**Dorota Rozmus**
Uniwersytet Ekonomiczny w Katowicach

# PORÓWNANIE STABILNOŚCI ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA IDEI METODY BAGGING

## Wprowadzenie

Pierwotnie podejście zagregowane (wielomodelowe) z dużym powodzeniem było stosowane w dyskryminacji i regresji w celu podniesienia dokładności predykcji. Zasadnicza idea tego podejścia polega na tym, że w pierwszym kroku są budowane liczne różniące się między sobą pojedyncze modele, które następnie za pomocą różnych operatorów są łączone w model zagregowany. W dyskryminacji najczęściej stosowanym operatorem jest głosowanie majoryzacyjne, co oznacza, że jest wybierana ta klasa, która najczęściej była wskazywana przez pojedyncze modele; natomiast w regresji najczęściej stosuje się uśrednianie wartości teoretycznych zmiennej $y$. Wśród najbardziej znanych metod agregacji należy wymienić: *bagging* [Breiman 1996], który jest oparty na losowaniu kolejnych prób boostrapowych oraz *boosting* [Freund 1999] polegający na nadawaniu wyższych wartości wag błędnie sklasyfikowanym obiektom.

W ostatnich latach analogiczne propozycje pojawiły się także w taksonomii, aby zapewnić większą poprawność i stabilność wyników grupowania [Fern i Brodley 2003; Fred 2002; Fred i Jain 2002; Strehl i Gosh 2002]. Zagadnienie agregacji w taksonomii może zostać sformułowane następująco: mając wyniki wielokrotnie przeprowadzonej klasyfikacji, znajdź zagregowany podział ostateczny o lepszej jakości. Liczne badania w tej dziedzinie ustanowiły już nowy obszar w tradycyjnej taksonomii. Istnieje wiele możliwości zastosowania idei podejścia zagregowanego w dziedzinie uczenia bez nauczyciela, wśród których jako najpopularniejsze należy wymienić:

1. Łączenie wyników grupowania uzyskanych za pomocą różnych metod.
2. Uzyskanie różniących się między sobą klasyfikacji z zastosowaniem różnych podzbiorów danych, np. poprzez losowanie bootstrapowe.
3. Stosowanie różnych podzbiorów zmiennych.
4. Wielokrotne zastosowanie tego samego algorytmu z różnymi wartościami parametrów lub punktami startowymi (np. losowo wybranymi zalążkami skupień w metodzie *k*-średnich).

Algorytm taksonomiczny powinien charaktryzować się stabilnością, a więc powinien być odporny na niewielkie zmiany w zbiorze danych, czy też wartości parametrów tego algorytmu. Wiadomo jednakże również, że kluczem do sukcesu podejścia zagregowanego jest zróżnicowanie klasyfikacji składowych. Klasyfikacja zagregowana, która została zbudowana na różniących się między sobą elementach składowych jest bardziej dokładna i stabilna niż pojedyncze metody taksonomiczne. W niniejszym badaniu uwaga zostanie skupiona na stabilności metod taksonomicznych. Głównym celem tego artykułu jest porównanie stabilności zagregowanych algorytmów taksonomicznych, a także relacji między stabilnością i dokładnością; przy czym pod uwagę zostanie wzięta tylko specyficzna klasa metod agregacji, które są oparte na idei metody *bagging*.

## 1. Metoda *bagging* w taksonomii

Metoda *bagging* jest pewną ogólną koncepcją, w ramach której narodziły się szczegółowe rozwiązania zaproponowane m.in. przez Hornika [2005], Dudoid i Fridlyand [2003] oraz Leischa [1999]. Pierwszy krok we wszystkich tych metodach jest taki sam: polega na konstrukcji *B* prób bootstrapowych i zastosowaniu do nich pojedynczego algorytmu taksonomiczego w celu uzyskania klasyfikacji składowych wchodzących w skład klasyfikacji ostatecznej. Poszczególne warianty tej metody różnią się natomiast w kroku drugim, czyli w kroku agregacji wyników.

### Propozycja Leischa

Leisch [1999] zaproponował, by w pierwszym kroku na podstawie każdej podpróby bootstrapowej dokonać grupowania przy zastosowaniu tzw. bazowej metody taksonomicznej, którą jest jedna z metod iteracyjno-optymalizacyjnych, np. algorytm *k*-średnich. W kolejnym etapie ostateczne centra skupień są prze-

kształcane w nowy zbiór danych obejmujący $B \times K$ obserwacji ($K$ to liczba skupień w metodzie bazowej), który jest poddawany podziałowi za pomocą metod hierarchicznych. Uzyskany dendrogram jest podstawą ostatecznego podziału – obserwacje z pierwotnego zbioru są przydzielane do tej grupy, której środek ciężkości znajduje się w minimalnej odległości Euklidesowej.

