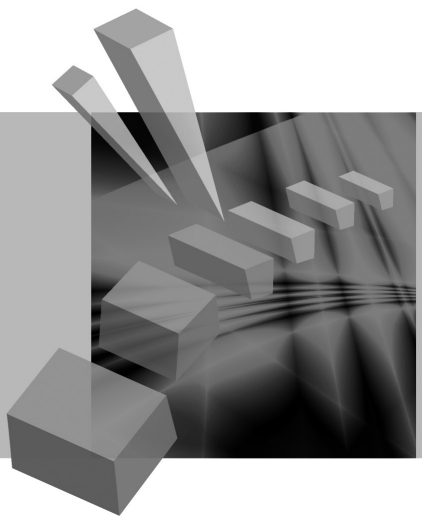
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

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

Wstęp..... 9

Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokołowski: Sejm VI kadencji – maszynka do głosowania 11
Barbara Pawełek, Adam Sagan: Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I 19
Jan Paradysz: Nowe możliwości badania koniunktury na rynku pracy 29
Krzysztof Najman: Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze 41
Kamila Migdał-Najman: Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym 48
Aleksandra Matuszewska-Janica, Dorota Witkowska: Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych 58
Iwona Foryś, Ewa Putek-Szeląg: Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim 67
Joanna Banaś, Małgorzata Machowska-Szewczyk: Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych..... 77
Marta Jarocka: Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni 85
Anna Zamojska: Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych..... 95
Dorota Rozmus: Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody *bagging* 106
Ewa Wędrowska: Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur..... 115
Katarzyna Wójcik, Janusz Tuchowski: Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ... 124
Małgorzata Misztal: Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych..... 135
Anna Czapkiewicz, Beata Basiura: Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych 146
Tomasz Szubert: Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej” 154

Marcin Szymkowiak: Konstrukcja estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości 164
Wojciech Roszka: Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie 174
Justyna Brzezińska: Metody wizualizacji danych jakościowych w programie **R** 182
Agata Sielska: Regionalne zróżnicowanie potencjału konkurencyjnego polskich gospodarstw rolnych w województwach po akcesji do Unii Europejskiej 191
Mariusz Kubus: Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych 201
Beata Basiura: Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości 209
Katarzyna Wardzińska: Wykorzystanie metody obwiedni danych w procesie klasyfikacji przedsiębiorstw 217
Katarzyna Dębkowska: Modelowanie upadłości przedsiębiorstw oparte na próbach niezbilansowanych 226
Danuta Tarka: Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska .. 235
Artur Czech: Zastosowanie wybranych metod doboru zmiennych diagnostycznych w badaniach konsumpcji w ujęciu pośrednim 246
Beata Bal-Domańska: Ocena relacji zachodzących między inteligentnym rozwojem a spójnością ekonomiczną w wymiarze regionalnym z wykorzystaniem modeli panelowych 255
Mariola Chrzanowska: *Ordinary kriging* i *inverse distance weighting* jako metody szacowania cen nieruchomości na przykładzie warszawskiego rynku 264
Adam Depta: Zastosowanie analizy wariancji w badaniu jakości życia na podstawie kwestionariusza SF-36v2 272
Maciej Beręsewicz, Tomasz Klimanek: Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach cen mieszkań 281
Karolina Paradysz: Benchmarkowa analiza estymacji dla małych obszarów na lokalnych rynkach pracy 291
Anna Gryko-Nikitin: Dobór parametrów w równoległych algorytmach genetycznych dla problemu plecakowego 301
Tomasz Ząbkowski, Piotr Jałowiecki: Zastosowanie reguł asocjacyjnych do analizy danych ankietowych w wybranych obszarach logistyki przedsiębiorstw przetwórstwa rolno-spożywczego 311
Agnieszka Przedborska, Małgorzata Misztal: Zastosowanie metod statystyki wielowymiarowej do oceny wydolności stawów kolanowych u pacjentów z chorobą zwyrodnieniową leczonych operacyjnie 321
Dorota Perło: Rozwój zrównoważony w wymiarze gospodarczym, społecznym i środowiskowym – analiza przestrzenna 331

Ewa Putek-Szeląg, Urszula Gierałtowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej .. 342

