PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Krzysztof Najman
Uniwersytet Gdański

SYMULACYJNA ANALIZA WPŁYWU WYBORU KRYTERIUM OPTYMALNOŚCI PODZIAŁU OBIEKTÓW NA JAKOŚĆ UZYSKANEJ KLASYFIKACJI W ALGORYTMACH *K*-ŚREDNICH

1. Wstęp

Metody podziałowe należą do najstarszych metod analizy skupień. Ich idea polega na tym, że wśród wszystkich możliwych podziałów obiektów na skupienia poszukuje się takiego, który maksymalizuje ich podobieństwo w skupieniach i maksymalizuje niepodobieństwo między skupieniami. Następuje podział zbioru n obiektów na określoną liczbę K rozłącznych skupień w taki sposób, aby każdy obiekt był przydzielony tylko do jednego skupienia. Podział taki jest nazywany ***K*-podziałem**. Optymalnego[1] podziału poszukuje się iteracyjnie, przesuwając obiekty między tworzonymi skupieniami, stąd dawniej metody te nazywano iteracyjnymi algorytmami transferu (*iterative relocation algorithm*) lub też technikami przemieszczeń.

Liczba wszystkich możliwych podziałów n obiektów na K skupień jest bardzo duża (zob. [Gordon 1981]). Aby przeszukać przestrzeń możliwych podziałów zaledwie 19 obiektów na 8 skupień, należy dokonać ok. $1{,}7\times10^{12}$ podziałów. Dla współczesnych problemów, gdzie często grupuje się tysiące obiektów, możliwych podziałów jest praktycznie nieskończenie wiele. Stąd istnieje potrzeba znalezienia algorytmu pozwalającego zoptymalizować podział obiektów w możliwie małej liczbie badanych podziałów.

Pierwsza propozycja[2] rozwiązania tego problemu pojawiła się w roku 1953, kiedy R.L. Thorndike opublikował artykuł *Who belongs in the family?*[3]. Thorndike zauważył, że aby rozpocząć poszukiwanie optymalnego podziału, należy zdefinio-

[1] Optymalizacja wymaga podania kryterium optymalności. Formalne kryteria optymalności podziału będą podstawą badań w dalszej części artykułu.

[2] Warto tu dodać, że większość monografii poświęconych analizie skupień artykułu R.L. Thorndike'a nie zauważa (np. [Anderberg 1973; Gordon 1999]).

[3] Zob. [Thorndike 1953].

wać kryterium jego optymalności. Zaproponował, aby tym kryterium była maksymalna wartość wariancji międzygrupowej, zdefiniowana następująco:

$$SK_M = SK_C - \sum_{i=1}^{K} SK_W, \tag{1}$$

gdzie wariancja międzygrupowa SK_M jest zdefiniowana jako różnica wariancji[4] całkowitej SK_C i sumy zmienności wewnątrz skupień SK_W:

$$SK_C = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \bar{x}_j\right)^2, \tag{2}$$

$$SK_{W(g)} = \sum_{i=1}^{n_g}\sum_{j=1}^{p}\left(x_{ijg} - \bar{x}_{jg}\right)^2. \tag{3}$$

Wskazał także, że aby zbudować algorytm podziału obiektów na skupienia, należy ustalić wstępne, prototypowe centra skupień. Zaproponował, aby tymi centrami stały się obiekty najbardziej od siebie oddalone, co powinno być ustalone na podstawie macierzy kwadratów odległości euklidesowych. Pozostał jeszcze jeden ważny aspekt. Należy rozstrzygnąć, w którym momencie trzeba przenieść obiekt do innego skupienia. Czy powinno to nastąpić po zmierzeniu zmiany wariancji międzygrupowej po przesunięciu danego obiektu do kolejnych skupień, czy w danej iteracji przesunięty powinien być tylko jeden obiekt, którego przesunięcie w największym stopniu zwiększa wariancję międzygrupową? R.L. Thorndike uznał, że pierwszy wariant będzie właściwy. Ostatecznie algorytm Thorndike'a można zapisać następująco: 1) ustal liczbę skupień K; 2) wyznacz macierz kwadratów odległości euklidesowych między wszystkimi obiektami; 3) znajdź K obiektów najbardziej od siebie oddalonych, będą one stanowić wstępne centra skupień; 4) wyznacz międzygrupową sumę kwadratów odległości euklidesowych; 5) przesuń obiekt do każdego z K skupień i każdorazowo wyznacz międzygrupową sumę kwadratów odległości euklidesowych; pozostaw obiekt w skupieniu, dla którego suma ta jest największa; 6) przelicz nowe centra skupień; 7) powtórz punkty 4 i 5 dla każdego obiektu; 8) powtarzaj punkty 4, 5, 6 aż dla żadnego podziału międzygrupowa suma kwadratów nie będzie wzrastać.

