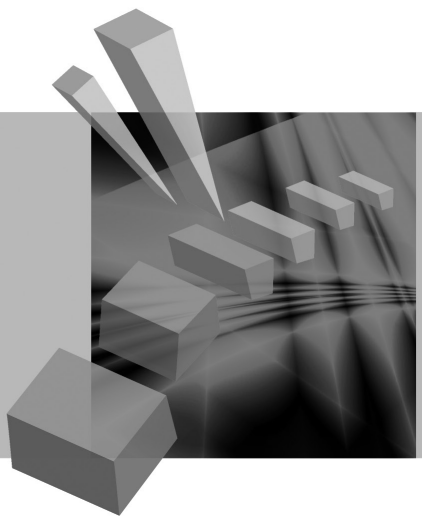
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

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

Summaries

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

Dorota Rozmus
Uniwersytet Ekonomiczny w Katowicach

PORÓWNANIE DOKŁADNOŚCI TAKSONOMICZNEJ METODY PROPAGACJI PODOBIEŃSTWA ORAZ ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA IDEI METODY *BAGGING*

Streszczenie: Podczas stosowania metod taksonomicznych w jakimkolwiek zagadnieniu klasyfikacji ważną kwestią jest zapewnienie wysokiej poprawności wyników grupowania. Od niej bowiem zależeć będzie skuteczność wszelkich decyzji podjętych na ich podstawie. 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, czyli łączenia wyników uzyskanych w wyniku wielokrotnego grupowania (*cluster ensemble*) oraz taksonomiczna metoda propagacji podobieństwa (*affinity propagation clustering*). Głównym celem artykułu jest porównanie dokładności zagregowanych algorytmów taksonomicznych opartych na idei metody *bagging* oraz taksonomicznej metody propagacji podobieństwa.

Słowa kluczowe: taksonomia, podejście zagregowane, taksonomiczna metoda propagacji podobieństwa, dokładność grupowania.

1. Wstęp

Stosowanie metod taksonomicznych w jakimkolwiek zagadnieniu grupowania wymaga jednocześnie zapewnienie wysokiej dokładności wyników podziału. Ona bowiem warunkuje skuteczność wszelkich decyzji podjętych na podstawie uzyskanych rezultatów. Przez pojęcie *dokładność grupowania* należy rozumieć zdolność metody do rozpoznawania rzeczywistej struktury klas. Dlatego też w literaturze wciąż proponowane są nowe rozwiązania, których zadaniem jest poprawa dokładności grupowania w stosunku do tradycyjnie stosowanych metod (np. *k*-średnich, hierarchicznych). Przykładami mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz stosunkowo niedawno zaproponowana metoda propagacji podobieństwa. Podejście zagregowane w taksonomii można sformułować następująco:

dysponując wynikami wielokrotnie przeprowadzonego grupowania, znajdź zagregowany ostateczny sposób podziału. Metoda propagacji podobieństwa natomiast to metoda, która wśród wszystkich obiektów w zbiorze danych przesyła odpowiednie informacje i w ten sposób identyfikuje tzw. reprezentantów, wokół których tworzy grupy obiektów podobnych do siebie.

Głównym celem tego artykułu jest porównanie dokładności taksonomicznej metody propagacji podobieństwa [Frey, Dueck 2007] oraz zagregowanych algorytmów taksonomicznych opartych na idei metody *bagging* [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999].

2. Metoda *bagging* w taksonomii

Metoda *bagging* w taksonomii jest pewną ogólną koncepcją, w ramach której narodziło się kilka szczegółowych rozwiązań. Pierwszy etap we wszystkich algorytmach jest taki sam – polega 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

W algorytmie zaproponowanym przez Leischa [1999] w pierwszym kroku na podstawie każdej podpróby bootstrapowej określane są rezultaty grupowania przy zastosowaniu tzw. bazowej metody taksonomicznej, którą jest jedną 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 najbliżej.

Szczegółowo algorytm zaproponowany przez Leischa przebiega w następująco:

1. Z pierwotnego *n*-elementowego zbioru $G = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ wylosuj *B* prób bootstrapowych $G_n^1, G_n^2, ..., G_n^B$, losując za każdym razem *n* obserwacji przy wykorzystaniu schematu losowania ze zwracaniem.

