Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

PROBLEMATYKA DOBORU MIARY ODLEGŁOŚCI W KLASYFIKACJI SPEKTRALNEJ DANYCH SYMBOLICZNYCH

Wprowadzenie

Zagadnienie doboru odpowiedniej miary odległości stanowi, obok problematyki doboru liczby klas, jeden z kluczowych kroków w klasyfikacji spektralnej.

Celem artykułu jest przetestowanie przydatności siedmiu różnych miar odległości dla danych symbolicznych w przypadku zastosowania klasyfikacji spektralnej dla danych tego typu. W badaniach symulacyjnych wykorzystano dane symboliczne interwałowe o znanej strukturze klas obiektów wygenerowane z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` oraz zbiory danych o nietypowych strukturach klas wygenerowane z zastosowaniem funkcji pakietu `mlbench`. Dla każdego modelu wygenerowano 40 zbiorów danych, przeprowadzono klasyfikację spektralną z zastosowaniem danej miary odległości. Otrzymane rezultaty porównano ze znaną strukturą klas z wykorzystaniem skorygowanego indeksu Randa.

1. Klasyfikacja spektralna

W analizie danych symbolicznych opracowano wiele różnych metod klasyfikacji (hierarchicznych i iteracyjno-optymalizacyjnych)[1]. Możliwe jest także zastosowanie klasycznych metod analizy skupień, o ile bazują one na macierzach odległości. Niemniej jednak istotne jest modyfikowanie istniejących rozwiązań dla danych klasycznych na potrzeby danych symbolicznych i rozwijanie nowych metod klasyfikacji danych symbolicznych.

[1] Zob. np. R. Verde: Clustering methods in symbolic data analysis. W: Classification, Clustering and Data Mining Applications. Eds. D. Banks et al. Springer-Verlag, Heidelberg 2004, s. 299-317; A. Dudek: Metody analizy danych symbolicznych w badaniach ekonomicznych. Wydawnictwo UE we Wrocławiu, Wrocław 2013, s. 66-79.

Nazwa klasyfikacji spektralnej nawiązuje do jednego z podstawowych kroków tej metody, w którym wyznacza się spektrum macierzy Laplace'a. W matematyce zbiór wartości własnych macierzy nazywa się widmem (spektrum) macierzy[2]. Podstawowy algorytm klasyfikacji spektralnej zaproponowano w pracy Ng, Jordan i Weiss[3]. Modyfikacje tego algorytmu zaproponowano m.in. w pracach: Shorteed[4], Walesiaka i Dudka[5], Walesiaka[6].

W pracy von Luxburg przedstawiono badania porównawcze, z których wynika, że klasyfikacja spektralna często daje znacznie lepsze rezultaty niż tradycyjne metody klasyfikacji. Wynika to z faktu, że nie przyjmuje się w niej żadnych założeń co do kształtu skupień. Dodatkowo klasyfikacja spektralna w większości prezentowanych tam przypadków daje lepsze rezultaty dla skupień o nietypowych kształtach[7].

Klasyfikacja spektralna dla danych symbolicznych interwałowych składa się z następujących kroków[8]:

1. Konstrukcja tablicy danych symbolicznych $\mathbf{V} = \left[v_{ij}\right]$ o wymiarach $n \times m$ ($i = 1, \ldots, n$ – numer obiektu, $j = 1, \ldots, m$ – numer zmiennej).

2. Zastosowanie estymatora jądrowego do obliczenia macierzy podobieństw $\mathbf{A} = \left[A_{ik}\right]$ (*affinity matrix*) między obiektami. Najczęściej do wyznaczenia macierzy $\mathbf{A}$ wykorzystywany jest estymator gaussowski[9]:

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

gdzie:

d_{ik} – odległość między i-tym i k-tym obiektem symbolicznym,

σ – parametr skali (szerokość pasma – *kernel width*), $A_{ii} = 0$.

---

[2] Cyt. za: M. Walesiak: Zagadnienie doboru liczby klas w klasyfikacji spektralnej. „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu" 2013, nr 278, s. 33-34.

[3] A. Ng, M. Jordan, Y. Weiss: On spectral clustering: Analysis and algorithm. W: Advances in Neural Information Processing Systems 14. Eds. T. Dietterich, S. Becker, Z. Ghahramani. MIT Press, Cambridge 2002, s. 849-856.

[4] S. Shorteed: Learning in spectral clustering. Rozprawa doktorska. Univeristy of Washington 2006.