Dalsze badania toczyły się często niezależnie w wielu ośrodkach na całym świecie. Podstawowy algorytm był wielokrotnie modyfikowany[5]. Badania te prowadzone są także współcześnie. Główne nurty badań skupiają się na najważniejszych elementach algorytmu:

1) kolejności obiektów w zbiorze danych (*różne metody porządkowania*),

[4] Dowód, że wariancję można przedstawić w kategoriach odległości, przedstawili w roku 1965 A.W.F. Edwards i L. Cavalli-Sforza.

[5] Zob. [Anderberg 1973; Dagnelie 1975; Gordon A.D. 1999].

2) wyborze kryterium optymalności podziału (*budowa miar heterogeniczności i izolacji: wewnątrzgrupowa suma kwadratów, suma odległości, suma odchyleń od centrum, średnica, miary oddzielenia, przecięcia*),

3) ustaleniu wstępnych centrów skupień (*K-pierwszych, K-najbardziej odległych, K-losowych, różne metody losowania obiektów),*

4) wyeliminowaniu wpływu wartości nietypowych na wynik grupowania (*testowanie występowania wielowymiarowych wartości nietypowych),*

5) przyspieszeniu zbieżności algorytmu.

Dziś w literaturze te modyfikacje podstawowego algorytmu, w których centrum skupienia jest wyznaczane jako abstrakcyjny punkt będący jego środkiem, nazywa się metodą *K*-średnich.

Celem prezentowanych badań jest analiza wpływu wyboru kryterium grupowania na jakość uzyskanej klasyfikacji. Ocenie zostanie poddane 8 kryteriów. Oceniany będzie także wpływ 4 różnych metod początkowego ustalenia centrów skupień, łącznie z różnymi uporządkowaniami obiektów w zbiorach, na jakość uzyskanej klasyfikacji.

2. Kryteria optymalności podziału

Do klasycznych (wewnętrznych) kryteriów optymalności podziału obiektów w algorytmie *K*-średnich można zaliczyć[6]: miary oparte na macierzy odległości (minimalizacja sumy kwadratów odległości między obiektami w skupieniach, maksymalizacja sumy kwadratów odległości między obiektami z różnych skupień) i miary oparte na macierzy danych[7] (trace(W), trace(cov(W)), det(W)). Mierniki te są stosowane jako wewnętrzne kryterium optymalności podziału obiektów metody *K*-średnich. Stosowanie tych prostych kryteriów może być przyczyną krytyki samej metody, ponieważ w efekcie grupowania uzyskuje się skupienia będące hiperkulami. Nie można więc rozdzielić metodą *K*-średnich zbiorów, które w dowolnym wymiarze są nieseparowalne liniowo.

Warto zwrócić uwagę, że istnieje także wiele zewnętrznych kryteriów optymalności podzialu. Uzyskana struktura grupowa może być oceniona po pogrupowaniu obiektów. Do tego celu może posłużyć wiele wskaźników. Z doświadczeń autora[8] wynika, że szczególnie kilka z nich może charakteryzować się dużym potencjałem poprawy jakości grupowania. Są to[9]: indeks Daviesa-Bouldina, indeks Silhouette, indeks Calinskiego-Harabasza, indeks Hartigana, trace($W^{-1}B$), det(T)/det(W). Wskaźniki te, wraz z formalizmem matematycznym, były szczegó-

[6] Zob. [Ball, Hall 1965].

[7] Niektóre własności tych wskaźników w ustalaniu optymalnej liczby skupień zostały szerzej omówione w pracy [Najman, Najman 2006].

[8] Wiele z nich jest syntetycznie omówionych w pracach [Milligan, Cooper 1985; Najman, Najman 2005; 2006; Najman 2007].

