

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

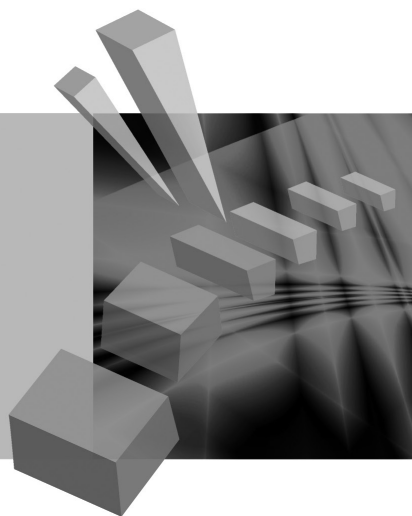
**RESEARCH PAPERS**

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](http://www.ibuk.pl), [www.ebscohost.com](http://www.ebscohost.com),

The Central and Eastern European Online Library [www.ceeol.com](http://www.ceeol.com),

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

[http://kangur.uek.krakow.pl/bazy\\_ae/bazekon/nowy/index.php](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](http://www.wydawnictwo.ue.wroc.pl)

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2013

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

## Spis treści

<b>Wstęp</b> .....	9
<b>Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski:</b> Sejm VI kadencji – maszynka do głosowania .....	11
<b>Barbara Pawelek, Adam Sagan:</b> Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I .....	19
<b>Jan Paradysz:</b> Nowe możliwości badania koniunktury na rynku pracy .....	29
<b>Krzysztof Najman:</b> Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze .....	41
<b>Kamila Migdał-Najman:</b> Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym .....	48
<b>Aleksandra Matuszewska-Janica, Dorota Witkowska:</b> Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych .....	58
<b>Iwona Foryś, Ewa Putek-Szeląg:</b> Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim .....	67
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk:</b> Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych.....	77
<b>Marta Jaročka:</b> Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni .....	85
<b>Anna Zamojska:</b> Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych.....	95
<b>Dorota Rozmus:</b> Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i> .....	106
<b>Ewa Wędrowska:</b> Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur.....	115
<b>Katarzyna Wójcik, Janusz Tuchowski:</b> Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ...	124
<b>Małgorzata Misztal:</b> Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych.....	135
<b>Anna Czapkiewicz, Beata Basiura:</b> Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych .....	146
<b>Tomasz Szubert:</b> Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej” .....	154

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

<b>Ewa Putek-Szeląg, Urszula Gieraltowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej..</b>	342
--	-----

## Summaries

<b>Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: VIth-term Sejm – a voting machine .....</b>	18
<b>Barbara Pawelek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model .....</b>	28
<b>Jan Paradysz: New possibilities for studying the situation on the labour market .....</b>	40
<b>Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data.....</b>	47
<b>Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering .....</b>	57
<b>Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees .....</b>	66
<b>Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship...</b>	76
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables .....</b>	84
<b>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.....</b>	94
<b>Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures.....</b>	105
<b>Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea.....</b>	114
<b>Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements .....</b>	123
<b>Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis .....</b>	134
<b>Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ...</b>	145
<b>Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series.....</b>	153
<b>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 .....</b>	162
<b>Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures .....</b>	173

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

**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 zależy 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 określone 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  $G = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  wylosuj  $B$  prób bootstrapowych  $G_n^1, G_n^2, \dots, G_n^B$ , 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  $B \times K$  załączków skupień  $c_{11}, c_{12}, \dots, c_{1K}, c_{21}, \dots, c_{BK}$ , gdzie  $K$  oznacza liczbę skupień w metodzie bazowej, a  $c_{bk}$  jest  $k$ -tym załączkiem znalezionym na podstawie podpróby  $G_n^b$ .

3. Załączki skupień uzyskane na podstawie kolejnych prób bootstrapowych przekształć w nowy zbiór danych  $C^B = \{c_{11}, \dots, c_{BK}\}$ .



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  $C_1^B, \dots, C_m^B$ , gdzie  $1 \leq m \leq BK$ .

5. Każdą obserwację  $x_i$  z pierwotnego zbioru danych  $G$  przydziel do tej grupy, w której znajduje się najbliższej leżący załazek  $c(x_i)$ , 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  $G = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , uzyskując w ten sposób etykiety klas  $T(x_i, G) = \hat{y}_i$  dla każdej obserwacji  $x_i, i = 1, \dots, n$ .