[5] M. Walesiak, A. Dudek: Odległość GDM dla danych porządkowych a klasyfikacja spektralna. „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu" 2009, nr 84, s. 9-19.

[6] M. Walesiak: Klasyfikacja spektralna a skale pomiaru zmiennych. „Przegląd Statystyczny" 2012, z. 1, s. 13-31.

[7] Zob. np. U. von Luxburg, O. Bousquet, M. Belkin: Limits of spectral clustering. W: Advances in Neural Information Processing Systems (NIPS) 17. Eds. L. Saul, Y. Weiss, L. Bottou. MIT Press, Cambridge, MA, 2005, s. 857-864.

[8] Opracowano na podstawie prac: M. Walesiak, A. Dudek: Odległość GDM…, op. cit., s. 12-14; A. Dudek: Metody analizy…, op. cit., s. 78; M. Walesiak: Zagadnienie doboru…, op. cit., s. 34-35.

[9] A. Karatzoglou: Kernel methods. Software, algorithms and applications. Rozprawa doktorska. Technische Universität Wien 2006, s. 26.

Innymi przykładami estymatorów jądrowych, które mogą być zastosowane w tym kroku, są m.in.: jądro wielomianowe, jądro liniowe, jądro w postaci tangensa hiperbolicznego, jądro Bessela, jądro ANOVA czy jądro łańcuchowe (dla danych tekstowych)[10].

W artykule przetestowano siedem różnych miar odległości (z zastosowaniem estymatora jądrowego wyrażonego wzorem (1)), które można zastosować w przypadku zmiennych symbolicznych interwałowych[11]:

a) Miara Ichino-Yaguchiego (U_2):

$$\sqrt[q]{\sum_{j=1}^{m} \phi\left(v_{ij}, v_{kj}\right)^q}, \tag{2}$$

gdzie:

$\phi(v_{ij}, v_{kj}) = \left|v_{ij} \oplus v_{kj}\right| - \left|v_{ij} \otimes v_{kj}\right| + \gamma\left(2 \cdot \left|v_{ij} \oplus v_{kj}\right| - \left|v_{ij}\right| - \left|v_{kj}\right|\right)$,

v_{ij}, v_{kj} – oznacza realizację j-tej zmiennej symbolicznej w i-tym oraz k-tym obiekcie,

$\oplus$ oraz $\otimes$ są rozszerzeniem pojęcia sumy i iloczynu kartezjańskiego na zmienne symboliczne,

$\left|\ \right|$ – dla zmiennych interwałowych oznacza długość przedziałów, dla zmiennych wielowariantowych liczbę wariantów (kategorii),

γ – parametr ustalany arbitralnie przez badacza (zwykle $\gamma = 0{,}5$).

b) Znormalizowana miara Ichino-Yaguchiego (U_3):

$$\sqrt[q]{\sum_{j=1}^{m} \psi\left(v_{ij}, v_{kj}\right)^q}, \tag{3}$$

gdzie:

$\psi\left(v_{ij}, v_{kj}\right) = \phi\left(v_{ij}, v_{kj}\right) / \left|V_j\right|$,

$\left|V_j\right|$ – zbiór obrazów zmiennej symbolicznej, pozostałe oznaczenia jak we wzorze (2).

c) Miara de Carvalho, która jest modyfikacją odległości Ichino-Yaguchiego (SO_3):

$$\sqrt[q]{\sum_{j=1}^{m} \frac{1}{m}\left[\psi\left(v_{ij}, v_{kj}\right)\right]^q}, \tag{4}$$

[10] M. Walesiak: Zagadnienie doboru…, op. cit., s. 35.

[11] Inne miary odległości dla danych symbolicznych można znaleźć np. w: A. Dudek: Metody analizy…, op. cit., s. 51-61.

gdzie:
$\psi(v_{ij}, v_{kj}) = \phi(v_{ij}, v_{kj}) / \mu(v_{ij} \oplus v_{kj})$, $\mu(v_{ij}, v_{kj})$ – oznacza długość przedziału dla zmiennych interwałowych – w pozostałych przypadkach jest to $|\ |$, pozostałe oznaczenia jak we wzorze (2).

d) Miara de Carvalho oparta na pojęciu potencjału opisowego obiektu symbolicznego (SO_3):

$$\pi(A_i \oplus A_k) - \pi(A_i \oplus A_k) + \gamma[2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k)], \tag{5}$$

gdzie:
π – potencjał opisowy obiektu symbolicznego:

$$\pi(A_i) = \prod_{k=1}^{m} \mu(v_{ik}), \tag{6}$$

pozostałe oznaczenia jak we wzorach (2) i (4).

