2 macierz podobieństwa zgodnie ze wzorem (3). Rysunki 1b i 1c prezentują odpowiednio obiekty z macierzy **E** o wymiarach 80×2 (krok 5) oraz obiekty ze znormalizowanej macierzy $\mathbf{Y} = [y_{ij}]$ o wymiarach 80×2 (krok 6). W przypadku dobrze separowalnych skupień współrzędne obiektów z tych samych skupień w macierzy **Y** (krok 6) są takie same.

Rysunek 1d prezentuje uporządkowane składowe wektora **K** w metodzie Girolamiego służącej do ustalenia optymalnej liczby klas. Rysunek ten wskazuje dwa dominujące elementy tego wektora, więc zbiór obiektów należy podzielić na dwie klasy.

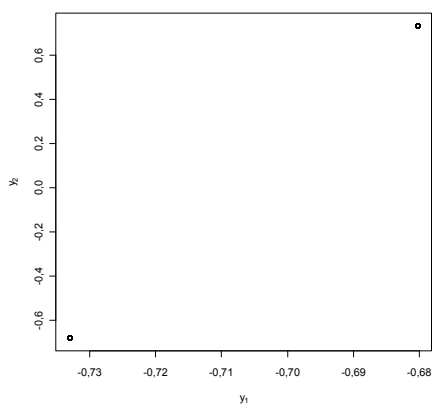
Macierz **Y** stanowi punkt wyjścia zastosowania metody *k*-średnich do podziału zbioru 80 obiektów na dwie klasy. Następnie obliczono skorygowany indeks Randa między wynikiem podziału zbioru obiektów metodą klasyfikacji spektralnej a znaną strukturą klas, który wynosi w tym przypadku 1.



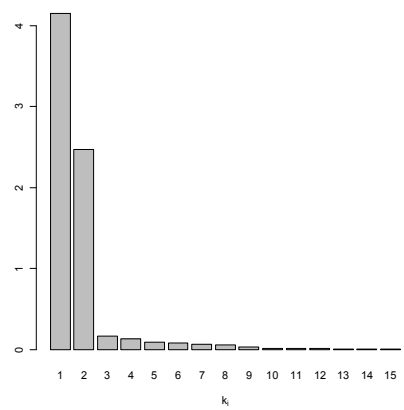
a) graficzna prezentacja zbioru danych porządkowych



b) zbiór danych w przestrzeni dwóch wektorów własnych macierzy Laplace'a



c) zbiór danych w przestrzeni dwóch wektorów własnych macierzy Laplace'a po normalizacji



d) uporządkowane składowe wektora **K** w metodzie Girolamiego służącej do ustalenia optymalnej liczby klas

Rys. 1. Wybrane etapy klasyfikacji spektralnej dla przykładowego zbioru danych porządkowych wygenerowanego z wykorzystaniem funkcji `cluster.Gen` z pakietu `clusterSim`

Źródło: opracowanie własne.

Przykład 2. Przeprowadzono klasyfikację spektralną, wyznaczając w kroku 2 macierz podobieństwa zgodnie ze wzorem (3), 27 nieruchomości lokalowych na jeleniogórskim rynku nieruchomości opisanych 6 zmiennymi (zob. tab. 1). Nieruchomość 1 jest wyceniana, natomiast nieruchomości od 2 do 27 to nieruchomości porównywalne, dla których znane są ceny transakcyjne.

Tabela 1. Macierz danych (27 nieruchomości opisanych 6 zmiennymi)

Nr nieruchomości	x_1	x_2	x_3	x_4	x_5	x_6
1	1	2	3	3	1	3
2	3	2	1	3	2	2
3	1	1	1	4	1	1
4	4	2	3	3	2	3
5	1	1	2	4	1	2
6	2	2	2	3	1	3
7	3	1	1	3	2	2
8	2	1	1	4	1	1
9	1	2	2	4	1	2
10	2	3	3	3	1	3
11	1	1	1	4	1	1
12	2	2	3	4	1	2
13	2	1	1	3	1	1
14	2	1	1	3	2	3
15	1	1	2	3	2	1
16	3	2	2	3	1	1
17	2	3	3	3	2	3
18	2	4	2	4	1	2
19	3	2	2	3	2	1
20	3	3	3	3	1	3
21	2	2	2	3	1	1
22	1	2	2	4	1	2
23	1	1	1	4	1	2
24	2	3	2	3	1	2
25	3	3	3	2	2	3
26	3	2	3	1	2	3
27	4	2	3	1	2	3

Źródło: [Pawlukowicz 2006, s. 238].

Mieszkalne nieruchomości lokalowe zostały opisane takimi zmiennymi, jak (zob. [Pawlukowicz 2006, s. 238]):

1. Lokalizacja środowiskowa nieruchomości gruntowej, z którą związany jest lokal mieszkalny (1 – bardzo dobra, 2 – dobra, 3 – dostateczna, 4 – nieodpowiednia, 5 – zła).

2. Standard użytkowy lokalu mieszkalnego (1 – wysoki, 2 – średni, 3 – niski, 4 – zły).

3. Warunki bytowe występujące na nieruchomości gruntowej, z którą związany jest lokal mieszkalny (1 – dobre, 2 – przeciętne, 3 – złe).