[9] Gdzie: W – wewnętrzna macierz rozrzutu (*pooled-within groups scatter matrix*), B – międzygrupowa macierz rozrzutu (*between groups scatter matrix*), T – łączna macierz rozrzutu (*total scatter matrix*).

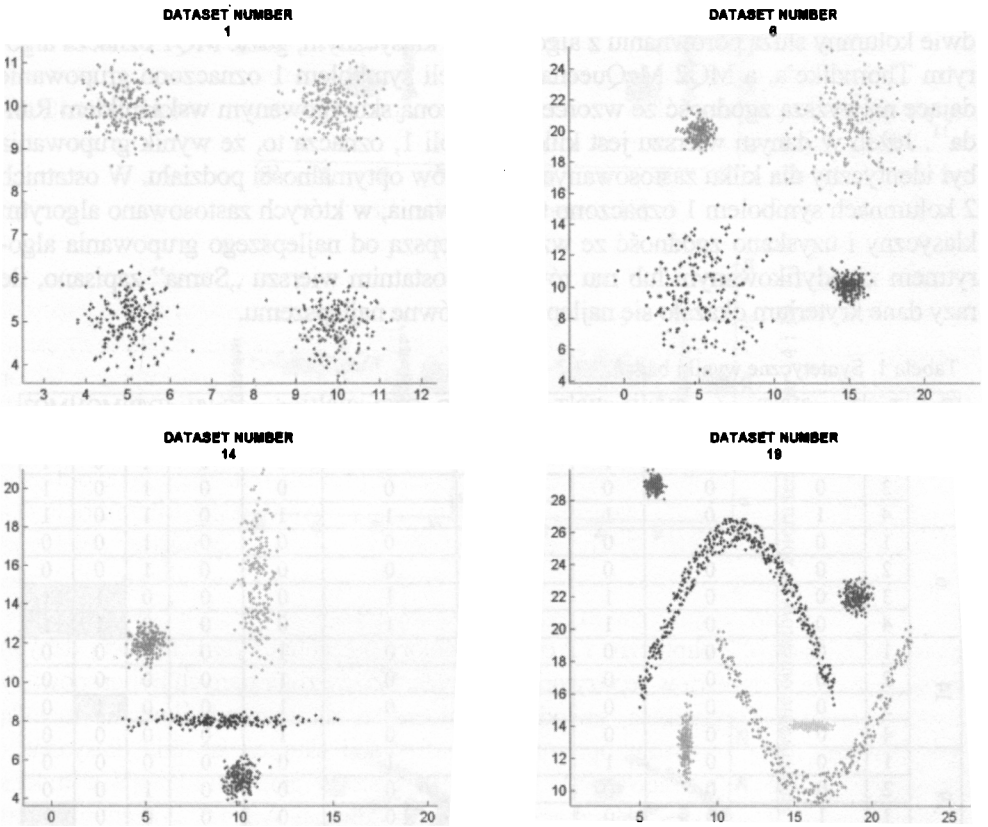
łowo opisane, a ich przydatność w ocenie jakości grupowania przebadana w pracach [Najman, Najman 2005; 2006; Najman 2007]. W niniejszej pracy ich zapis zostanie pominięty. Wydaje się, że wskaźniki te mogą potencjalnie stanowić wewnętrzne kryterium optymalności podziału obiektów w metodzie *K*-średnich. Autor stawia więc tezę: jeżeli powyższe wskaźniki charakteryzują się wysoką zdolnością do prawidłowego opisu jakości struktury grupowej (kryterium zewnętrzne), to powinny ją także prawidłowo wskazywać w procesie grupowania metodą *K*-średnich (będąc kryterium wewnętrznym).