Summaries

Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokołowski: VIth-term Sejm – a voting machine 18
Barbara Pawełek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model 28
Jan Paradysz: New possibilities for studying the situation on the labour market 40
Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data..... 47
Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering 57
Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees..... 66
Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship... 76
Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables 84
Marta Jarocka: The impact of the method of the selection of diagnostic variables on the result of linear ordering on the example of ranking of universities in Poland..... 94
Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures..... 105
Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea..... 114
Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements 123
Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis 134
Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ... 145
Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series..... 153
Tomasz Szubert: Factors differentiating the level of satisfaction with life and the life's values of people with and without disabilities in the light of the "Social Diagnosis" survey 162
Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures 173

Wojciech Roszka: Joint characteristics' estimation of variables not jointly observed ... 181
Justyna Brzezińska: Visualizing categorical data in **R** ... 190
Agata Sielska: Regional diversity of competitiveness potential of Polish farms after the accession to the European Union ... 200
Mariusz Kubus: Regularized linear probability model as a filter ... 208
Beata Basiura: The Ward method in the application for classification of Polish voivodeships with different distances ... 216
Katarzyna Wardzińska: Application of Data Envelopment Analysis in company classification process ... 225
Katarzyna Dębkowska: Modeling corporate bankruptcy based on unbalanced samples ... 234
Danuta Tarka: Influence of the features selection method on the results of objects classification using environmental data ... 245
Artur Czech: Application of chosen methods for the selection of diagnostic variables in indirect consumption research ... 254
Beata Bal-Domańska: Assessment of relations occurring between smart growth and economic cohesion in regional dimension using panel models 263
Mariola Chrzanowska: Ordinary kriging and inverse distance weighting as methods of estimating prices based on Warsaw real estate market ... 271
Adam Depta: Application of analysis of variance in the study of the quality of life based on questionnaire SF-36v2 ... 280
Maciej Beręsewicz, Tomasz Klimanek: Using indirect estimation with spatial autocorrelation in dwelling price surveys ... 290
Karolina Paradysz: Benchmark analysis of small area estimation on local labor markets ... 300
Anna Gryko-Nikitin: Selection of various parameters of parallel evolutionary algorithm for knapsack problems ... 310
Tomasz Ząbkowski, Piotr Jałowiecki: Application of association rules for the survey of data analysis in the selected areas of logistics in food processing companies ... 320
Agnieszka Przedborska, Małgorzata Misztal: Using multivariate statistical methods to assess the capacity of the knee joint among the patients treated surgically for osteoarthritis ... 330
Dorota Perło: Sustainable development in the economic, social and environmental dimensions – spatial analysis ... 341
Ewa Putek-Szeląg, Urszula Gierałtowska: Analysis and diagnosis of the volume of renewable energy production in Poland compared to EU countries ... 352

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Beata Basiura
AGH Akademia Górniczo-Hutnicza w Krakowie

METODA WARDA W ZASTOSOWANIU KLASYFIKACJI WOJEWÓDZTW POLSKI Z RÓŻNYMI MIARAMI ODLEGŁOŚCI

Streszczenie: W niniejszej pracy proponuje się porównanie wyników klasyfikacji województw Polski z zastosowaniem różnych funkcji celu. Badanie starano się przeprowadzić na podstawie danych empirycznych z uzasadnieniem wyboru miary podobieństwa. Porównano wyniki otrzymane przy zastosowaniu klasycznej metody Warda oraz algorytmu proponowanego przez formułę Lance'a-Williamsa. Wydaje się, że stosowanie różnych miar odległości w klasyfikacji województw Polski metodą Warda daje porównywalne jakości klasyfikacji. Na tle zaproponowanych odległości wyróżnia się ważona odległość euklidesowa.

Słowa kluczowe: metoda Warda, jakość klasyfikacji, miary niepodobieństwa.