2. Na podstawie każdego podzbioru za pomocą metod iteracyjno-optymalizacyjnych (np. *k*-średnich) dokonaj 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. Zalążki skupień uzyskane na podstawie kolejnych prób bootstrapowych przekształć w nowy zbiór danych $C^B = \{c_{11}, ..., c_{BK}\}$.

4. Do tak skonstruowanego zbioru zastosuj hierarchiczną metodę taksonomiczną, uzyskując w ten sposób dendrogram. Dokonując cięcia dendrogramu na określonym poziomie, uzyskuje się grupy obiektów podobnych $C_1^B, ..., C_m^B$, gdzie $1 \le m \le BK$.

5. Każdą obserwację x_i z pierwotnego zbioru danych *G* przydziel do tej grupy, w której znajduje się najbliżej leżący zalążek $c(x_i)$, uzyskując w ten sposób ostateczny sposób podziału.

Propozycja Dudoit i Fridlyand

W metodzie *bagging* w wersji zaproponowanej przez Dudoit i Fridlyand [2003] stosuje się algorytm iteracyjno-optymalizacyjny do oryginalnego zbioru danych i poszczególnych prób bootstrapowych. Następnie dokonuje się permutacji etykiet klas w poszczególnych podpróbach tak, by zachodziła jak największa zbieżność z podziałem obiektów z oryginalnego zbioru danych. Ostatni krok to zastosowanie głosowania majoryzacyjnego w celu określenia grupowania zagregowanego.

Poszczególne 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 podziałem 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 zagregowane 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* to zbiór wszystkich możliwych podziałów zagregowanych, *dist* – odległość Euklidesowa, $(c_1, ..., c_B)$ – grupowania wchodzące w skład podziału zagregowanego.

3. Taksonomiczna metoda propagacji podobieństwa

Frey i Dueck [2007] opisują tę metodę (*affinity propagation)* jako algorytm, który wśród obiektów w zbiorze danych identyfikuje tzw. reprezentantów (*exemplars*) i wokół nich tworzy grupy obiektów podobnych. Metoda ta działa poprzez jednoczesne rozpatrywanie wszystkich obiektów w zbiorze jako potencjalnych reprezentantów. Wymieniając informacje pomiędzy obiektami, działa aż do momentu, gdy zostanie znaleziony odpowiedni zbiór reprezentantów i odpowiadający mu podział obiektów. Celem metody jest maksymalizacja sumy podobieństw między obiektami i ich reprezentantami.

Poszczególne etapy metody propagacji podobieństwa można przedstawić w następujących krokach:

1. Określenie macierzy podobieństw między obiektami przy zastosowaniu ujemnego kwadratu odległości euklidesowej:

$$s(i,k) = -\left\| x_i - x_k \right\|^2. \tag{4}$$

2. Ustalenie tzw. preferencji (*preferences*):

$$s(k,k) = p, \tag{5}$$

które dla każdej obserwacji wskazują tendencję do tego, by stała się ona reprezentantem.

3. Między obserwacjami wymieniane są dwa rodzaje informacji:

a. Odpowiedniość *r*(*i, k*) (*responsibility*) przesyłana jest od obserwacji x_i do potencjalnego reprezentanta x_k i odzwierciedla, jak bardzo obserwacja otrzymująca informację powinna być reprezentantem dla obserwacji wysyłającej informację, biorąc pod uwagę informację płynącą od innych punktów będących potencjalnymi reprezentantami dla obserwacji x_i.

b. Osiągalność $a(i, k)$ (*availability*) przesyłana jest od potencjalnego reprezentanta x_k do obserwacji x_i i odzwierciedla, jak bardzo obserwacja wysyłająca informację powinna być reprezentantem dla obserwacji otrzymującej informację, biorąc pod uwagę informację płynącą od innych obserwacji, dla których x_k jest potencjalnym reprezentantem.

4. Inicjując algorytm, przyjmuje się, że początkowe wartości osiągalności przyjmują wartość $a(i, k) = 0$.

5. Następnie wartości odpowiedniości i osiągalności obliczane są z formuł:

$$r(i,k) \leftarrow s(i,k) - \max_{k':k' \neq k} \{a(i,k') + s(i,k')\}, \tag{6}$$

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum_{i':i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}, & \text{gdy } i \neq k \\ \sum_{i':i' \neq i} \max\{0, r(i',k)\}, & \text{gdy } i = k \end{cases}. \tag{7}$$

6. Procedura przesyłania informacji może zostać zakończona po:
- określonej liczbie iteracji;
- jeśli zmiana w przesyłanej informacji będzie mniejsza niż jakaś z góry ustalona wartość;
- gdy przez określoną liczbę iteracji nie ma zmian w przesyłanej informacji.