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

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  $1, \dots, K$ . Znajdź permutację  $\tau^b \in PR_K$  maksymalizującą:

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

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:

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

**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:

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

gdzie:  $C$  to zbiór wszystkich możliwych podziałów zagregowanych,  $\text{dist}$  – odległość Euklidesowa,  $(c_1, \dots, c_B)$  – 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:

$$s(i, k) = -\|x_i - x_k\|^2. \quad (4)$$

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

$$s(k, k) = p, \quad (5)$$

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 odpowiedności i osiągalności obliczane są z formuł:

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

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

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  $\hat{\mathbf{c}} = (\hat{c}_1, \dots, \hat{c}_N)$  odbywa się według reguły:

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

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<sup>1</sup>. 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.

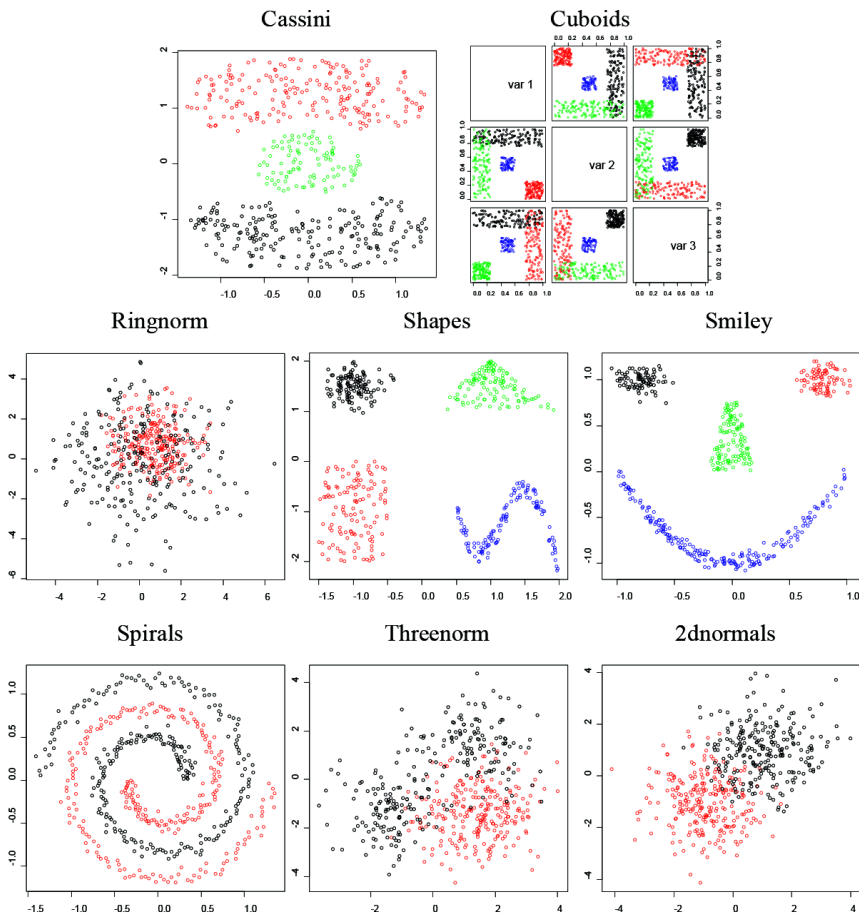
W metodzie *bagging* według Leischa jako metodę bazową zastosowano metodę *k*-średnich, natomiast ostatecznego grupowania dokonano z zastosowaniem: metody najbliższego sąsiedztwa (*single*), najdalszego sąsiedztwa (*complete*), średniej odległości między skupieniami (*average*), środka ciężkości (*centroid*),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

<sup>1</sup> Zbiory zaczerpnięte zostały z pakietu `mlbench` z programu **R**.

**Tabela 1.** Charakterystyka zastosowanych zbiorów danych

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

Źródło: opracowanie własne.

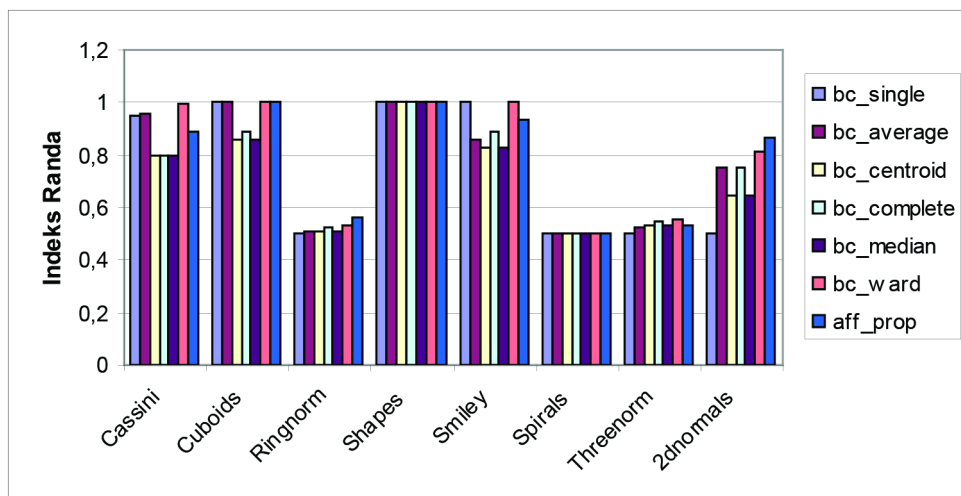


**Rys. 1.** Struktura zastosowanych zbiorów danych

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<sup>2</sup>.