Algorytm zaproponowany przez Leischa przebiega w następujących krokach:

1. Z pierwotnego $N$-elementowego zbioru $G$ 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 zbioru 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 $C^B = C^B(K) = \{c_{11}, ..., c_{BK}\}$.
4. Do tak skonstruowanego zbioru należy zastosować hierarchiczną metodę taksonomiczną, uzyskując w ten sposób dendrogram.
5. Niech $c(x_i)$ oznacza zalążek znajdujący się najbliżej obserwacji $x_i$, $i$ = 1, …, $n$. Podział na grupy pierwotnego zbioru danych jest określany 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 Duidoid i Fridlyand

Metoda *bagging* w wersji zaproponowanej przez Dudoid i Fridlyand [2003] stosuje algorytmy iteracyjno-optymalizacyjne do oryginalnego zbioru danych i poszczególnych prób bootstrapowych, a po dokonaniu permutacji etykiet klas w wynikach grupowania uzyskanych na podstawie każdej podpróby tak, by zachodziła jak największa zbieżność z klasyfikacją obiektów z oryginalnego zbioru danych, stosuje głosowanie majoryzacyjne w celu określenia ostatecznej klasyfikacji zagregowanej.

Kroki zaproponowanego przez nich algorytmu można ująć według następującego schematu. Dla założonej liczby klas *K*:

1. Zastosuj iteracyjno-optymalizacyjny algorytm taksonomiczny *T* do pierwotnego zbioru danych *G*, 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 $G_n^b = (x_1^b, ..., x_n^b)$.
3. Zastosuj algorytm taksonomiczny *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 pojedynczego algorytmu taksonomicznego uzyskuje się klasyfikacje składowe. Klasyfikacja zagregowana natomiast jest uzyskiwana 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 klasyfikacji zagregowanych,
*dist* – odległość Euklidesowa,
$(c_1,...,c_B)$ – klasyfikacje wchodzące w skład klasyfikacji zagregowanej.

## 2. Miary stabilności i dokładności

W celu zbadania stabilności i dokładności zastosowano koncepcję miar zaproponowanych przez Kunchevę i Vetrova [2006]. Mierniki te są oparte na skorygowanym indeksie Randa (*AR*), którego definicja jest następująca [Hubert i Arabie 1985]: niech $A$ i $B$ będą wynikami dwóch różnych klasyfikacji zbioru $Z$ posiadającego $N$ elementów. Przez $l_A$ oznaczmy liczbę klas w klasyfikacji $A$, natomiast przez $l_B$ – liczbę klas w klasyfikacji $B$; $N_{ij}$ to liczba obiektów znajdujących się w klasie $i$ w grupowaniu $A$ i w klasie $j$ w klasyfikacji $B$; $N_{i\bullet}$ to liczba obserwacji w klasie $i$ w klasyfikacji $A$, natomiast $N_{\bullet j}$ to liczba obserwacji w klasie $j$ w klasyfikacji $B$. Skorygowany indeks Randa jest dany wzorem:

$$AR(A,B) = \frac{\sum_{i=1}^{l_A}\sum_{j=1}^{l_B}\binom{N_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}, \tag{4}$$

gdzie:

$$t_1 = \sum_{i=1}^{l_A}\binom{N_{i\bullet}}{2}, \tag{5}$$

$$t_2 = \sum_{j=1}^{l_B}\binom{N_{\bullet j}}{2}, \tag{6}$$

$$t_3 = \frac{2t_1 t_2}{N(N-1)}. \tag{7}$$

1. STABILNOŚĆ DLA PAR KLASYFIKACJI ZAGREGOWANYCH (ang. *pairwise ensemble stability*):

$$S_{agr} = \frac{2}{Z \cdot (Z-1)} \sum_{\substack{1 \le z,l \le Z \\ z<l}}^{Z} AR(P_z^{agr}, P_l^{agr}), \tag{8}$$

gdzie:
$Z$ – liczba klasyfikacji zagregowanych,
$AR$ – skorygowany indeks Randa,
$P_z^{agr}$ – klasyfikacja na podstawie $z$-tej klasyfikacji zagregowanej,
$P_l^{agr}$ – klasyfikacja na podstawie $l$-tej klasyfikacji zagregowanej.