1. Wstęp

Problem grupowania obiektów jest problemem odkrywania struktury grupowej na podstawie zaobserwowanych danych. Najczęstsze zastosowanie w praktyce ma zdefiniowanie pewnej funkcji jakości klasyfikacji i szukanie algorytmu, który pozwoli na maksymalizację lub minimalizację tej funkcji. Istnieje wiele różnych możliwości wyboru takiej funkcji. Stosowane są klasyfikacje na podstawie odległości punktów danych od środków grup czy różnych kryteriów związanych z podobieństwem wewnątrzgrupowym lub niepodobieństwem pomiędzy grupami. Hierarchiczne metody aglomeracji są jedną z takich metod klasyfikacji, a na ich wyniki ogromny wpływ ma wybór miary podobieństwa obiektów. W szczególności w klasycznej metodzie Warda podkreślany jest wymóg kwadratu odległości euklidesowej jako miary podobieństwa obiektów. Mimo znanych uogólnień tej metody, opartej na funkcji celu, stosowanie różnych miar odległości nie jest zalecane. W literaturze przedmiotu podkreśla się, że stosowanie innych miar odległości nie ma interpretacji geometrycznej [Jain, Dubes 1988; Gatnar, Walesiak (red.) 2004; Walesiak, Gatnar 2009]. Natomiast niektórzy autorzy proponują i weryfikują stosowanie klasyfikacji metodą Warda na podstawie innych funkcji celu [Batagelj 1988; Szekely, Rizzo 2005] i innych miar odległości pomiędzy obiektami [Mirkin 2005]. Motywacją podjęcia tematu jest pró-

ba porównania wyników klasyfikacji metodą Warda przy zastosowaniu algorytmu klasycznego, stosującego minimalizację funkcji celu, z wynikami uzyskanymi przy zastosowaniu algorytmu Lance'a-Williamsa [Lance, Williams, 1967] na przykładzie klasyfikacji województw Polski. Badanie starano się przeprowadzić na podstawie danych empirycznych z uzasadnieniem wyboru miary podobieństwa.

2. Metoda Warda

2.1. Klasyczny algorytm grupowania obiektów metodą Warda

Metoda Warda to hierarchiczna metoda aglomeracyjna klasyfikacji obiektów, w której kryterium wyboru pary zbiorów łączonych w danym kroku jest wartością optymalną pewnej funkcji celu. Jak wiadomo, wiele jest możliwości wyboru funkcji celu. Jedną z nich, najbardziej popularną, jest suma kwadratów odchyleń poszczególnych elementów skupienia od środka ciężkości tego skupienia. Taką interpretację podał w swoim artykule J.H. Ward [1963]. Podstawową ideą opisywanej metody jest połączenie, w każdym kroku aglomeracji, takich dwóch podzbiorów, dla których funkcja celu dana równaniem (1) jest najmniejsza.

$$E = \sum_k E_k = \sum_k \sum_{i=1}^{n_k} (z_{ik} - \overline{z}_k)^2. \tag{1}$$

Na przykładzie podziału S zbioru obiektów Ω zawierającego podzbiory C_u oraz C_v można opisać myśl przewodnią tej metody. Punktem wyjścia jest odpowiedź na pytanie, w jaki sposób zmieni się błąd kwadratowy podziału S, oznaczony jako $W(S, c)$, jeśli połączymy dwa skupienia C_u oraz C_v. Porównując dwie klasyfikacje: podział S oraz podział $S(C_u, C_v)$, które różnią się tylko tym, że skupienia C_u and C_v są połączone w podziale $S(C_u, C_v)$, otrzyma się różnicę błędów kwadratowych daną wzorem (2).

$$d_W(C_u, C_v) = W(S(C_u, C_v), \overline{C}_{uv}) - W(S, \overline{C}), \tag{2}$$

przy czym przyjmuje się, że:

$\overline{C}_{uv}$ to środek klasy $C_u \cup C_v$. Ponieważ te dwa podziały na podzbiory różnią się tylko tym, że w podziale drugim skupienia C_u oraz C_v są połączone w jeden zbiór, to różnica błędów kwadratowych będzie różnicą błędów wyznaczonych dla skupienia łącznego $C_u \cup C_v$ oraz błędów wyznaczonych osobno dla skupień C_u i C_v, co można zapisać wzorem (3), w którym $\overline{C}_u, \overline{C}_v$ to środki klas C_u, C_v.

$$d_W(C_u, C_v) = W(S(C_u, C_v), \overline{C}_{uv}) - W(S_u, \overline{C}_u) - W(S_v, \overline{C}_v). \tag{3}$$

Na tej podstawie stwierdza się, że sumaryczny błąd kwadratowy połączonych skupień jest sumą kwadratowych błędów wyjściowych skupień i odległości Warda pomiędzy tymi skupieniami.

$$d_W(C_u, C_v) = \frac{n_u \cdot n_v}{n_u + n_v} d_2^2(\overline{C}_u, \overline{C}_v). \tag{4}$$

Zgodnie z algorytmem Warda odległość pomiędzy skupieniami można zapisać wzorem (4), przyjmując n_u jako liczność skupienia C_u, n_v jako liczność C_v, $\overline{C}_u$ i $\overline{C}_v$ jako środki skupień odpowiednio C_u i C_v, a d_2 jako odległość euklidesową.

