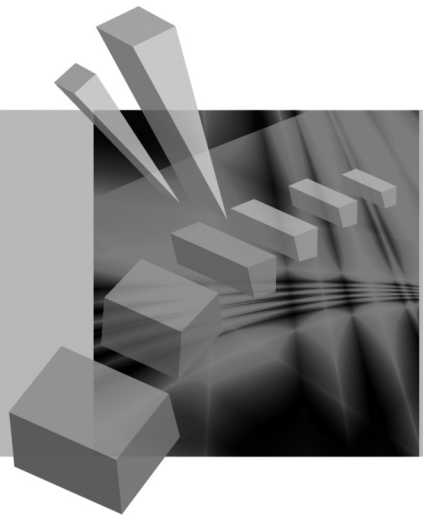
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w 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

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Dorota Rozmus
Uniwersytet Ekonomiczny w Katowicach

PORÓWNANIE DOKŁADNOŚCI TAKSONOMII SPEKTRALNEJ ORAZ ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA IDEI METODY *BAGGING*

Streszczenie: Kiedy stosuje się metody taksonomiczne w jakimkolwiek zagadnieniu klasyfikacji, ważną kwestią jest zapewnienie wysokiej poprawności wyników grupowania. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod. Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz algorytmy spektralne. Głównym celem tego artykułu jest porównanie dokładności zagregowanych algorytmów taksonomicznych opartych na idei metody *bagging* [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999] oraz spektralnego algorytmu taksonomicznego zaproponowanego przez Ng i in. [2001].

Słowa kluczowe: taksonomia, zagregowane algorytmy taksonomiczne, taksonomia spektralna, dokładność.

1. Wstęp

Kiedy stosuje się metody taksonomiczne w jakimkolwiek zagadnieniu klasyfikacji, ważną kwestią jest zapewnienie wysokiej dokładności wyników grupowania. Od niej bowiem zależeć będzie skuteczność wszelkich decyzji podjętych na ich podstawie. Przez pojęcie *dokładność grupowania* należy rozumieć zdolność metody do rozpoznawania rzeczywistej struktury klas. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod (np. *k*-średnich, metod hierarchicznych). Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz algorytmy spektralne. Taksonomia spektralna polega na zastosowaniu wartości własnych pochodzących ze spektralnej dekompozycji macierzy podobieństwa opisującej badane obiekty. Podejście zagregowane w taksonomii można natomiast sformułować następująco: mając wyniki wielokrotnie przeprowadzonego grupowania, znajdź zagregowany podział ostateczny.

Głównym celem tego artykułu jest porównanie dokładności zagregowanych i spektralnych algorytmów taksonomicznych. W badaniu pod uwagę wzięta zostanie tylko specyficzna klasa metod agregacji, która oparta jest na idei metody *bagging* [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999]. Natomiast jako algorytm spektralny zastosowana będzie metoda zaproponowana przez Ng i in. [2001].

2. Metoda *bagging* w taksonomii

Metoda *bagging* w taksonomii jest pewną ogólną ideą, w ramach której narodziło się kilka szczegółowych rozwiązań. Generalnie polega ona na losowaniu B prób bootstrapowych i dokonywaniu ich grupowania w celu uzyskania podziałów składowych, które będą agregowane. Różnice w poszczególnych rozwiązaniach polegają na zastosowaniu różnych operatorów agregacji.

Propozycja Leischa

Leisch [1999] zaproponował, by w pierwszym kroku na podstawie każdej podpróby bootstrapowej określane były rezultaty grupowania przy zastosowaniu tzw. bazowej metody taksonomicznej, którą jest jedna z metod iteracyjno-optymalizacyjnych, np. *k*-średnich. W kolejnym etapie ostateczne centra skupień przekształcane są w nowy zbiór danych obejmujący $B \times K$ obserwacji (K to liczba skupień w metodzie bazowej), który poddawany jest podziałowi za pomocą metod hierarchicznych. Uzyskany dendrogram jest podstawą ostatecznego podziału – obserwacje z pierwotnego zbioru przydzielane są do tej grupy, której środek ciężkości znajduje się w minimalnej odległości euklidesowej.