3. Eksperyment badawczy

W celu zweryfikowania powyższej tezy zostało zaprojektowane badanie. Przygotowano 4 zbiory danych o znanej strukturze grupowej, charakteryzujących się różnymi konfiguracjami skupień. Zbiór pierwszy to zbiór kontrolny – niezależnie od przyjętego kryterium metoda *K*-średnich powinna bezbłędnie podzielić obiekty na skupienia. Zbiór drugi charakteryzuje się tym, że maksymalne odległości między obiektami w „dużych" skupieniach są takie same jak odległości między centrami wszystkich skupień. Zbiór trzeci złożony jest z 2 skupień owalnych i 2 skupień elipsoidalnych. Minimalne odległości wzdłuż krótszej osi w skupieniach elipsoidalnych są równe maksymalnym odległościom w skupieniach owalnych. Jednocześnie maksymalne odległości między obiektami w skupieniach elipsoidalnych wzdłuż dłuższej osi są większe niż największe odległości między obiektami ze skupień owalnych. Zbiór czwarty zawiera 6 skupień, w tym 2 nieseparowalne liniowo, 2 eliptyczne i 2 owalne. Zbiory prezentuje rys. 1. Numery danych odnoszą się do ich numeru w repozytorium danych. Każdy zbiór danych był grupowany klasyczną metodą Thorndike'a, metodą McQueena, a także zmodyfikowaną metodą uwzględniającą jako kryterium optymalnej konfiguracji skupień kryteria klasyczne i proponowane: trace(W), trace(Cov(W)), det(W), trace($W^{-1}B$), det(T)/det(W), indeks Hartigana, indeks Calinskiego-Harabasza, Daviesa-Bouldina. W prezentowanym badaniu zrezygnowano z testowania indeksu Silhouette ze względu na znaczny czas obliczeń wymagany do jego wyznaczenia. Aby istniała możliwość porównań działania algorytmu, ustalono arbitralnie, że jakość uzyskanej struktury grupowej będzie testowana po maksymalnie 10 iteracjach przesuwania wszystkich obiektów do skupień. Zgodność uzyskanego grupowania ze znanym wzorcem testowano indeksami: skorygowanym wskaźnikiem Randa, Jaccarda i Fowklesa-Mallowsa.

Ponieważ znana jest wrażliwość algorytmu *K*-średnich na początkową konfigurację centrów skupień, każdorazowo testowano 4 procedury wyznaczania prototypowych centów skupień i konfiguracje obiektów: losowe centra skupień będące obiektami, losowe przyporządkowanie obiektów do skupień, centra skupień rozłożone równomiernie według wartości zmiennych, losowe przyporządkowanie obiektów do skupień, obiekty uporządkowane według niemalejących wartości pierwszej zmiennej, centra skupień wybrane systematycznie, obiekty przyporządkowane do skupień według najmniejszej kwadratowej odległości euklidesowej, obiekty upo-

rządkowane według niemalejących wartości pierwszej zmiennej, centra skupień stanowi K pierwszych obiektów, obiekty przyporządkowane do skupień losowo.



Rys. 1. Zbiory obiektów testowych

Źródło: opracowanie własne.

Centra skupień zostały wyznaczone na poziomie środków ciężkości skupień i będą przeliczane każdorazowo po przeniesieniu obiektu do nowego skupienia. Dokumentacja badania była prowadzona tabelarycznie i graficznie.

4. Wyniki badań

Przedstawienie kompletu tablic wynikowych w publikacji jest niemożliwe ze względu na ich znaczną objętość[10]. Wyniki zostaną zaprezentowane syntetycznie w

[10] Komplet tablic znajduje się na stronie internetowej: http://panda.bg.univ.gda.pl/~Najman/grupowanie.html.

tab. 1. W tabeli w kolejnych wierszach zamieszczono wyniki dla zbiorów danych, numer repozytorium 1, 6, 14, 19. Do każdego z nich przyporządkowano 4 procedury ustalania prototypów centrów skupień i konfiguracji obiektów. Kolejne kolumny prezentują wyniki uzyskane dla grupowań zmodyfikowaną metodą *K*-średnich. Ostatnie dwie kolumny służą porównaniu z algorytmem klasycznym, gdzie MQ1 oznacza algorytm Thorndike'a, a MQ2 McQueena. W tabeli symbolem 1 oznaczono grupowanie dające najwyższą zgodność ze wzorcem, mierzoną skorygowanym wskaźnikiem Randa[11]. Jeżeli w danym wierszu jest kilka symboli 1, oznacza to, że wynik grupowania był identyczny dla kilku zastosowanych kryteriów optymalności podziału. W ostatnich 2 kolumnach symbolem 1 oznaczono te grupowania, w których zastosowano algorytm klasyczny i uzyskano zgodność ze wzorcem lepszą od najlepszego grupowania algorytmem zmodyfikowanym lub mu równą. W ostatnim wierszu „Suma" zapisano, ile razy dane kryterium okazało się najlepsze lub równe najlepszemu.