Miara ta ocenia stabilność klasyfikacji zagregowanych poprzez ocenę podobieństwa wyników grupowania, które na ich podstawie zostały uzyskane.

2. PRZECIĘTNA DOKŁADNOŚĆ KLASYFIKACJI ZAGREGOWANEJ (ang. *average ensemble accuracy*):

$$A_{agr} = \frac{1}{Z} \sum_{z=1}^{Z} AR(P_z^{agr}, P^T), \tag{9}$$

gdzie: $P^T$ – rzeczywiste etykiety klas.

Miara ta jest uśrednioną po wszystkich klasyfikacjach zagregowanych miarą dokładności i mierzy podobieństwo między ostateczną klasyfikacją zagregowaną a prawdziwymi etykietami klas.

## 3. Badania empiryczne

W badaniach zastosowano sztucznie generowane zbiory danych, które standardowo są wykorzystywane w badaniach porównawczych w taksonomii[1]. Są to takie zbiory, w których przynależność obiektów do klas jest znana. Ich krótka

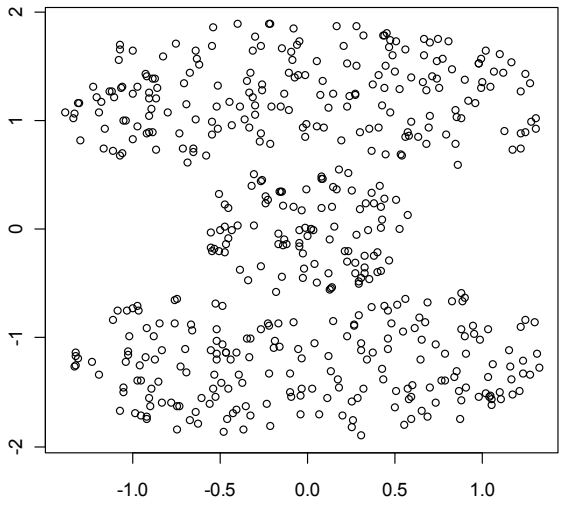
[1] Zbiory zaczerpnięte zostały z pakietu `mlbench` z programu **R**.

charakterystyka znajduje się w tabeli 1, natomiast struktura jest pokazana na rys. 1-8. Zbiory *Cassini*, *Cuboids*, *Shapes*, *Smiley* oraz *Spirals* należą do zbiorów o wyraźnie separowalnych klasach, natomiast *2dnormals*, *Ringnorm* i *Threenorm* posiadają nakładające się na siebie, trudno separowalne klasy.

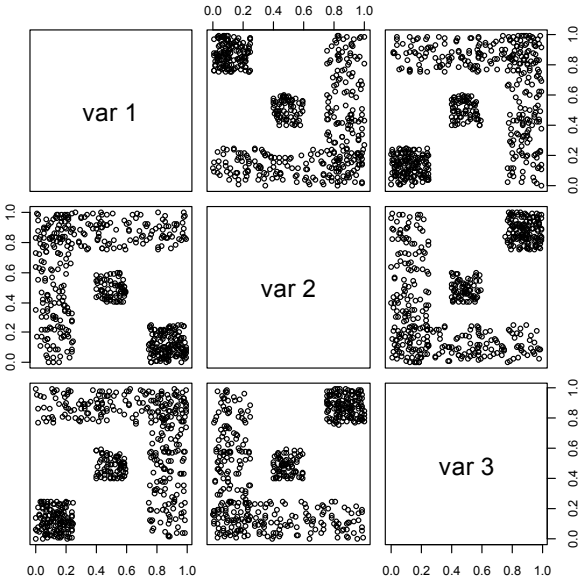
Tabela 1

Charakterystyka zastosowanych zbiorów danych

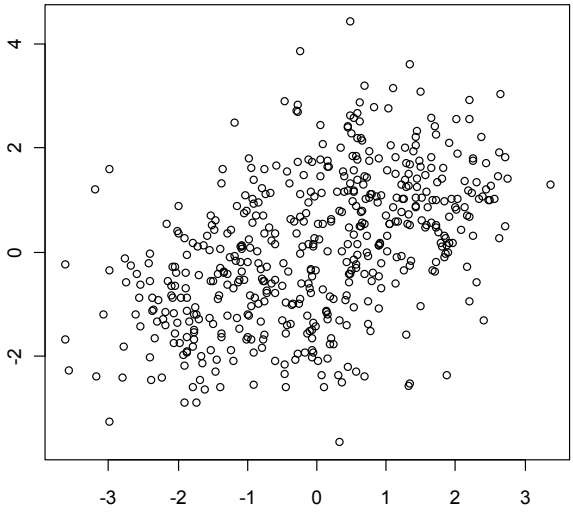
| Zbiór danych | Liczba obiektów | Liczba cech | Liczba klas |
|---|---|---|---|
| *Cassini* | 500 | 2 | 3 |
| *Cuboids* | 500 | 3 | 4 |
| *2dnormals* | 500 | 2 | 2 |
| *Ringnorm* | 500 | 2 | 2 |
| *Shapes* | 500 | 2 | 4 |
| *Smiley* | 500 | 2 | 4 |
| *Spirals* | 500 | 2 | 2 |
| *Threenorm* | 500 | 2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