e) Znormalizowana miara de Carvalho oparta na pojęciu potencjału opisowego obiektu symbolicznego:

$$[\pi(A_i \oplus A_k) - \pi(A_i \oplus A_k) + \gamma[2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k)]] / \pi(A^E), \tag{7}$$

gdzie:
$\pi(A^E)$ – oznacza potencjał opisowy najbardziej ogólnego obiektu symbolicznego (w rozumieniu potencjału opisowego), pozostałe oznaczenia jak we wzorze (5).

f) Znormalizowana miara de Carvalho oparta na pojęciu potencjału opisowego obiektu symbolicznego – postać druga miary (SO_5):

$$[\pi(A_i \oplus A_k) - \pi(A_i \oplus A_k) + \gamma[2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k)]] / \pi(A_i \oplus A_k), \tag{8}$$

gdzie:
oznaczenia jak we wzorze (6).

g) Miara Hausdorffa (H):

$$\left[\sum_{j=1}^{m} \left(\max\left\{\left|\overline{v}_{ij} - \overline{v}_{kj}\right|, \left|\underline{v}_{ij} - \underline{v}_{kj}\right|\right\}\right)^2\right]^{\frac{1}{2}}, \tag{9}$$

gdzie:
$\overline{v}_{ij}, \overline{v}_{kj}$ $(\underline{v}_{ij}, \underline{v}_{kj})$ – oznaczają odpowiednio górne (dolne) krańce przedziału zmiennej symbolicznej interwałowej.

Parametr skali (σ), podobnie jak w przypadku klasyfikacji spektralnej dla danych klasycznych, ma kluczowe znaczenie dla klasyfikacji spektralnej. Poszukiwana jest taka wartość parametru skali, która dla zadanej liczby klas będzie minimalizować zmienność wewnątrzklasową. Jest to heurystyczna metoda poszukiwania minimum lokalnego[12].

3. Obliczenie diagonalnej macierzy $\mathbf{D}$, na głównej przekątnej tej macierzy znajdują się sumy każdego wiersza z macierzy $\mathbf{A}$, a poza nią są zera.

4. Konstrukcja znormalizowanej macierzy Laplace'a[13]:

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} \tag{10}$$

5. 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, gdzie u – liczba klas, wektorów własnych tworzy macierz $\mathbf{E} = \left[e_{ij}\right]$ o wymiarach $n \times u$.

6. Przeprowadzenie normalizacji macierzy $\mathbf{E}$ zgodnie ze wzorem:

$$y_{ij} = e_{ij} \Big/ \sqrt{\sum_{j=1}^{u} e_{ij}^{\ 2}}\ , \tag{11}$$

gdzie:
$i = 1, \ldots, n$ – 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.

7. Macierz $\mathbf{Y}$ stanowi punkt wyjścia zastosowania jednej z klasycznych metod analizy skupień (zwykle jest to metoda *k*-średnich).

2. Dane symboliczne

Obiekty symboliczne mogą być opisywane przez następujące rodzaje zmiennych symbolicznych[14]:

a) ilorazowe,
b) przedziałowe,

[12] M. Walesiak, Zagadnienie doboru …, op. cit., s. 41.
[13] Własności tej macierzy zaprezentowano m.in. w pracy: U. von Luxburg: A tutorial on spectral clustering. Max Planck Institute for Biological Cybernetics, Technical Report TR-149, 2006.
[14] Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data. Eds. H.-H. Bock, E. Diday. Springer Verlag, Berlin 2000, s. 2-3.

c) porządkowe,
d) nominalne,
e) interwałowe, których realizacją są przedziały liczbowe rozłączne lub nierozłączne;
f) wielowariantowe, gdzie realizacją zmiennej jest więcej niż jeden wariant (liczba lub kategoria);
g) wielowariantowe z wagami, gdzie realizacją zmiennej oprócz wielu wariantów są dodatkowo wagi (lub prawdopodobieństwa) dla każdego z wariantów zmiennej dla danego obiektu,
h) interwałowe z wagami (histogramowe).

Przykłady zmiennych symbolicznych wraz z ich realizacjami zawarto w tabeli 1.