Szczegółowo algorytm zaproponowany przez Leischa przebiega w następujących krokach:

1. Z pierwotnego n-elementowego zbioru $G = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ należy wylosować B prób bootstrapowch $G_n^1, G_n^2, ..., G_n^B$, losując n obserwacji przy wykorzystaniu schematu losowania ze zwracaniem.

2. Na podstawie każdego podzbioru za pomocą metod iteracyjno-optymalizacyjnych (np. *k*-średnich) dokonuje się podziału na grupy obserwacji podobnych do siebie, uzyskując w ten sposób $B \times K$ zalążków skupień $c_{11}, c_{12}, ..., c_{1K}, c_{21}, ..., c_{BK}$ gdzie K oznacza liczbę skupień w metodzie bazowej, a c_{bk} jest k-tym zalążkiem znalezionym na podstawie podpróby G_n^b.

3. Niech zalążki skupień uzyskane na podstawie kolejnych prób bootstrapowych utworzą nowy zbiór danych $C^B = \{c_{11}, ..., c_{BK}\}$.

4. Do tak skonstruowanego zbioru należy zastosować hierarchiczną metodę taksonomiczną, uzyskując w ten sposób dendrogram.

5. Podział na grupy pierwotnego zbioru danych określany jest w ten sposób, że dendrogram uzyskany na podstawie zbioru C^B jest cięty na określonym przez

badacza poziomie, co prowadzi do uzyskania grup obiektów podobnych C_1^B, ..., C_m^B, gdzie $1 \le m \le BK$. Każda obserwacja x_i z pierwotnego zbioru danych *G* jest przydzielana do tej grupy, w której znajduje się najbliżej leżący zalążek $c(x_i)$.

Propozycja Dudoit i Fridlyand

Metoda *bagging* w wersji zaproponowanej przez Dudoit i Fridlyand [2003] stosuje algorytm iteracyjno-optymalizacyjny do oryginalnego zbioru danych i poszczególnych prób bootstrapowych. Następnie, po dokonaniu permutacji etykiet klas w poszczególnych podpróbach tak, by zachodziła jak największa zbieżność z podziałem obiektów z oryginalnego zbioru danych, stosuje głosowanie majoryzacyjne w celu określenia ostatecznego grupowania zagregowanego.

Kroki zaproponowanego przez nich algorytmu można ująć według poniższego schematu.

Dla założonej liczby klas *K*:

1. Zastosuj iteracyjno-optymalizacyjny algorytm taksonomiczny *T* do pierwotnego zbioru danych $G = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, uzyskując w ten sposób etykiety klas $T(x_i, G) = \hat{y}_i$ dla każdej obserwacji x_i, $i = 1, \ldots, n$.

2. Skonstruuj *b*-tą próbę bootstrapową $G_n^b = \{\mathbf{x}_1^b, ..., \mathbf{x}_n^b\}$.

3. Zastosuj metodę taksonomiczną *T* do skonstruowanej próby bootstrapowej G_n^b, uzyskując podział na klasy: $T(x_i^b, G_n^b)$ dla każdej obserwacji w zbiorze G_n^b.

4. Dokonaj permutacji etykiet klas przyznanych obserwacjom w próbie bootstrapowej G_n^b tak, by zachodziła jak największa zbieżność z klasyfikacją obiektów z oryginalnego zbioru danych *G*. Niech PR_K oznacza zbiór wszystkich permutacji zbioru liczb całkowitych *1*,…, *K*. Znajdź permutację $\tau^b \in PR_K$ maksymalizującą:

$$\sum_{i=1}^{n} I(\tau(T(x_i^b, G_n^b)) = T(x_i^b, G)), \tag{1}$$

gdzie $I(\cdot)$ to funkcja wskaźnikowa równa 1, gdy zachodzi prawda, 0 w przypadku przeciwnym.