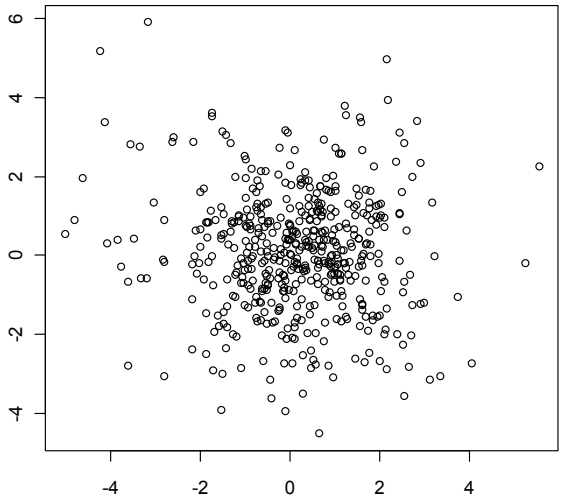

Rys. 1. Zastosowane zbiory danych – zbiór *Cassini*



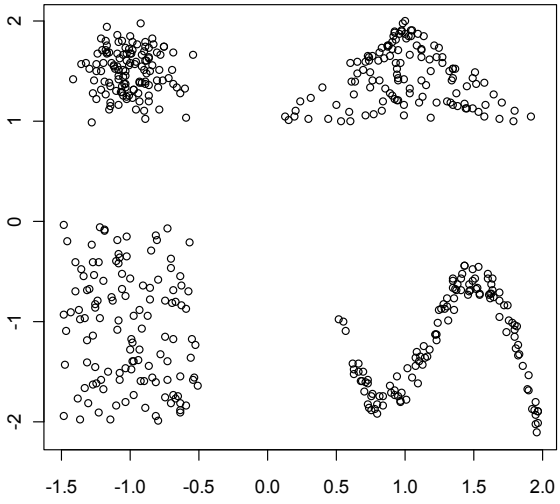

Rys. 2. Zastosowane zbiory danych – zbiór *Cuboids*



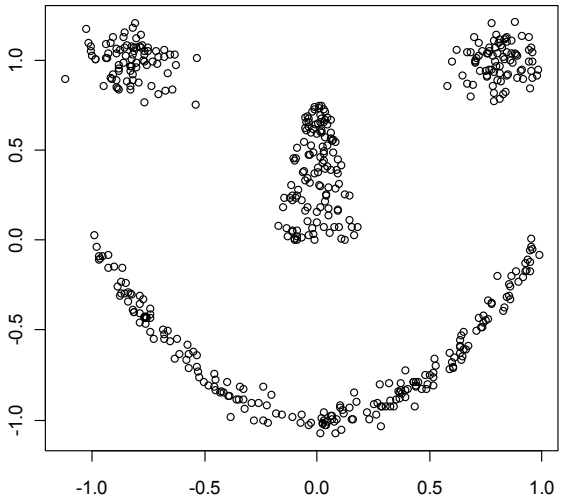

Rys. 3. Zastosowane zbiory danych – zbiór *2dnormals*



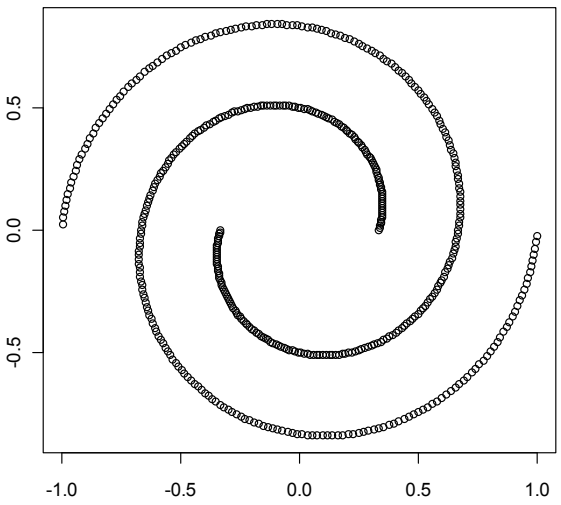

Rys. 4. Zastosowane zbiory danych – zbiór *Ringnorm*



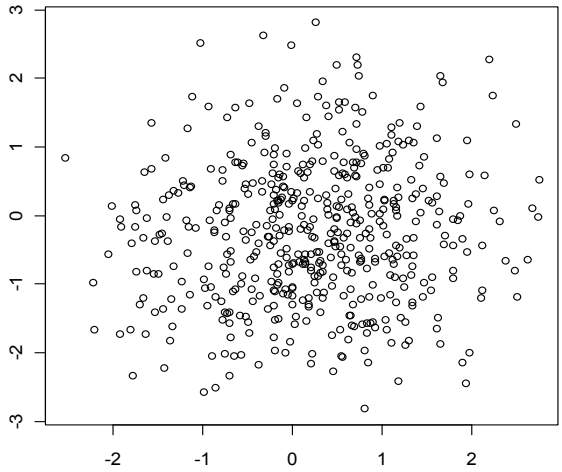

Rys. 5. Zastosowane zbiory danych – zbiór *Shapes*




Rys. 6. Zastosowane zbiory danych – zbiór *Smiley*




Rys. 7. Zastosowane zbiory danych – zbiór *Spirals*



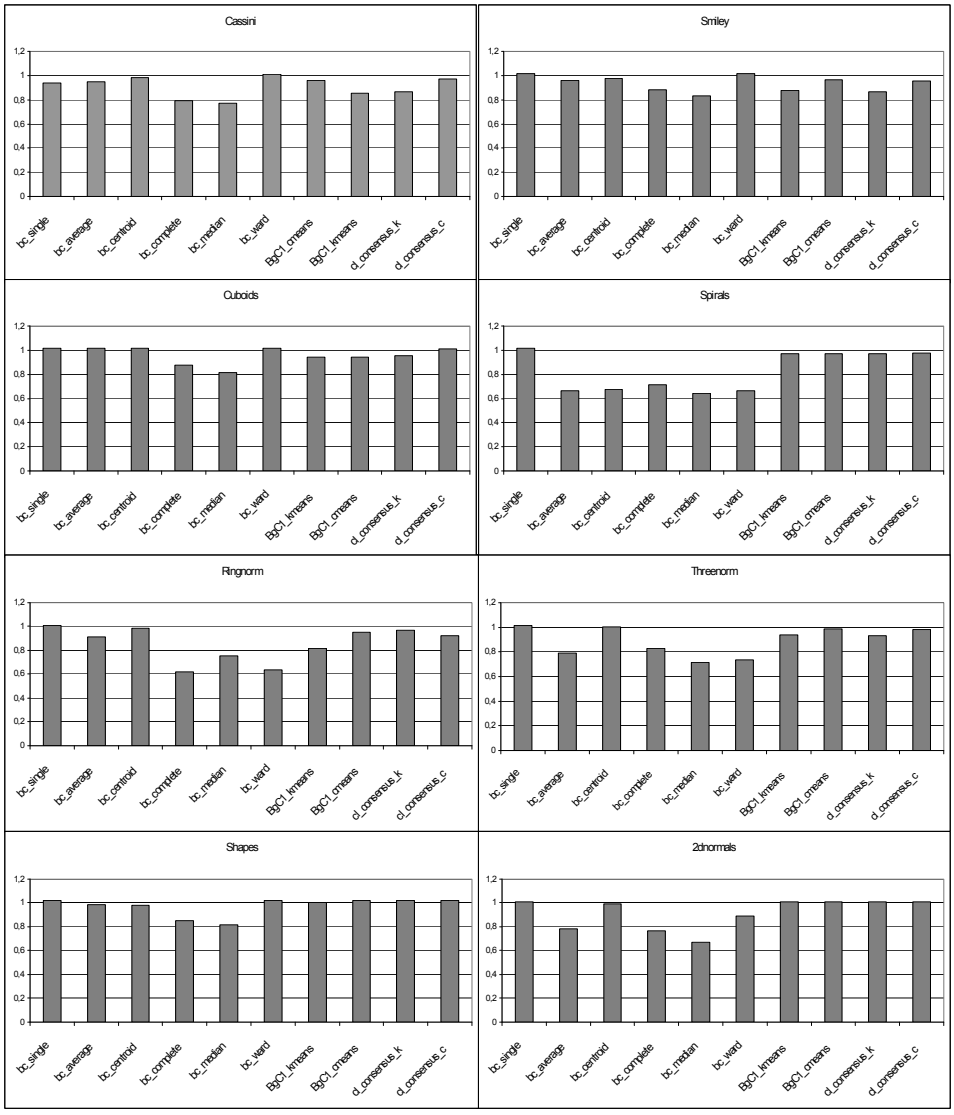

Rys. 8. Zastosowane zbiory danych – zbiór *Threenorm*