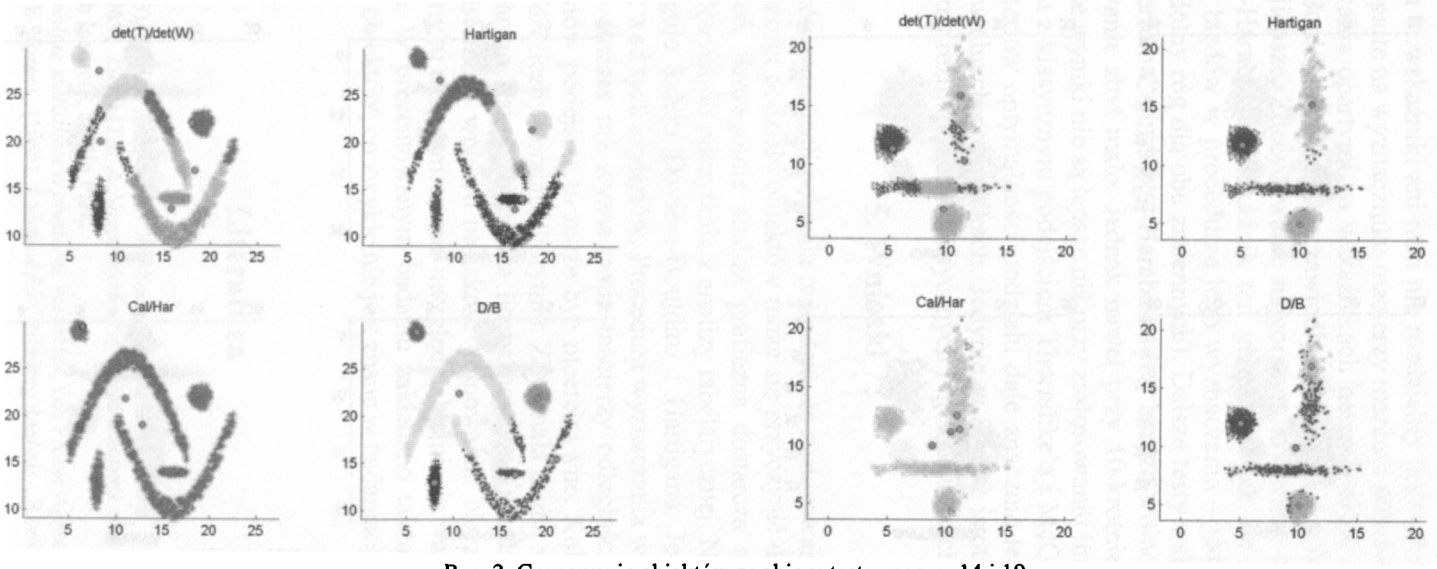
Tabela 1. Syntetyczne wyniki badań

Dane	Test	Trace(W)	Trace(cov(W))	Det(W)	Trace($W^{-1}B$)	Det(T)/Det(W)	Hartigan	Cal/Har	D/B	MQ1	MQ2
1	1	1	0	1	0	1	0	0	1	0	1
	2	1	0	1	0	1	1	0	1	0	1
	3	0	0	0	0	0	0	0	1	0	1
	4	1	0	1	0	1	1	0	1	0	1
6	1	0	0	0	0	0	0	0	1	0	0
	2	0	0	0	0	0	0	0	1	0	0
	3	0	0	1	0	1	0	0	0	1	1
	4	0	0	1	0	1	0	0	0	1	1
14	1	0	0	0	0	0	1	0	0	0	0
	2	0	0	0	0	0	1	0	0	0	0
	3	0	0	0	0	0	1	0	0	1	0
	4	0	0	0	0	0	1	0	0	0	0
19	1	0	0	1	0	1	0	0	0	0	0
	2	0	0	0	0	0	0	0	1	0	0
	3	1	0	0	0	0	0	0	0	0	1
	4	0	0	0	0	0	0	0	1	0	0
Suma		4	0	6	0	6	6	0	8	3	7

Źródło: opracowanie własne.