5. Powtórz kroki 2-4 *B* razy. Ostatecznie zaklasyfikuj *i*-tą obserwację, stosując głosowanie majoryzacyjne, zatem przydzielając ją do tej klasy, dla której zachodzi:

$$\arg\max_{1 \le k \le K} \sum_{b: x_i \in G_n^b} I(\tau^b(T(x_i, G_n^b)) = k). \tag{2}$$

Propozycja Hornika

W metodzie tej po skonstruowaniu *B* prób bootstrapowych i zastosowaniu do nich algorytmu taksonomicznego uzyskuje się podziały składowe. Grupowanie za-

gregowane natomiast jest uzyskiwane za pomocą tzw. podejścia optymalizacyjnego, które ma za zadanie zminimalizować funkcję o postaci:

$$\sum_{b=1}^{B} dist(c, c_b)^2 \Rightarrow \min_{c \in C}, \tag{3}$$

gdzie: C – zbiór wszystkich możliwych podziałów zagregowanych,
$dist$ – odległość euklidesowa,
$(c_1, ..., c_B)$ – grupowania wchodzące w skład podziału zagregowanego.

3. Taksonomia spektralna

Taksonomia spektralna polega na zastosowaniu wartości własnych pochodzących ze spektralnej dekompozycji macierzy podobieństwa opisującej badane obiekty. Następnie największe wartości własne oraz odpowiadające im wektory własne są wykorzystywane do ostatecznego podziału obserwacji. W literaturze zaproponowano kilka metod spektralnych, a każda z nich w nieco inny sposób stosuje wektory własne [Kannan i in. 2004; Ng i in. 2001; Shi, Malik 2000]. W niniejszym badaniu zastosowana zostanie metoda zaproponowana przez Ng i in. [2001].

Dany jest zbiór obserwacji $G = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ w przestrzeni R^l, który należy podzielić na k grup.

1. Skonstruuj macierz podobieństwa (*affinity matrix*) $A \in R^{n \times n}$, której elementy są zdefiniowane jako:

$$A_{ij} = \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right), \tag{4}$$

gdy $i \neq j$ oraz $A_{ii} = 0$. σ to parametr skalujący dobierany przez badacza.

2. Zdefiniuj $\boldsymbol{D}$ jako macierz diagonalną, której element (i, i) jest sumą i-tego wiersza macierzy $\boldsymbol{A}$ i na jej podstawie skonstruuj macierz:

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{5}$$

3. Znajdź k pierwszych wektorów własnych $(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_k)$ macierzy $\boldsymbol{L}$ i zestawiając je w kolumny, skonstruuj macierz:

$$\mathbf{Z} = \left[\mathbf{z}_1, ..., \mathbf{z}_k\right] \in R^{n \times k}. \tag{6}$$

4. Skonstruuj macierz $\boldsymbol{Y}$ przez normalizację każdego wiersza macierzy $\boldsymbol{Z}$ tak, by miały jednakową długość, tj.:

$$y_{ij} = z_{ij} / \left(\sum\nolimits_j z_{ij}^2\right)^{1/2}. \tag{7}$$

5. Traktując każdy wiersz macierzy $\boldsymbol{Y}$ jako punkt w przestrzeni R^k, podziel je na k grup z zastosowaniem metody k-średnich (lub innej).

6. Ostatecznie przydziel każdą pierwotną obserwację x_i do j-tej grupy wtedy i tylko wtedy, gdy i-ty wiersz macierzy $\boldsymbol{Y}$ został przydzielony do j-tej grupy.

4. Badania empiryczne

W celu porównania dokładności grupowania badanych metod zastosowano miarę opartą na indeksie Randa:

$$Acc = \frac{1}{Z}\sum_{z=1}^{Z} R(P_z, P^T), \tag{8}$$

gdzie: Z – liczba badanych podziałów,
R – indeks Randa,
P_z – grupowanie na podstawie z-tego podziału,
P^T – rzeczywiste etykiety klas.

Miara ta jest uśrednioną po wszystkich badanych podziałach miarą dokładności i ocenia podobieństwo między ostatecznym grupowaniem zagregowanym a prawdziwymi etykietami klas.