2.2. Algorytm Lance'a-Williamsa

Metoda Warda implementowana jest najczęściej rekurencyjnie, poprzez modyfikację macierzy odległości zgodnie z algorytmem Lance'a-Williamsa [Lance, Williams 1967]. Algorytm ten wykonywany jest następująco:

Krok 1. Zakłada się, że każdy obiekt stanowi osobną grupę, i wyznacza się macierz odległości pomiędzy wszystkimi obiektami.

Krok 2. W macierzy odległości szuka się pary skupień najbardziej podobnych, dla których odległość jest najmniejsza w całej macierzy. Obiekty tych grup utworzą nowe skupienie.

Krok 3. W macierzy odległości wykreśla się jedną kolumnę i jeden wiersz, a następnie według odpowiedniej reguły przelicza się odległość nowego skupienia od pozostałych klas.

Krok 4. Kroki 1-3 powtarzane są, aż wszystkie obiekty znajdą się w jednej klasie.

Modyfikacja macierzy w kroku 3 wykonywana jest według wzoru (5),

$$\begin{aligned} d(C_s \cup C_t, C_i) = {} & \frac{n_s + n_i}{n_i + n_s + n_t} d(C_s, C_i) + \frac{n_t + n_i}{n_i + n_s + n_t} d(C_t, C_i) + \\ & + \frac{-n_i}{n_i + n_s + n_t} d(C_s, C_t), \end{aligned} \tag{5}$$

w którym C_s, C_t są grupami łączonymi w nowy zbiór, C_i jest dowolnym innym podzbiorem, wszystkie $d(C_t, C_s)$ są odległościami pomiędzy grupami C_t i C_s z macierzy z kroku poprzedniego, a wszystkie n_i oznaczają liczności odpowiednich skupień.

2.3. Uogólniona metoda Warda

W literaturze przedmiotu można znaleźć uogólnienia algorytmu Warda. Batagelj [1988] wprowadził uogólnioną miarę niepodobieństwa dwóch klas zdefiniowaną następująco:

$$D^W(C_u, C_v) = \frac{w(C_u) \cdot w(C_v)}{w(C_u \cup C_v)} d(\tilde{C}_u, \tilde{C}_v), \tag{6}$$

gdzie $w(C)$ jest wagą skupienia C, d jest miarą niepodobieństwa pomiędzy obiektami klas, natomiast $\tilde{C}$ jest uogólnionym środkiem skupienia C. Mirkin [2005, s.132]

proponuje stosowanie algorytmu Lance'a-Williamsa z różnymi miarami odległości pomiędzy obiektami zbioru Ω. Szekely i Rizzo [2005] zaproponowali specjalną miarę odległości nazwaną *e-distance* pozwalającą na łączne mierzenie odległości pomiędzy skupieniami i wewnątrz skupień. W efekcie stosowania tej odległości uzyskuje się grupy bardziej jednorodne.

3. Wybrane funkcje celu

Większość programów komputerowych pozwala na stosowanie metody Warda z dowolną macierzą odległości. Interesujące są zatem wyniki otrzymanej w ten sposób klasyfikacji. W niniejszej pracy pokazano wykorzystanie algorytmu metody Warda opartego na wprowadzonej wzorem (1) funkcji celu oraz algorytmu Lance'a-Williamsa z zastosowaniem różnych miar odległości obiektów. Zaprezentowano zastosowanie procedury na trzech zbiorach danych związanych ze wskaźnikami zatrudnienia w poszczególnych województwach Polski. Zaproponowano, aby w każdym kroku łączyć te dwie grupy, dla których po połączeniu wybrana odległość wszystkich punktów nowego skupienia od nowego środka będzie najmniejsza. Zastosowana została następująca funkcja celu $E_k = \sum_{i=1}^{n_k} d(z_{ik}, \overline{z}_k)$, przy czym przyjęto jako odległość d, co następuje:

$d = d^{\alpha}$ – jako odległość euklidesową w potędze α,
d – jako ważoną odległość euklidesową,
d – jako odległość Canbrerra,
d – jako *e-distance* daną wzorem (7)

$$e(C_i, C_j) = \frac{n_1 n_2}{n_1 + n_2}\left(\frac{2}{n_1 n_2} \sum_{k=1}^{n_1}\sum_{l=1}^{n_2} \hat{d}^{\alpha}(c_{ki} - c_{lj}) - \frac{1}{n_1^2} \sum_{k=1}^{n_1}\sum_{l=1}^{n_1} \hat{d}^{\alpha}(c_{ki} - c_{li}) - \frac{1}{n_2^2} \sum_{k=1}^{n_2}\sum_{l=1}^{n_2} \hat{d}^{\alpha}(c_{kj} - c_{lj}) \right), \tag{7}$$

gdzie $\hat{d}^{\alpha}$ jest odległością euklidesową podniesioną do potęgi α, a n_1, n_2 to liczności grup.

4. Badanie empiryczne

4.1. Zbiór danych

W badaniu empirycznym porównane zostały wyniki klasyfikacji w trzech przypadkach. W przykładzie pierwszym, jednowymiarowym, zastosowano zaprezentowaną procedurę do klasyfikacji województw Polski charakteryzowanych liczbą wolnych miejsc pracy. Przykład drugi prezentuje klasyfikacje województw pod kątem liczby wolnych miejsc pracy i procentowego współczynnika przyjęć do pracy. W przykła-

dzie trzecim rozważano zbiór województw z pięcioma wskaźnikami zatrudnienia, takimi jak: liczba wolnych miejsc pracy (w tys.), nowo utworzone miejsca pracy (w tys.), zlikwidowane miejsca pracy (w tys.), współczynnik zatrudnienia (w %), współczynnik zatrudnienia kobiet (w %). Wszystkie dane pochodzą z rocznika statystycznego województw wydanego przez GUS w roku 2010. Zgodnie z regułami klasyfikacji danych wszystkie zmienne poddane zostały standaryzacji.

4.2. Wyniki

Przeprowadzona klasyfikacja województw dla trzech różnych zbiorów danych, omówionych powyżej, pozwoliła podzielić zbiór województw na trzy skupienia. Wyniki grupowania nie były jednoznaczne i zależały od wyboru miary odległości pomiędzy skupieniami.

Podstawowym pytaniem postawionym na początku niniejszej pracy było, która klasyfikacja najlepiej wykrywa strukturę grupową zbioru obiektów. Dla każdej klasyfikacji dokonano oceny jakości na podstawie znanych współczynników oceny jakości. Wybrane zostały cztery współczynniki: indeks Calińskiego i Harabasza [1974], indeks Davies-Bouldina [1979], Silhouette indeks [Kaufman, Rousseeuw 1990] oraz współczynnik korelacji kofenetycznej [Sokal, Rohlf 1962]. We wszystkich indeksach wykorzystano odległość zastosowaną w funkcji celu. W obliczeniach niektórych indeksów wykorzystano procedury z pakietu ClusterSim programu *R*-project [Walesiak, Dudek, 2012]. Wybrane wyniki uzyskane w poszczególnych przykładach zawierają tab. 1, 2, 3.

Tabela 1. Wybrane wartości współczynników jakości klasyfikacji wyznaczone dla przykładu pierwszego

Odległość	Metoda	Indeks Calińskiego-Harabasza	Indeks DB	Silhouette indeks	Współczynnik korelacji kofenetycznej
Kwadrat odległości Euklidesa	klasyczna	89,49	0,54	0,74	0,91
	L-W	89,49	0,47	0,74	0,91
Odległość euklidesowa	klasyczna	89,49	0,78	0,58	0,94
	L-W	80,78	0,49	0,55	0,91
Odległość euklidesowa w potędze $\alpha = 0,5$	klasyczna	11,92	1,74	0,32	0,69
	L-W	80,78	0,49	0,41	0,83
Odległość euklidesowa w potędze $\alpha = 1,5$	klasyczna	89,49	0,63	0,68	0,92
	L-W	80,78	0,49	0,60	0,92
Odległość euklidesowa ważona	klasyczna	53,05	0,31	0,50	0,94
	L-W	80,78	0,49	0,55	0,91
e-distance $\alpha = 0,5$	L-W	80,78	0,49	0,55	0,77
e-distance $\alpha = 1,5$	L-W	80,78	0,49	0,55	0,93
Canberra	klasyczna	11,92	0,65	0,52	0,92
Canberra	L-W	4,05	1,59	0,63	0,94

Źródło: opracowanie własne.