W badaniach empirycznych zastosowano 50 klasyfikacji zagregowanych, a wszystkie obliczenia zostały powtórzone 50 razy, by uzyskać bardziej dokładne i wiarygodne rezultaty. W metodzie *bagging* zaproponowanej przez Leischa po skonstruowaniu 10 prób bootstrapowych jako bazowy iteracyjno-optymalizacyjny algorytm taksonomiczny zastosowano metodę *k*-średnich z wartością parametru $k = 50$[2], a po przekształceniu ostatecznych zalążków skupień do postaci zbioru danych obejmującego 500 obserwacji dokonano podziału za pomocą następujących hierarchicznych metod taksonomicznych[3]: najbliższego sąsiedztwa (bclust_single), najdalszego sąsiedztwa (bclust_complete), centroidy (bclust_centroid), mediany (bclust_median), średniej odległości (bclust_mean), warda (bclust_ward). Obliczenia zostały wykonane w programie **R** z zastosowaniem funkcji `bclust` z pakietu `e1071`.

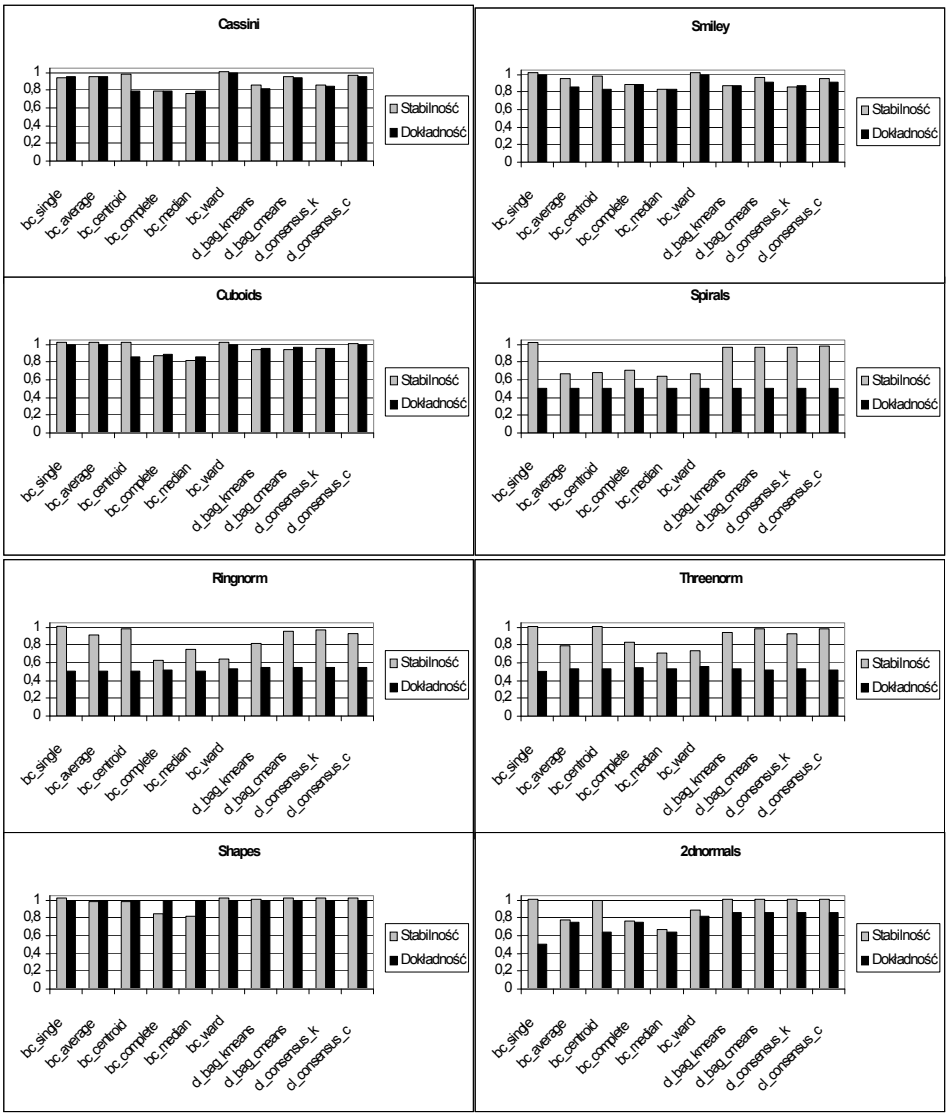
W metodzie *bagging* w wersji zaproponowanej przez Dudoid i Fridlyand oraz przez Hornika po skonstruowaniu 25 prób bootstrapowych zastosowano dwa algorytmy, a mianowicie metodę *k*-średnich oraz *c*-średnich, która jest rozmytą wersją metody *k*-średnich opracowaną przez Bezdeka [1981]. Metoda Dudoid i Fridlyand jest oprogramowana w programie **R** pod nazwą funkcji `cl_bag` w pakiecie `clue` (na rysunkach zastosowano nazwy cl_bag_kmeans oraz cl_bag_cmeans), natomiast metodę Hornika można znaleźć w tym samym pakiecie pod nazwą `cl_consensus` (na rysunkach oznaczenie cl_consensus_k odnosi się do metody agregacji, gdzie na poszczególnych próbach bootstrapowych była stosowana metoda *k*-średnich, a cl_consensus_c – metoda *c*-średnich).