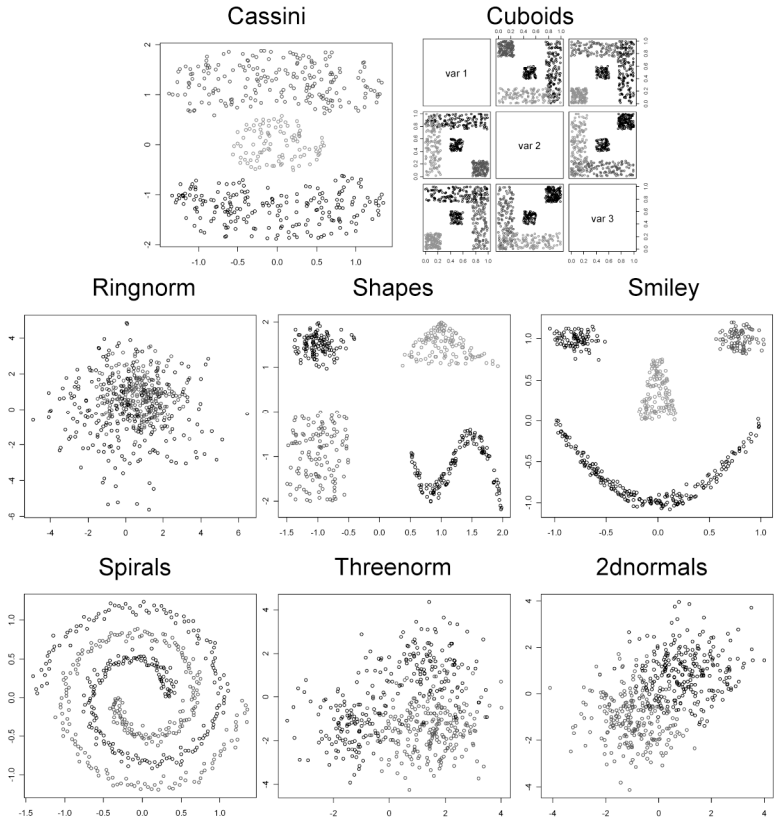
W badaniach zastosowano sztucznie generowane zbiory danych, które standardowo wykorzystywane są w badaniach porównawczych w taksonomii[1]. Są to takie zbiory, w których przynależność obiektów do klas jest znana. Ich krótka charakterystyka znajduje się w tab. 1, natomiast struktura przedstawiona jest na rys. 1.

Tabela 1. Charakterystyka zastosowanych zbiorów danych

Zbiór danych	Liczba obiektów	Liczba cech	Liczba klas
Cassini	500	2	3
Cuboids	500	3	4
Ringnorm	500	2	2
Shapes	500	2	4
Smiley	500	2	4
Spirals	500	2	2
Threenorm	500	2	2
2dnormals	500	2	2

Źródło: opracowanie własne.

[1] Zbiory zaczerpnięte zostały z pakietu `mlbench` z programu **R**.



Rys. 1. Struktura zastosowanych zbiorów danych

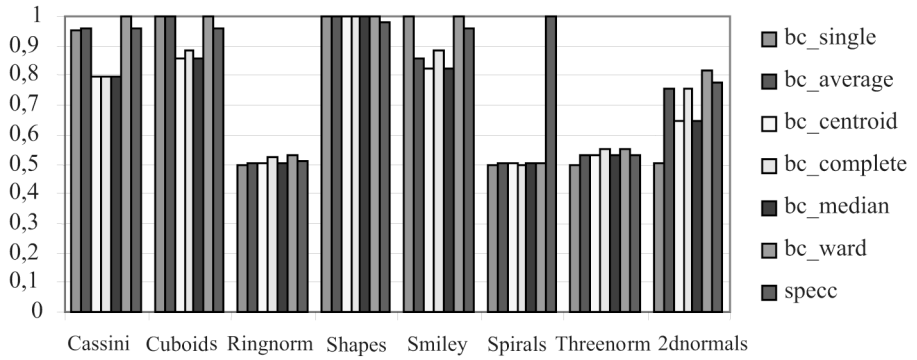
Źródło: opracowanie własne na podstawie programu **R**.