Tabela 2. Wybrane wartości współczynników jakości klasyfikacji wyznaczone dla przykładu drugiego

Odległość	Metoda	Indeks Calińskiego-Harabasza	Indeks DB	Silhouette indeks	Współczynnik korelacji kofenetycznej
Kwadrat odległości Euklidesa	klasyczna	12,98	2,55	0,58	0,78
	L-W	13,54	0,63	0,55	0,74
Odległość euklidesowa	klasyczna	12,99	3,41	0,44	0,84
	L-W	12,99	0,80	0,45	0,83
Odległość euklidesowa w potędze $\alpha = 0,5$	klasyczna	12,99	3,53	0,30	0,81
	L-W	12,99	0,80	0,30	0,81
Odległość euklidesowa w potędze $\alpha = 1,5$	klasyczna	12,15	2,47	0,59	0,82
	L-W	12,15	0,48	0,59	0,81
Odległość euklidesowa ważona	klasyczna	12,99	0,29	0,59	0,80
	L-W	12,99	0,80	0,53	0,81
e-distance $\alpha = 0,5$	L-W	12,99	0,80	0,45	0,75
e-distance $\alpha = 1,5$	L-W	12,99	0,80	0,45	0,85
Canberra	klasyczna	7,14	1,54	0,28	0,52
Canberra	L- W	4,27	2,13	0,18	0,79

Źródło: opracowanie własne.

Tabela 3. Wybrane wartości współczynników jakości klasyfikacji wyznaczone dla przykładu trzeciego

Odległość	Metoda	Indeks Calińskiego-Harabasza	Indeks DB	Silhouette indeks	Współczynnik korelacji kofenetycznej
Kwadrat odległości Euklidesa	klasyczna	8,37	21,30	0,22	0,36
	L-W	8,37	1,05	0,50	0,48
Odległość euklidesowa	klasyczna	8,37	7,76	0,34	0,55
	L-W	6,73	1,05	0,29	0,56
Odległość euklidesowa w potędze $\alpha = 0,5$	klasyczna	6,61	3,63	0,14	0,49
	L-W	6,73	1,05	0,19	0,61
Odległość euklidesowa w potędze $\alpha = 1,5$	klasyczna	8,37	12,45	0,43	0,40
	L-W	8,37	1,05	0,43	0,52
Odległość euklidesowa ważona	klasyczna	6,24	0,31	0,32	0,83
	L-W	6,73	1,05	0,29	0,56
e-distance $\alpha = 0,5$	L-W	6,73	1,05	0,29	0,55
e-distance $\alpha = 1,5$	L-W	6,73	1,05	0,29	0,56
Canberra	klasyczna	6,61	1,05	0,38	0,82
Canberra	L-W	6,61	1,36	0,38	0,84

Źródło: opracowanie własne.

O lepszej jakości klasyfikacji mówią wyższe wartości indeksu Calińskiego i Harabasza, wyższe wartości współczynnika korelacji kofenetycznej oraz niższe indeksy Davies-Bouldina. Silhouette indeks z przedziału od 0,5 do 0,7 świadczy o poważnej strukturze klas, natomiast wartości wyższe niż 0,7 charakteryzują silną strukturę klas [Gatnar, Walesiak (red.) 2004].

Na podstawie przeprowadzonych badań można stwierdzić, że dla przykładu pierwszego wyższe współczynniki korelacji kofenetycznej otrzymuje się przy zastosowaniu metody klasycznej i odległości euklidesowej, jej kwadratu lub odległości euklidesowej z uwzględnieniem wag. Indeks Calińskiego i Harabasza także przyjmuje wyższe wartości dla metody klasycznej. Najniższą wartość indeksu Davies-Bouldina zaobserwować można w przypadku zastosowania ważonej odległości euklidesowej w metodzie klasycznej.