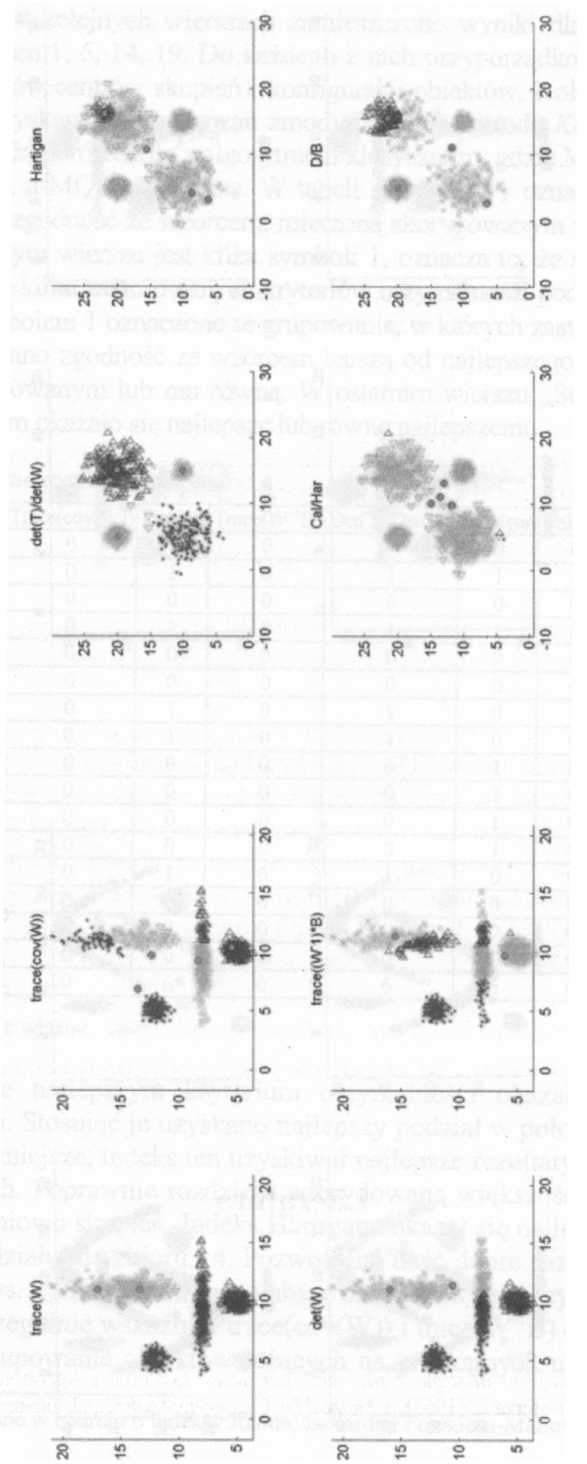
Zdecydowanie najlepszym kryterium optymalności okazało się kryterium Daviesa-Bouldina. Stosując je uzyskano najlepszy podział w połowie przypadków. Co jednak najważniejsze, indeks ten uzyskiwał najlepsze rezultaty także w najtrudniejszych zbiorach. Poprawnie rozdzielił zdecydowaną większość obiektów z nieseparowalnych liniowo skupień. Indeks Hartigana okazał się najlepszym kryterium optymalności podziału dla zbioru 14. Pozwolił na dość dobre rozdzielenie skupień elipsoidalnych (rys. 2). Zdecydowanie słabsze okazały się indeksy oparte na macierzy rozrzutu. Szczególnie wskaźniki trace(cov(W)) i trace($W^{-1}B$) charakteryzowały się własnością grupowania obiektów leżących na przekątnych układu współrzęd-

[11] Wyniki uzyskane w oparciu o indeksy Randa, Jaccarda i Fowklesa-Mallowsa były zgodne.



Rys. 2. Grupowanie obiektów ze zbioru testowego nr 14 i 19

Źródło: opracowanie własne.



Rys. 3. Grupowanie obiektów ze zbioru testowego nr 6 i 14

Źródło: opracowanie własne.

nych (rys. 3). Oba te wskaźniki ani razu nie rozdzieliły prawidłowo skupień. Wyniki grupowania oparte na wyznaczniku macierzy rozrzutu dobrze się uzupełniały z wynikami grupowania opartymi na wskaźnikach heterogeniczności skupień. Gdy indeksy Daviesa-Bouldina i Hartigana zawiodły, wskaźniki det(W) i det(T)/det(W) okazywały się najlepsze. Zdecydowanie najgorszym kryterium okazało się kryterium Calinskiego-Harabasza. Wskaźnik ten generował tak złe wyniki, że autor długo poszukiwał błędów w procedurze jego wyznaczania – jednak bez rezultatu (por. rys. 2 lewy dolny róg dla obu zmiennych). Dalsze testy wskazały, że metoda *K*-średnich z kryterium Calinskiego-Harabasza jest bardzo wolno zbieżna i 10 iteracji to zdecydowanie zbyt mało. Jednak nawet przy 10-krotnie większej liczbie iteracji uzyskiwane wyniki nie są lepsze niż przy zastosowaniu innych kryteriów.