W metodzie *bagging* według Leischa jako metodę bazową zastosowano metodę *k*-średnich, natomiast ostatecznego grupowania dokonano z zastosowaniem: metody najbliższego sąsiedztwa (*single*), najdalszego sąsiedztwa (*complete*), średniej odległości między skupieniami (*average*), środka ciężkości (*centroid*), mediany (*median*), Warda (*Ward*). W metodzie Dudoit i Fridlyand oraz Hornika utworzono 50 prób bootstrapowych, a na ich podstawie określano podziały składowe z zastosowaniem metody *k*-średnich oraz *c*-średnich, która jest rozmytą wersją metody *k*-średnich opracowaną przez Bezdeka [1981]. Natomiast agregacja przebiegała z zastosowaniem równania 2 w metodzie Dudoid i Fridlyand oraz 3 w metodzie Hornika[2].

[2] Na rysunkach 3 i 4 stosowano skróty cl_bagg_k i cl_consensus_k , jeżeli grupowania składowe określane były z zastosowaniem metody *k*-średnich, oraz cl_bagg_c i cl_consensus_c, gdy wykorzystywano metodę *c*-średnich.

W taksonomii spektralnej macierz ***Y*** grupowana była z zastosowaniem metody *k*-średnich.

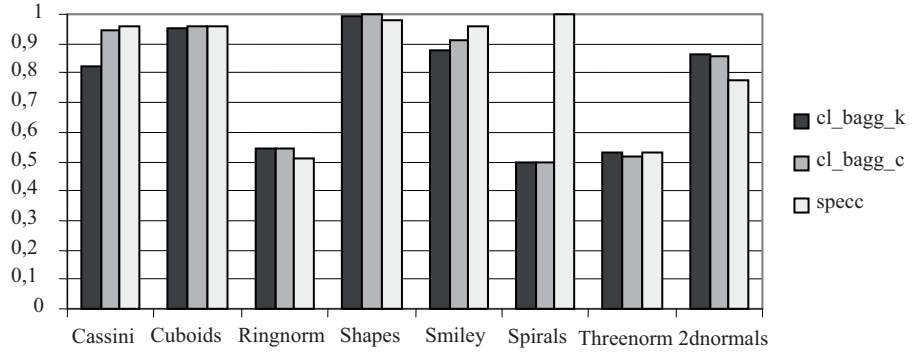
Każde podejście stosowano 50 razy i ich wyniki były potem badane pod względem dokładności.



Rys. 2. Porównanie dokładności metody *bagging* według Leischa oraz podejścia spektralnego

Źródło: opracowanie własne.

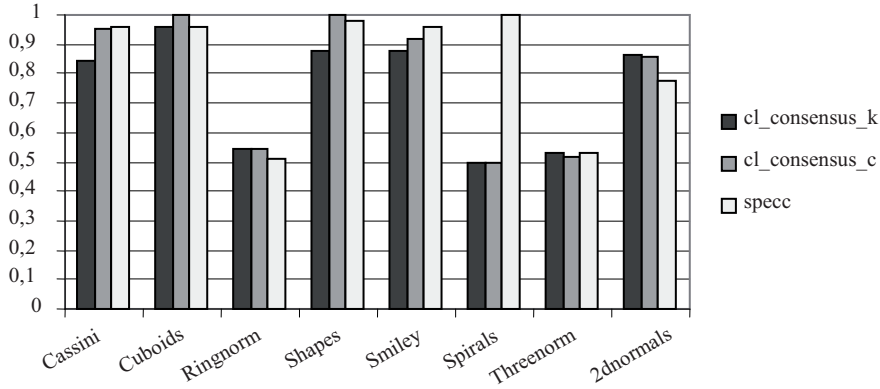
W przypadku metody *bagging* według Leischa (rys. 2) można zaobserwować, że taksonomia spektralna (specc) zawsze daje niższą dokładność niż metoda bc_ward, z wyjątkiem zbioru *Spirals*. Dla zbiorów *Cassini*, *Cuboids*, *Smiley* oraz *2dnormals* specc daje wyższą dokładność niż zagregowane metody taksonomiczne bc_centroid, bc_complete oraz bc_median. Porównywalną dokładność taksonomii spektralnej z metodami zagregowanymi można zaobserwować dla zbiorów *Ringnorm*, *Shapes* oraz *Threenorm*. Natomiast dla zbioru *Spirals* taksonomia spektralna daje znacznie lepszą dokładność w porównaniu z metodami zagregowanymi.