Tabela 1

Przykłady zmiennych symbolicznych wraz z realizacjami

Zmienna	Realizacje	Typ zmiennej symbolicznej
Preferowana cena samochodu (w zł)	<27000, 42000>; <35000, 50000> <20000, 30000>; <25000, 37000>	interwałowa (przedziały nierozłączne)
Rozważana pojemność silnika (w cm^3)	<1000, 1200>; <1300, 1400> <1500, 1800>; <1900, 2200>	interwałowa (przedziały rozłączne)
Wybrany kolor	{niebieski, czerwony, żółty} {zielony, czarny, szary, biały}	wielowariantowa
Preferowana marka samochodu	{Toyota (0,3); Volvo (0,7)} {Audi (0,6); Skoda (0,4)} {VW (1,0)}	wielowariantowa z wagami

Niezależnie od typu zmiennej w analizie danych symbolicznych możemy mieć do czynienia ze zmiennymi strukturalnymi[15]. Do tego typu zmiennych zalicza się zmienne hierarchiczne – w których *a priori* ustalone są reguły decydujące o tym, czy dana zmienna opisuje dany obiekt, czy nie; zmienne taksonomiczne – w których ustalone są *a priori* realizacje danej zmiennej; zmienne logiczne – tj. takie, dla których ustalono *a priori* reguły logiczne lub funkcyjne decydujące o wartościach zmiennej.

W analizie danych symbolicznych wyróżnia się dwa typy obiektów symbolicznych:
– obiekty symboliczne pierwszego rzędu – obiekty rozumiane w sensie „klasycznym" (obiekty elementarne), np. konsument, przedsiębiorstwo, produkt, pacjent czy gospodarstwo domowe,
– obiekty symboliczne drugiego rzędu – obiekty utworzone w wyniku agregacji zbioru obiektów symbolicznych pierwszego rzędu, np. grupa konsumentów preferująca określony produkt, region geograficzny (jako wynik agregacji podregionów).

[15] Ibid., s. 2-3, 33-37.

3. Badania symulacyjne

Dla celów badania symulacyjnego z wykorzystaniem siedmiu zaprezentowanych miar odległości przygotowano cztery zbiory danych o znanej strukturze klas. Dla każdego ze zbiorów i każdej miary odległości przeprowadzono 40 symulacji. W celu wybrania ostatecznej liczby klas zastosowano indeks sylwetkowy pozwalający na ocenę prawidłowego zaklasyfikowania poszczególnych obiektów do klas w postaci[16]:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}, \tag{12}$$

gdzie:

$a(i) = \sum_{k \in \{P_r \setminus i\}} d_{ik} / (n_r - 1)$ – oznacza średnią odległość *i*-tego obiektu od pozostałych obiektów należących do klasy P_r,

$b(i) = \min_{s \neq r} \{d_{iP_s}\}$, d_{iP_s} – średnia odległość *i*-tego obiektu od obiektów należących do klasy P_s $\left(d_{iP_s} = \sum_{k \in P_s} d_{ik} / n_s\right)$,

$r, s = 1, \ldots, u$ – numer klasy,

u – liczba klas.

Ogólna jakość klasyfikacji oraz prawidłowość wyodrębnienia poszczególnych klas są mierzone jako[17]: $S(P_r) = \sum_{k \in P_r} S(i)/n_r$ oraz $S(P) = \sum_r S(P_r)/u$.

Do porównania rezultatów ze znaną strukturą klas wykorzystano skorygowany indeks Randa[18] w postaci:

$$R_{HA} = \frac{R - E(R)}{R_{\max} - E(R)}, \tag{13}$$

gdzie:

$R = 1 - N \Big/ \binom{n}{2}$,

$R_{\max}$ – maksymalna wartość miary Randa $(R_{\max} = 1)$,

[16] Szerzej o tym indeksie oraz innych indeksach służących wyborowi liczby klas pisze np. M. Walesiak: Metody klasyfikacji. W: Metody statystycznej analizy wielowymiarowej w badaniach marketingowych. Red. E. Gatnar, M. Walesiak. Wydawnictwo AE we Wrocławiu, Wrocław 2004, s. 338- 343.

[17] Ibid., s. 343.

[18] Zob. np. M. Walesiak: Problemy decyzyjne w procesie klasyfikacji zbioru obiektów. „Prace Naukowe Akademii Ekonomicznej we Wrocławiu" 2004, nr 1010, s. 60-61.