W porównaniu z klasycznym podejściem Thorndike'a i McQueena zastosowanie badanych kryteriów optymalności podziału daje znacznie lepsze rezultaty we wszystkich nietrywialnych przypadkach. Jedynie w 3 z 12 istotnych przypadków uzyskano identyczny rezultat metod klasycznych i nieklasycznych.

5. Wnioski

Analiza wyników symulacji pozwala twierdzić, że zastosowanie nieklasycznych kryteriów optymalności podziału obiektów może się przyczynić do poprawy jakości uzyskanych skupień. Szczególnie analiza graficzna dostarcza szeregu ciekawych wniosków niemożliwych do odczytania z analizy tabelarycznej. Na szczególne zainteresowanie zasługuje indeks Daviesa-Bouldina i Hartigana. Jednocześnie należy wskazać na szereg wad tych kryteriów. Procedura wyznaczania wartości tych indeksów jest znacznie dłuższa niż wyznaczenia macierzy odległości. Dla grupowania dużej liczby obiektów podejście to może być nieefektywne. Kolejnym problemem jest ustalenie *a priori* liczby iteracji algorytmu. Zbyt mała liczba sprawia, że uzyskane skupienia są bardzo słabe. Zbyt duża liczba powoduje bardzo znaczny wzrost czasu analiz bez istotnego wpływu na jakość grupowania. Na zakończenie należy dodać, że opisywany tu eksperyment nie uwzględniał istnienia wartości nietypowych w zbiorze danych. W prezentowanym badaniu zakładano także znajomość liczby skupień w zbiorze obiektów, co zwykle nie jest znane w badaniach empirycznych.

Literatura

Anderberg M.R. (1973), *Cluster Analysis for Applications*, Academic Press.

Ball G., Hall D.J. (1965), *ISODATA, A Novel Method of Data Analysis and Pattern Classification*, Stanford Research Institute, Menlo Park.

Dagnelie P. (1975), *Analyse statistique à plusieres variables*, Les Presse Agronomique, Gambloux.

Edwards A.W.F., Cavalli-Sforza L. (1965), *A Method for Clustering Analysis*, „Biometrics", 21, s. 362-375.

Gordon A.D. (1981), *Classification. Methods for the Exploratory Analysis of Multivariate Data*, Chapman and Hall, London.

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

Milligan G.W., Cooper M.C. (1985), *An Examination of Procedures for Determining the Number of Clusters in Data Set*, „Psychometrika", 50(2), s. 159-179.

Milligan G.W. (1980), *An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms*, „Psychometrika", 45, s. 325-342.

Najman K., Najman K. (2005), *Analityczne metody ustalania liczby skupień*, [w:] Taksonomia 12, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław, s. 265-273.

Najman K., Najman K. (2006), *Analityczne metody ustalania liczby skupień w rozmytych zbiorach danych*, [w:] Taksonomia 13, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1126, AE, Wrocław, s. 159-167.

Najman K. (2007), *Metody ustalania liczby skupień w binarnych zbiorach danych*, [w:] Taksonomia 14, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław, s. 321-329.

Thorndike R.L. (1953),*Who Belongs in the Family?*, „Psychometrika", vol. 18.

THE INFLUENCE ANALYSIS OF SELECTION OF THE OPTIMUM PARTITIONING OBJECTS CRITERION FOR *K*-MEANS ALGORITHMS

Summary

The paper presents the modification of partitioning *k*-means algorithm. There are proposed the application of a nonclassical indices as the inner optimal criterion of the partitioning the objects: index Davies-Bouldin, index Calinski-Harabasz, index Hartigan, trace($W^{-1}B$) and det(T)/det(W). On the basis of simulation research critical opinions of quality of partitions that are obtained are maid. There are presented weak and strong points of modified k-maens algorithm. The results are compared with the classical Thorndike's and McQueen's algorithms.