Rys. 3. Porównanie dokładności metody *bagging* według Duidoit i Fridlyand oraz podejścia spektralnego

Źródło: opracowanie własne.

Rysunek 3 porównujący rezultaty dla metody *bagging* według Dudoit i Fridlyand oraz taksonomii spektralnej pozwala stwierdzić, że specc daje lepsze rezultaty niż obydwa rozpatrywane warianty zagregowane tylko w przypadku zbiorów *Cassini*, *Smiley* i *Spirals*. Niższą dokładność taksonomii spektralnej w porównaniu z metodami zagregowanymi można zaobserwować dla zbiorów *Ringnorm*, *Shapes* i *2dnormals*. Natomiast dla zbiorów *Cuboids* i *Threenorm* obydwa podejścia dają porównywalne rezultaty.



Rys. 4. Porównanie dokładności metody *bagging* według Hornika oraz podejścia spektralnego

Źródło: opracowanie własne.

W przypadku metody *bagging* według Hornika i taksonomii spektralnej (rys. 4) można stwierdzić, że specc daje zdecydowanie wyższą dokładność dla zbioru *Spirals* i nieco lepszą dla zbiorów *Smiley* i *Cassini*. Dla pozostałych zbiorów dokładność specc można określić jako porównywalną lub nieco gorszą niż dokładność metod zagregowanych.

5. Podsumowanie

W świetle uzyskanych wyników można wyciągnąć następujące wnioski:

1. Dla zbiorów o strukturze podobnej do zbioru *Spirals* taksonomia spektralna jest bezkonkurencyjna z punktu widzenia dokładności.
2. Warto zastąpić zagregowane metody bc_complete, bc_centroid oraz bc_median przez taksonomię spektralną, bo na pewno uzyskamy nie gorszą dokładność.
3. Dla zbiorów z wyraźnie separowalnymi klasami (np. *Cassini*, *Smiley*) taksonomia spektralna niejednokrotnie jest bardziej dokładna niż metody zagregowane cl_bag_k oraz cl_consensus_k.
4. Dla zbiorów z trudno separowalnymi klasami (np. *Ringnorm*, *Threenorm, 2dnormals*) taksonomia spektralna daje porównywalne lub nieco gorsze rezultaty niż podejście zagregowane.

Literatura

Bezdek J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York 1981.

Dudoit S., Fridlyand J., *Bagging to improve the accuracy of a clustering procedure*, "Bioinformatics" 2003, vol. 19, no 9.

Hornik K., *A CLUE for CLUster ensembles*, "Journal of Statistical Software" 2005, no 14.

Kannan R., Vempala S., Vetta A., *On clustering – good, bad and spectral*, "Journal of the ACM" 2004, vol. 51, no 3.

Leisch F., *Bagged clustering*, „Adaptive Information Systems and Modeling in Economics and Management Science", Working Papers, SFB, 1999, no 51.

Ng A.Y., Jordan M.I., Weiss Y., *On spectral clustering: Analysis and an algorithm,* "Advances in Neural Information Processing Systems" 2001.

Shi J., Malik J., *Normalized cuts and image segmentation*, "IEEE Transactions on Pattern Analysis and Machine Intelligence" 2000, vol. *22*, no 8, http://www-2.cs.cmu.edu/»jshi/Grouping/.

COMPARISON OF ACCURACY OF SPECTRAL CLUSTERING AND CLUSTER ENSEMBLES STABILITY BASED ON BAGGING IDEA

Summary: High accuracy of the results is a very important task in any grouping problem (clustering). Therefore in the literature there are proposed methods and solutions that main aim is to give more accurate results than traditional clustering algorithms. The examples of such solutions can be cluster ensembles or spectral clustering algorithms. The main aim of the article is to compare the accuracy of spectral clustering and cluster ensembles. There will be considered cluster ensembles based on bagging idea [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999] and spectral algorithm proposed by Ng et al. [2001].

Keywords: taxonomy, cluster ensemble, spectral clustering, accuracy.