$E(R)$ – oczekiwana wartość miary Randa wyrażona wzorem:

$$E(R) = 1 + 2\sum_r \binom{n_{.r}}{2} \sum_s \binom{n_{s.}}{2} \Big/ \binom{n}{2}^2 - \left[\sum_r \binom{n_{.r}}{2} + \sum_s \binom{n_{s.}}{2} \right] \Big/ \binom{n}{2}, \quad (14)$$

gdzie:

$n_{.r}$ – liczba obiektów w klasie $P_r^{(t)}$,

$n_{s.}$ – liczba obiektów w klasie $P_s^{(q)}$.

Za pomocą funkcji `cluster.Gen` z pakietu `clusterSim` wygenerowano dwa modele:

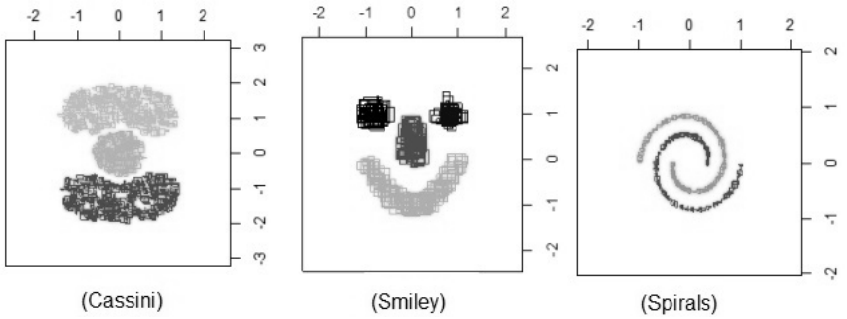
1. Zbiór 100 obserwacji podzielonych na dwie klasy o wydłużonym kształcie opisywane przez dwie zmienne symboliczne interwałowe. Środki ciężkości klas to (0, 0) oraz (1, 5) i macierz kowariancji Σ, w której $(\delta_{jj} = 1, \delta_{il} = -0{,}9)$.

2. Zbiór 250 obserwacji podzielonych na pięć niezbyt dobrze separowalnych klas opisywanych przez dwie zmienne symboliczne interwałowe. Środki ciężkości klas to (5, 5), (–3, 3), (3, –3), (0, 0), (–5, –5), i macierzy kowariancji Σ, gdzie $\delta_{jj} = 1\,(1 \le j \le 3)$, $\delta_{12} = \delta_{13} = -0{,}9$, $\delta_{23} = 0{,}9$.

Z wykorzystaniem pakietu `mlbench` (za pomocą funkcji `mlbench.cassini`, `mlbench.smiley` oraz `mlbench.spirals`) wygenerowano trzy zbiory klas (zob. rys. 1) opisywanych zmiennymi symbolicznymi interwałowymi:

1. Zbiór danych Cassini (zawierający 1000 obiektów podzielonych na trzy klasy).
2. Zbiór danych Smiley (zawierający 300 obiektów podzielonych na cztery klasy).
3. Zbiór danych Spirals (zawierający 300 obiektów podzielonych na dwie klasy).

Zbiory te zawierają struktury klas o nietypowych kształtach. Zostały one uzyskane w ten sposób, że za pomocą funkcji pakietu `mlbench` wygenerowano klasyczne zbiory danych (zawierające punkty) o tych samych nazwach. Następnie, zachowując oryginalny kształt skupień, dodano niewielkie odchylenia dla tych punktów celem otrzymania danych symbolicznych interwałowych.



Rys. 1. Zbiory danych otrzymane z wykorzystaniem pakietu `mlbench`

Wyniki badań symulacyjnych w postaci średnich wartości skorygowanego indeksu Randa obliczonego na podstawie wszystkich 40 symulacji z zastosowaniem danej miary odległości zestawiono w tabeli 2.

Tabela 2

Zestawienie wyników badań symulacyjnych

Nazwa modelu	Zmienne zakłócające	Miara odległości						
		H	U_2	U_3	SO_2	SO_3	SO_4	SO_5
Zbiór 1	brak	1	1	1	1	1	1	1
Zbiór 1	1	1	1	1	1	1	1	1
Zbiór 2	brak	0,95	0,99	0,94	0,87	0,90	0,90	0,89
Smiley	brak	0,66	0,87	0,88	0,88	0,90	0,89	0,89
Cassini	brak	0,87	0,90	0,90	0,90	0,90	0,90	0,90
Spirals	brak	0,88	0,91	0,92	0,93	0,93	0,94	0,94
Średnia		0,89	0,95	0,94	0,93	0,94	0,94	0,94

Podsumowanie

Klasyfikacja spektralna może z powodzeniem znaleźć zastosowanie w analizie skupień dla danych symbolicznych dzięki zastosowaniu odpowiedniej miary odległości dla danych tego typu.