[2] Autor metody zaleca, by wartość tego parametru była większa niż rzeczywista liczba skupień.
[3] W nawiasach zostały podane skróty nazw stosowane na rysunkach.

Rezultaty obliczeń widoczne na rys. 9 pozwalają stwierdzić, że w prawie wszystkich przypadkach najmniej stabilną okazała się metoda bclust_complete oraz bclust_median. Najwyższą stabilnością w przypadku większości zbiorów danych charakteryzują się metody: bclust_single, bclust_average oraz bclust_centroid (z wyjątkiem metod bclust_average oraz bclust_centroid dla zbioru *Spirals* oraz metody bclust_average dla zbioru *Threenorm* i *2dnormals*). Całkiem stabilne rezultaty można także zaobserwować dla reszty badanych metod z wyjątkiem metody bclust_ward dla zbiorów *Ringnorm*, *Threenorm* oraz *Spirals*.




Rys. 9. Stabilność poszczególnych metod opartych na idei *bagging* dla różnych zbiorów danych




Rys. 10. Relacje między stabilnością a dokładnością dla poszczególnych metod opartych na idei *bagging* dla różnych zbiorów danych

Wykresy na rys. 10 pokazujące relacje zachodzące między miarami stabilności i dokładności pozwalają stwierdzić brak generalnie obowiązującej zależności. Na przykład dla zbioru *Cassini* oraz *Cuboids* miary stabilności i dokładności osiągają niemalże ten sam poziom (z wyjątkiem metody bc_centroid). Podobnie miary te kształtują się także dla zbiorów *Shapes* oraz *Smiley* (z wyjąt-

kiem metod bclust_complete i bclust_median dla zbioru *Shapes* oraz metody bclust_median dla zbioru *Smiley*). Już dla zbioru *Ringnorm, Threenorm* oraz *Spirals* można jednak zaobserwować, że miary dokładności kształtują się na niemalże tym samym poziomie, natomiast miary stabilności zachowują się różnie dla różnych metod[4]. Na przykład dla cl_bag_cmeans, cl_bag_kmeans, cl_consensus_kmeans i cl_consensus_cmeans przyjmują dosyć duże wartości, a dla bclust_ward – stosunkowo niskie.

## Podsumowanie

Przechodząc do sformułowania uwag końcowych, należy na wstępie zauważyć, że wybór dobrego algorytmu taksonomicznego jest znacznie trudniejszy niż wybór dobrego algorytmu dyskryminacyjnego. Wynika to przede wszystkim z faktu, że w klasyfikacji wzorcowej mamy do czynienia z zagadnieniem uczenia z nauczycielem. W taksonomii natomiast nie znamy klas, do których należą obiekty, a tym samym brak jest określonej z góry struktury, która powinna zostać rozpoznana przez algorytm. W związku z tym, by ominąć ryzyko wyboru niewłaściwego algorytmu taksonomicznego, można zastosować podejście zagregowane celem połączenia wyników klasyfikacji różnych algorytmów. Każdy z nich ma swoje mocne i słabe strony, ale wydaje się, że ich łączne zastosowanie przyniesie efekt kompensacji.