7. Przydział obiektów do skupień $\hat{\mathbf{c}} = (\hat{c}_1, ..., \hat{c}_N)$ odbywa się według reguły:

$$\hat{c}_i = \arg\max_k [a(i,k) + r(i,k)], \tag{8}$$

gdzie $\hat{c}_i$ jest reprezentantem skupienia, do którego jest przydzielona obserwacja x_i.

4. Badania empiryczne

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 grup jest z góry znana. Ich krótka charakterystyka znajduje się w tab. 1, natomiast struktura przedstawiona jest na rys. 1.

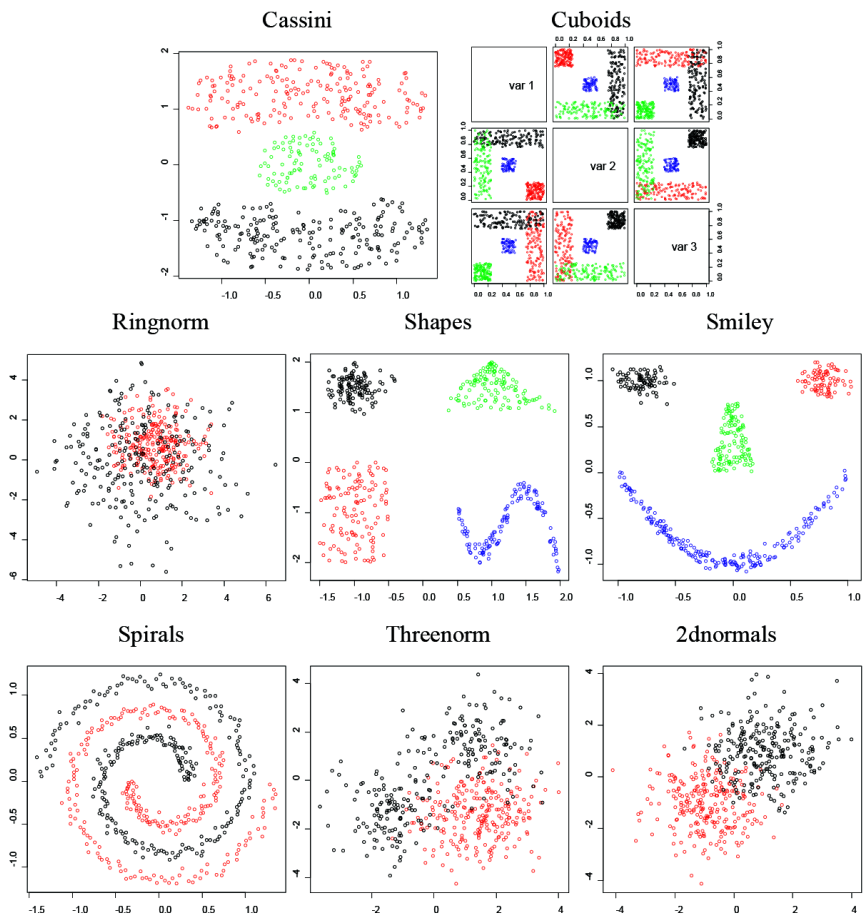
W metodzie *bagging* według Leischa jako metodę bazową zastosowano metodę *k*-średnich, natomiast ostatecznego grupowania dokonano z zastosowaniem: metody najbliższego sąsiedztwa (*single*), najdalszego sąsiedztwa (*complete*), średniej odległości między skupieniami (average), środka ciężkości (*centroid*), mediany (*median*), warda (*ward*). W metodzie Dudoit i Fridlyand oraz Hornika utworzono 50 prób bootstrapowych, na ich podstawie określano podziały składowe z zastosowaniem me-

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

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.