Dokładność grupowania była badana za pomocą indeksu Randa.

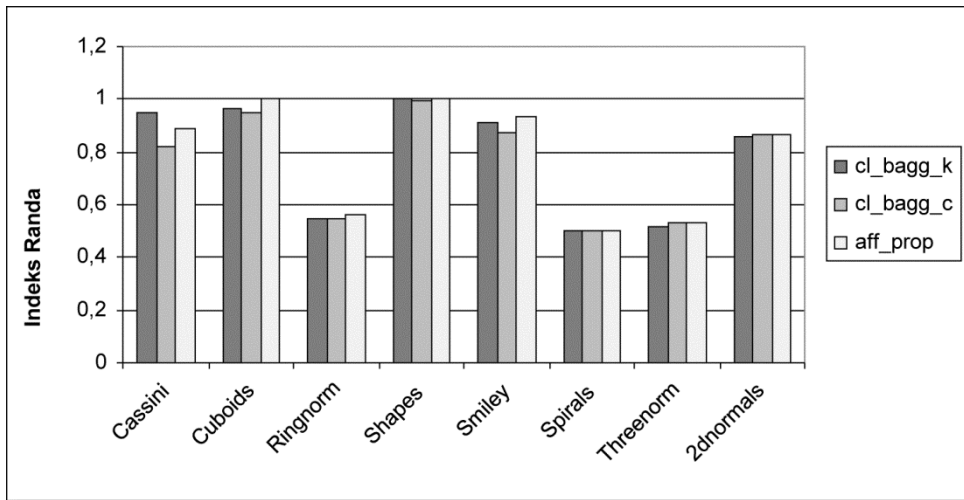


Rys. 2. Porównanie dokładności metody *bagging* według Leischa oraz metody propagacji podobieństwa

Źródło: opracowanie własne.

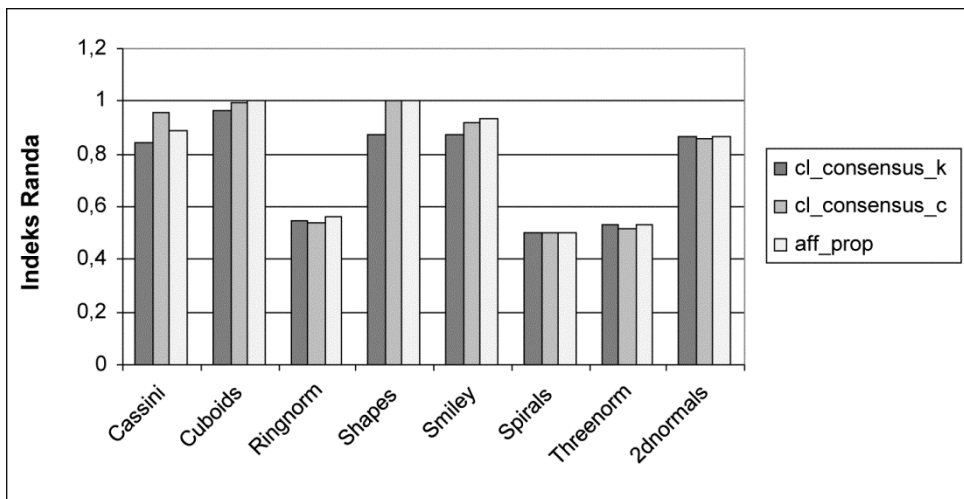
Wyniki empiryczne (rys. 2, 3, 4) nie wykazują, aby metoda propagacji podobieństwa dawała wyraźnie lepsze rezultaty niż metody zagregowane. Raczej są one porównywalne, chociaż można wskazać takie zbiory, dla których zauważalna jest nieznaczna przewaga metody propagacji podobieństwa nad podejściem zagregowanym, 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\_consensus\_k*. Podobną prawidłowość można też zaobserwować dla metody propagacji i wariantu *cl\_bagg\_k* w metodzie Dudoit i Fridlyand (z wyjątkiem zbioru *Cassini*).

<sup>2</sup> Na rysunkach 3 i 4 stosowano skróty *cl\_bagg\_k* i *cl\_consensus\_k*, jeżeli grupowania składowe określone 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.