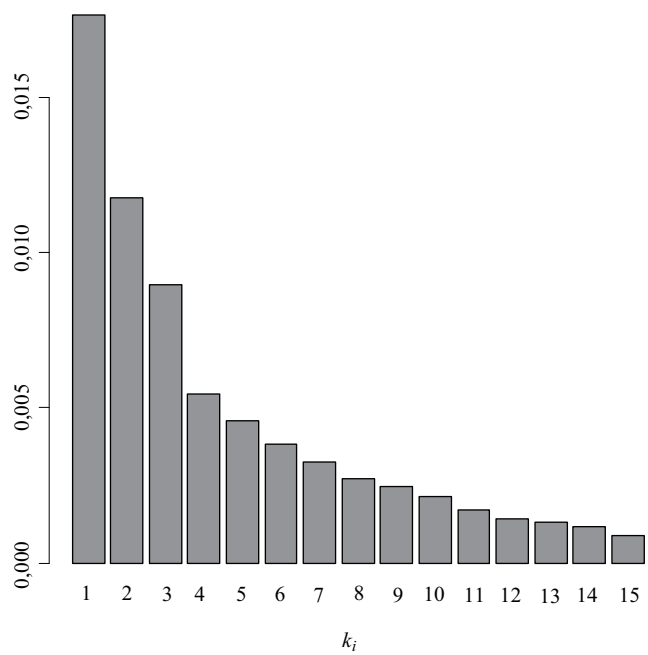
4. Położenie nieruchomości gruntowej, z którą związany jest lokal mieszkalny, w strefie miasta (1 – centralna, 2 – śródmiejska, 3 – pośrednia, 4 – peryferyjna).

5. Typ wspólnoty mieszkaniowej (1 – mała, 2 – duża).

6. Powierzchnia nieruchomości gruntowej, z którą związany jest lokal mieszkalny (1 – korzystna, 2 – akceptowana, 3 – niekorzystna).

Na podstawie danych z tab. 1 przeprowadzono klasyfikację spektralną 27 nieruchomości lokalowych na jeleniogórskim rynku nieruchomości opisanych 6 zmiennymi.

Rysunek 2 wskazuje trzy dominujące elementy tego wektora **K** w metodzie Girolamiego, więc zbiór obiektów należy podzielić na trzy klasy.



Rys. 2. Uporządkowane składowe wektora **K** w metodzie Girolamiego służącej do ustalenia optymalnej liczby klas

Źródło: opracowanie własne z wykorzystaniem programu R.

Ostatecznie otrzymano podział zbioru 27 nieruchomości na trzy klasy:

Lp.	Nr nieruchomości	Klasa	Lp.	Nr nieruchomości	Klasa
1	1	1	19	3	3

2	4	1	20	5	3
3	10	1	21	8	3
4	17	1	22	9	3
5	20	1	23	11	3
6	25	1	24	12	3
7	26	1	25	13	3
8	27	1	26	22	3
9	2	2	27	23	3
10	6	2			
11	7	2			
12	14	2			
13	15	2			
14	16	2			
15	18	2			
16	19	2			
17	21	2			
18	24	2			

W celu ułatwienia interpretacji wyników klasyfikacji spektralnej dla zmiennych z poszczególnych klas obliczono dominanty:

Klasa	Zmienna					
	x_1	x_2	x_3	x_4	x_5	x_6
1	3	NA	3	3	2	3
2	2	2	2	3	NA	NA
3	1	1	1	4	1	2

Nieruchomość wyceniana znalazła się w pierwszej klasie, zatem do jej wyceny należy wykorzystać dane z pozostałych nieruchomości w tej klasie (są to nieruchomości o numerach: 4, 10, 17, 20, 25, 26, 27).

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, s. 780-784.

Gordon A.D., *Classification*, Chapman & Hall/CRC, London 1999.

Karatzoglou A., *Kernel Methods. Software, Algorithms and Applications*, rozprawa doktorska, Uniwersytet Techniczny we Wiedniu, Wiedeń 2006.

Kendall M.G., *Discrimination and classification*, [w:] P.R. Krishnaiah (red.), *Multivariate Analysis I*, Academic Press, New York, London 1966, s. 165-185.

Macnaughton-Smith P., Williams W.T., Dale M.B., Mockett L.G., *Dissimilarity analysis: a new technique of hierarchical sub-division*, „Nature" 1964, 202, s. 1034-1035.

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.

Ng A., Jordan M., Weiss Y., *On spectral clustering: analysis and an algorithm*, [w:] T. Dietterich, S. Becker, Z. Ghahramani (red.), *Advances in Neural Information Processing Systems 14*, MIT Press, 2002, s. 849-856.

Pawlukowicz R., *Klasyfikacja w wyborze nieruchomości podobnych dla potrzeb wyceny rynkowej nieruchomości*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1100, Ekonometria 16, AE, Wrocław 2006, s. 232-240.

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, 1998, vol. I, s. 655-670.

Podani J., *Extending gowers general coefficient of similarity to ordinal characters*, „Taxon", 1999, 48, s. 331-340.

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, 2006, Germany, [w:] K.P. Jantke, G. Kreuzberger (red.), *Diskussionsbeiträge*, Institut für Medien und Kommunikationswisschaft, Technische Universität Ilmenau, July 2006, no 21, s. 61-69.

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

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

Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996.

Walesiak M., *Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów*, [w:] A. Zeliaś (red.), *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, AE, Kraków 2005, s. 185-203.

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

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

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

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

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

GDM DISTANCE FOR ORDINAL DATA AND SPECTRAL CLUSTERING

Summary: Spectral clustering methods are well known in literature. In the article the proposal of spectral clustering method for ordinal data, based on the procedure of Ng, Jordan and Weiss [2002], is presented. In the construction of affinity matrix we implement kernel function with GDM distance for ordinal data. Also the article gives the proposal of finding the best *kernel width* parameter σ in spectral clustering methods. The empirical research is presented at the end of article based on two types of data (simulated and real).