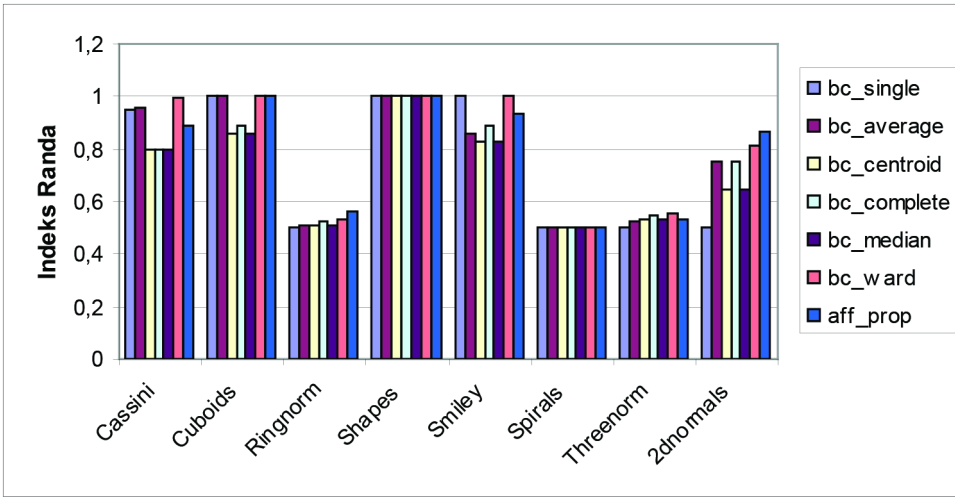


Rys. 1. Struktura zastosowanych zbiorów danych

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

tody *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 Dudoit i Fridlyand oraz 3 w metodzie Hornika[2].

Dokładność grupowania była badana za pomocą indeksu Randa.

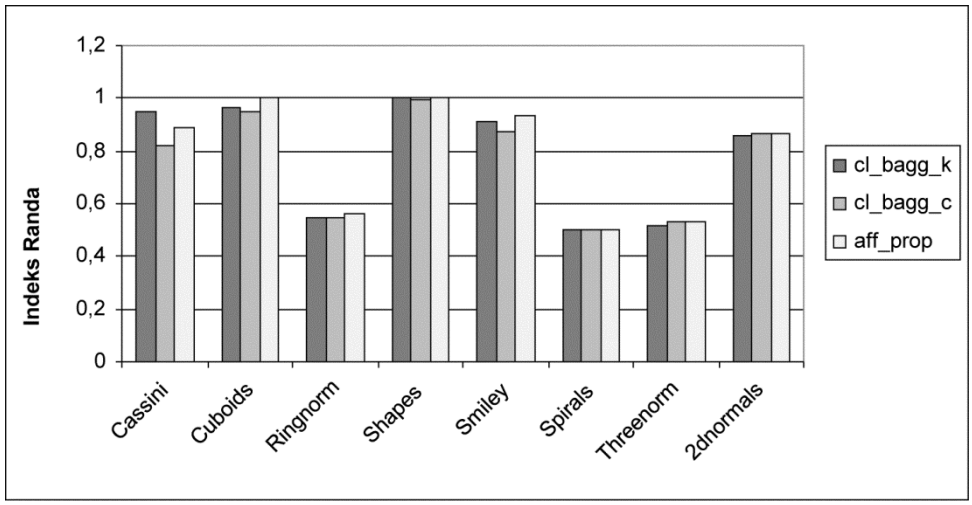


Rys. 2. Porównanie dokładności metody *bagging* według Leischa oraz metody propagacji podobieństwa

Źródło: opracowanie własne.