Drugą zaletą podejścia zagregowanego jest uniezależnienie wyników od wybranej metody, czy też wartości pewnych parametrów tych metod (np. początkowo wybranych zalążków skupień w metodzie *k*-średnich), a także zwiększenie odporności algorytmów taksonomicznych na szum i obserwacje oddalone. Agregacja wyników pozwala zatem na stabilizację rezultatów grupowania.

Wspomniane zalety powodują, że podejście to jest warte uwagi i tego, by spróbować zbadać relacje zachodzące między stabilnością i dokładnością zagregowanych algorytmów taksonomicznych. W przypadku gdyby między nimi zachodził wyraźny związek, mierniki stabilności mogłyby posłużyć jako wskazówka pomagająca wybrać najlepszą metodę podziału.

Z przeprowadzonych badań nad stabilnością zagregowanych metod taksonomicznych opartych na metodzie *bagging* wynika, że najbardziej stabilne okazały się metody: bclust_single, bclust_average, bclust_centroid, cl_bag_cmeans, cl_bag_kmeans, cl_consensus_kmeans oraz cl_consensus_cmeans. Najmniej

[4] Głównym punktem zainteresowania badań jest stabilność zagregowanych algorytmów taksonomicznych, dlatego przedstawiono wyniki nawet wtedy, gdy dokładność klasyfikacji nie osiągała wysokich wartości.

stabilne okazały się natomiast metody bclust_centroid oraz bclust_median; podczas gdy metoda bclust_ward dla niektórych zbiorów była bardzo stabilna (np. dla zbiorów *Cassini, Cuboids, Shapes* i *Smiley*), a dla niektórych stabilność była stosunkowo niska.

Z badań nad relacją między stabilnością i dokładnością w algorytmach opartych na metodzie *bagging* wynika, że nie da się sformułować jasnej i ogólnie obowiązującej zasady. Dla niektórych zbiorów danych stabilność i dokładność kształtuje się na zbliżonym do siebie poziomie, a dla niektórych stwierdza się brak jakiegokolwiek związku między nimi.

## Literatura


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

Breiman L. (1996): *Bagging Predictors*. „Machine Learning", No. 26(2).

Dudoit S., Fridlyand J. (2003): *Bagging to Improve the Accuracy of a Clustering Procedure*. „Bioinformatics", Vol. 19, No. 9.

Fern X.Z., Brodley C.E. (2003): *Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach*. „Proceedings of the 20th International Conference of Machine Learning".

Fred A. (2002): *Finding Consistent Clusters in Data Partitions*. „Proceedings of the International Workshop on Multiple Classifier Systems".

Fred N.L., Jain A.K. (2002): *Combining Multiple Clusterings Using Evidence Accumulation*. „IEEE Transactions on PAMI", No. 27(6).

Freund Y. (1999): *An Adaptive Version of the Boost by Majority Algorithm*. „Proceedings of the 12th Annual Conference on Computational Learning Theory".

Hornik K. (2005): *A CLUE for CLUster Ensembles*. „Journal of Statistical Software", No. 14.

Hubert L.., Arabie P. (1985): *Evaluating Object Set Partitions: Free Sort Analysis and Some Generalizations*. „Journal of Verbal Learning and Verbal Behaviour", No. 15.

Kuncheva L., Vetrov D. (2006): *Evaluation of Stability of k-means Cluster Ensembles with Respect to Random Initialization*. „IEEE Transactions On Pattern Analysis And Machine Intelligence", Vol. 28, No. 11.

Leisch F. (1999): *Bagged Clustering*. „Adaptive Information Systems and Modeling in Economics and Management Science", Working Paper 51.

Strehl A., Ghosh J. (2002): *Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions*. „Journal of Machine Learning Research", No. 3.

# COMPARISON OF STABILITY OF CLUSTER ENSEMBLES BASED ON BAGGING IDEA

## Summary

Ensemble approach has been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. One of the most popular method is bagging based on bootstrap samples. Recently, analogous techniques for cluster analysis have been suggested in order to increase classification accuracy, robustness and stability of the clustering solutions. Research has proved that, by combining a collection of different clusterings, an improved solution can be obtained.

A desirable quality of the method is the stability of a clustering algorithm with respect to small perturbations of data (e.g., data subsampling or resampling, small variations in the feature values) or the parameters of the algorithm (e.g., random initialization). Here, we look at the stability of the ensemble and carry out an experimental study to compare stability of cluster ensembles based on bagging idea.