W przykładzie drugim wyniki są mniej zróżnicowane. Znów odległość euklidesowa ważona wyróżnia się najniższym indeksem Davies-Bouldina. Indeks Calińskiego i Harabasza jest najwyższy dla kwadratu odległości Euklidesa zastosowanego w algorytmie Lance'a-Williamsa. Współczynnik korelacji kofenetycznej osiąga wartość największą dla odległości e-distance z wartością α =1,5.

W przykładzie trzecim współczynnik korelacji kofenetycznej jest najwyższy przy zastosowaniu odległości Canberra i ważonej odległości euklidesowej. Po raz trzeci zastosowanie w metodzie klasycznej odległości euklidesowej ważonej wyróżnia się najniższym indeksem Davies-Bouldina. Indeks Calińskiego i Harabasza wskazuje jako najlepszą odległość euklidesową, jej kwadrat oraz potęgę o wykładniku α =1,5.

5. Podsumowanie

W niniejszej pracy proponuje się porównanie wyników klasyfikacji województw Polski metodą Warda z zastosowaniem różnych funkcji celu. Badanie starano się przeprowadzić na podstawie danych empirycznych. W wyniku uzyskano podział zbioru województw na trzy skupienia. Otrzymane współczynniki jakości klasyfikacji nie wskazują, który z algorytmów: klasyczny czy Lance'a-Williamsa powinien być stasowany przy klasyfikacji województw. Na tle zaproponowanych miar odległości w rozważanych zbiorach danych najlepiej wypada ważona odległość euklidesowa. Niestety wydaje się, że wybrane współczynniki jakości klasyfikacji nie wskazują jednoznacznie, który algorytm i którą miarę odległości należy wybrać. Możliwe, że bardziej jednoznaczne wyniki otrzymano by przy zastosowaniu innej miary jakości klasyfikacji, np. miar opartych o entropię.

Literatura

Batagelj V., 1988, *General Ward and Related Clustering Problems*, Classification and Related Methods of Data Analysis, Amsterdam, pp. 67-74.

Calinski R.B., Harabasz J., 1974, *A dendrite method for cluster analysis*, "Communications in Statistics", vol. 3, 1-27.

Davies D.L., Bouldin D.W., 1979, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 2, pp. 224-227.

Gatnar E.,Walesiak M. (red.), 2004, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE, Wrocław.

Jain A., Dubes R., 1988, *Algorithms for Clustering Data*, Prentice Hall, New Jersey.

Kaufman L., Rousseeuw P.J., 1990, *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley, New York, pp. 83-88.

Lance G., Williams W.T., 1967, *A general theory of classificatory storing strategies i hierarchical systems*, "Computer Journal", nr 9.

Mirkin B., 2005, *Clustering for Data Mining*, Chapman&Hall/CRC.

Sokal R.R., Rohlf F.J., 1962, *The comparison of dendrograms by objective methods*, "Takson" no. 2, pp. 33-40.

Szekely G., Rizzo M., 2005, *Hierachical clustering vie Joit between-within distances: extending Ward's minimum variance method*, "Journal of Classification", vol. 22, pp. 151-183.

Walesiak M., Gatnar E., 2009, *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa.

Walesiak M., Dudek M., 2012, *Package 'clusterSim' in R project*, http://keii.ue.wroc.pl/clusterSim/index.html.

Ward J.H., 1963, *Hierarchical grouping to optimize an objective function*, "Journal of the American Statistical Association", no. 58, pp. 236-244.

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.12.2 (2011-02-25) R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

THE WARD METHOD IN THE APPLICATION FOR CLASSIFICATION OF POLISH VOIVODESHIPS WITH DIFFERENT DISTANCES

Summary: This paper proposes to compare the results of the classification of Polish voivodeships with different objective function. It was attempted to perform the study on the basis of empirical justification for the selection of the similarity measures. The results obtained using the classical method of Ward and the algorithm proposed by the Lance-Williams formula were compared. It seems that the use of different distance measures in the classification of Polish voivodeships using the Word method gives comparable classification quality. Against the background of the proposed distance the weighted Euclidean distance is distinguished.

Keywords: Ward's method, cluster validity, dissimilarity measure, hierarchical agglomerative method.