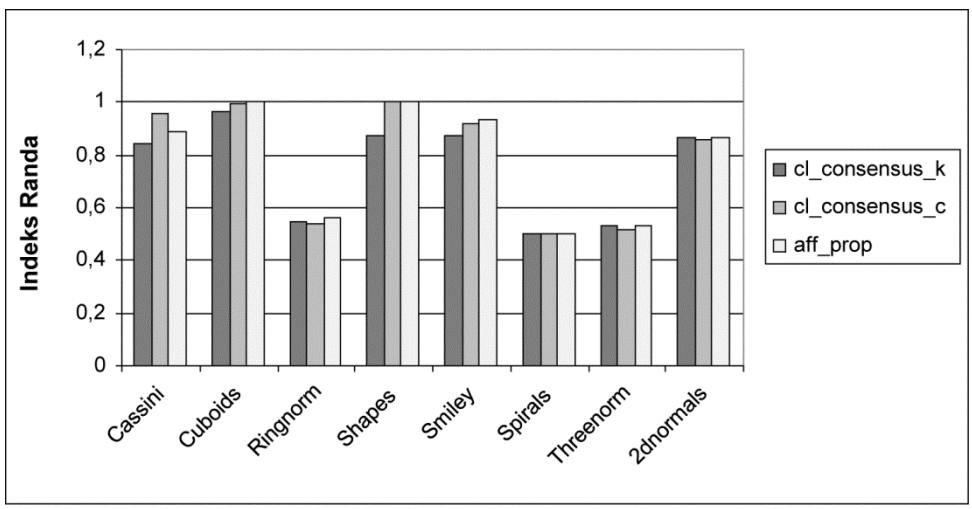
Wyniki empiryczne (rys. 2, 3, 4) nie wykazują, aby metoda propagacji podobieństwa dawała wyraźnie lepsze rezultaty niż metody zagregowane. Raczej są one porównywalne, chociaż można wskazać takie zbiory, dla których zauważalna jest nieznaczna przewaga metody propagacji podobieństwa nad podejściem zagregowanym, np. zbiory *Ringnorm*, *2dnormals* dla metody Leischa, *Cuboids* i *Smiley* dla metody Dudoit i Fridlyand, czy też *Ringnorm* i *Smiley* dla metody Hornika. Ponadto porównując metodę propagacji podobieństwa z metodą Hornika, można zauważyć, że zawsze metoda propagacji daje porównywalne lub lepsze rezultaty jak wariant cl_consensu_k. Podobną prawidłowość można też zaobserwować dla metody propagacji i wariantu cl_bagg_k w metodzie Dudoit i Fridlyand (z wyjątkiem zbioru *Cassini*).

[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.



Rys. 3. Porównanie dokładności metody *bagging* według Dudoit i Fridlyand oraz metody propagacji podobieństwa

Źródło: opracowanie własne.



Rys. 4. Porównanie dokładności metody *bagging* według Hornika oraz metody propagacji podobieństwa

Źródło: opracowanie własne.

5. Podsumowanie

W zagadnieniu taksonomii bardzo ważną kwestią jest zapewnienie wysokiej jakości grupowania, co powoduje, że w literaturze wciąż proponowane są nowe rozwiązania,

które mają być dokładniejsze niż metody tradycyjne (np. *k*-średnich, hierarchiczne). Przykładami takich rozwiązań mogą być metody polegające na zastosowaniu podejścia zagregowanego oraz stosunkowo niedawno zaproponowana metoda propagacji podobieństwa. Celem badań, które zaprezentowano w tym artykule, było porównanie dokładności tych dwóch podejść. Zaprezentowane wyniki empiryczne nie wykazały wyraźnej przewagi któregoś z tych dwóch sposobów grupowania. Wyniki analiz pozwalają określić dokładność obydwu podejść jako bardzo porównywalną.

Literatura

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

Bodenhofer U., Kothmeier A., Hochreiter S., *APCluster: an R package for affinity propagation clustering*, "Bioinformatics" 27(17):2463-2464, 2001. DOI: 10.1093/bioinformatics/btr406.

Dudoit S., Fridlyand J., *Bagging to improve the accuracy of a clustering procedure*, "Bioinformatics" 2003, vol. 19, no. 9, 1090-1099.

Frey B.J., Dueck D., *Clustering by passing messages between data points*, "Science", 315, 2007, 972-976. DOI: 10.1126/science.1136800.

Hornik K., *A CLUE for CLUster ensembles*, "Journal of Statistical Software", 2005, 14:65-72.

Leisch F., *Bagged clustering*, "Adaptive Information Systems and Modeling in Economics and Management Science", Working Papers, SFB, 1999, 51.

COMPARISON OF ACCURACY OF AFFINITY PROPAGATION CLUSTERING AND CLUSTER ENSEMBLES 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. Examples of such solutions can be cluster ensembles or affinity propagation method. The main aim of the article is to compare the accuracy of these two approaches. There will be considered cluster ensembles based on bagging idea [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999] and affinity propagation method proposed by Frey and Dueck [2007].

Keywords: taxonomy, cluster ensemble, affinity propagation, accuracy.