Podobnie jak w przypadku danych klasycznych, tak i w przypadku klasyfikacji spektralnej danych symbolicznych istotne znaczenie ma parametr σ (zob. wzór 1), który powinien minimalizować odległości wewnątrzklasowe przy zadanej liczbie klas.

Najlepsze wyniki dla analizowanych zbiorów danych i zastosowanego indeksu doboru liczby klas (zob. tabela 2) osiągnęła miara Ichino-Yaguchiego (SO_2). Zbliżone rezultaty (w sensie skorygowanego indeksu Randa) osiągnęły wszystkie miary znormalizowane. Zbliżone wyniki do miar znormalizowanych osiągnęła nieznormalizowana miara de Carvalho (SO_2), która jest modyfikacją odległości Ichino-Yaguchiego. Najgorsze wyniki otrzymano dla miary odległości Hausdorffa (H).

Celem dalszych prac będzie porównanie jakości otrzymanych wyników (w sensie skorygowanego indeksu Randa), jeżeli zastosowane zostaną inne miary odległości oraz inne indeksy służące doborowi liczby klas.

Literatura

Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data. Eds. H.-H. Bock, E. Diday. Springer Verlag, Berlin 2000.

Dudek A.: Metody analizy danych symbolicznych w badaniach ekonomicznych. Wydawnictwo UE we Wrocławiu, Wrocław 2013.

Karatzoglou A.: Kernel methods. Software, algorithms and applications. Rozprawa doktorska. Technische Universität Wien 2006.

Leisch F., Dimitriadou E.: mlbench package, 2010, www.r-project.org.

Luxburg U. von: A tutorial on spectral clustering. Max Planck Institute for Biological Cybernetics, Technical Report TR-149, 2006.

Luxburg U. von, Bousquet O., Belkin M.: Limits of spectral clustering. W: Advances in Neural Information Processing Systems (NIPS) 17. Eds. L. Saul, Y. Weiss, L. Bottou. MIT Press, Cambridge, MA, 2005.

Ng A., Jordan M., Weiss Y.: On spectral clustering: Analysis and algorithm. W: Advances in Neural Information Processing Systems 14. Eds. T. Dietterich, S. Becker, Z. Ghahramani. MIT Press, Cambridge 2002.

Shorteed S.: Learning in spectral clustering. Rozprawa doktorska. Univeristy of Washington 2006.

Verde R.: Clustering methods in symbolic data analysis. W: Classification, Clustering and Data Mining Applications. Eds. D. Banks, L. House, E.R. McMorris, P. Arabie, W. Gaul. Springer-Verlag, Heidelberg 2004.

Walesiak M., Dudek A.: clusterSim package, 2013, www.r-project.org.

Walesiak M., Dudek A.: Odległość GDM dla danych porządkowych a klasyfikacja spektralna. „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu" 2009, nr 84.

Walesiak M.: Klasyfikacja spektralna a skale pomiaru zmiennych. „Przegląd Statystyczny" 2012, z. 1.

Walesiak M.: Metody klasyfikacji. W: Metody statystycznej analizy wielowymiarowej w badaniach marketingowych. Red. E. Gatnar, M. Walesiak. Wydawnictwo AE we Wrocławiu, Wrocław 2004.

Walesiak M.: Problemy decyzyjne w procesie klasyfikacji zbioru obiektów. „Prace Naukowe Akademii Ekonomicznej we Wrocławiu" 2004, nr 1010.

Walesiak M.: Zagadnienie doboru liczby klas w klasyfikacji spektralnej. „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu" 2013, nr 278.

THE PROBLEM OF DISTANCE MEASURE SELECTION FOR SPECTRAL CLUSTERING OF SYMBOLIC DATA

Summary

Spectral clustering that was proposed by Ng, Jordan and Weiss, is not in fact a new clustering method, but rather a new way to prepare data set for clustering method. This method uses the idea of spectral decomposition.

The main aim of the paper is to present a possibility of application spectral clustering when dealing symbolic data, with a special focus on different distance measures that can be applied for this kind of data. In experiment studies artificial data sets with known

cluster structure were obtained with application of `clusterSim` and `mlbench` packages of R software. Each data set was clustered 40 times with application of each distance measure applied. Received results were compared with known cluster structure with application of adjusted Rand index.