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

**Dorota Rozmus**
Akademia Ekonomiczna w Katowicach

# WYKORZYSTANIE PODEJŚCIA ZAGREGOWANEGO W TAKSONOMII

## 1. Wstęp

Podejście wielomodelowe było w ostatnich latach z powodzeniem stosowane w klasyfikacji i regresji celem poprawy dokładności predykcji. Do najbardziej znanych metod agregacji modeli należą metody bootstrapowe (*bagging* i *boosting*). Generalnie idea tego podejścia polega na tym, że w pierwszym kroku budowane są liczne pojedyncze i różniące się między sobą modele, które następnie są łączone różnymi operatorami w model zagregowany. W regresji najczęściej uśredniane są wartości teoretyczne uzyskane na podstawie pojedynczych modeli, a w klasyfikacji wybierana jest ta klasa, która najczęściej była wskazywana przez pojedyncze modele[1]. Idea zastosowania podejścia wielomodelowego pojawiła się także w taksonomii po to, aby podnieść dokładność rozpoznawania poprawnej struktury klas i zmniejszyć zmienność wyników grupowania [Fred, Jain 2005; Kuncheva i in. 2006].

Zasadniczym celem artykułu jest porównanie zdolności rozpoznawania poprawnej struktury klas uzyskanych za pomocą klasycznych algorytmów taksonomicznych oraz przedstawionego w literaturze podejścia wielomodelowego.

## 2. Zastosowanie metody *bagging* w taksonomii

Jedną z ciekawszych propozycji w zakresie zastosowania podejścia wielomodelowego w taksonomii przedstawił Leisch [1999], proponując połączenie metod iteracyjno-optymalizacyjnych z hierarchicznymi. Zaczerpnął on generalną ideę algorytmu *bagging* [Breiman 1996], która polega na tworzeniu kolejnych podprób uczących przy wykorzystaniu schematu losowania ze zwracaniem i zastosowaniu do nich metod taksonomicznych. Na podstawie każdej takiej podpróby określane są rezultaty grupowania przy zastosowaniu tzw. bazowej metody taksonomicznej,

[1] Jest to tzw. podejście majoryzacyjne [Gatnar 2001].

którą jest jedna z metod iteracyjno-optymalizacyjnych, np. algorytm *k*-średnich. W kolejnym etapie ostateczne centra skupień przekształcane są w nowy zbiór danych, który poddawany jest podziałowi za pomocą metod hierarchicznych. Głównym celem zastosowania idei podejścia wielomodelowego była stabilizacja rezultatów uzyskiwanych za pomocą metod iteracyjno-optymalizacyjnych poprzez wielokrotne stosowanie algorytmu i agregację wyników. Na przykład metoda *k*-średnich jest niestabilna w tym sensie, że przy wielokrotnym jej stosowaniu za każdym razem znalezione będzie jedynie lokalne optimum. Zarówno początkowo wybierane zalążki centrów skupień, jak i niewielkie zmiany w zbiorze danych mają wpływ na postać tego lokalnego optimum. Przez powtórne zastosowania tego algorytmu uzyskuje się różne rezultaty grupowania, które – przeciętnie rzecz biorąc – powinny być niezależne od wartości początkowo wybranych zalążków skupień.

Algorytm przebiega w następujących krokach:

1. Z pierwotnego $N$-elementowego zbioru uczącego $G$ 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 uczącego 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 $c_{ij}$ jest $j$-tym zalążkiem znalezionym na podstawie podpróby $G_N^i$.

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

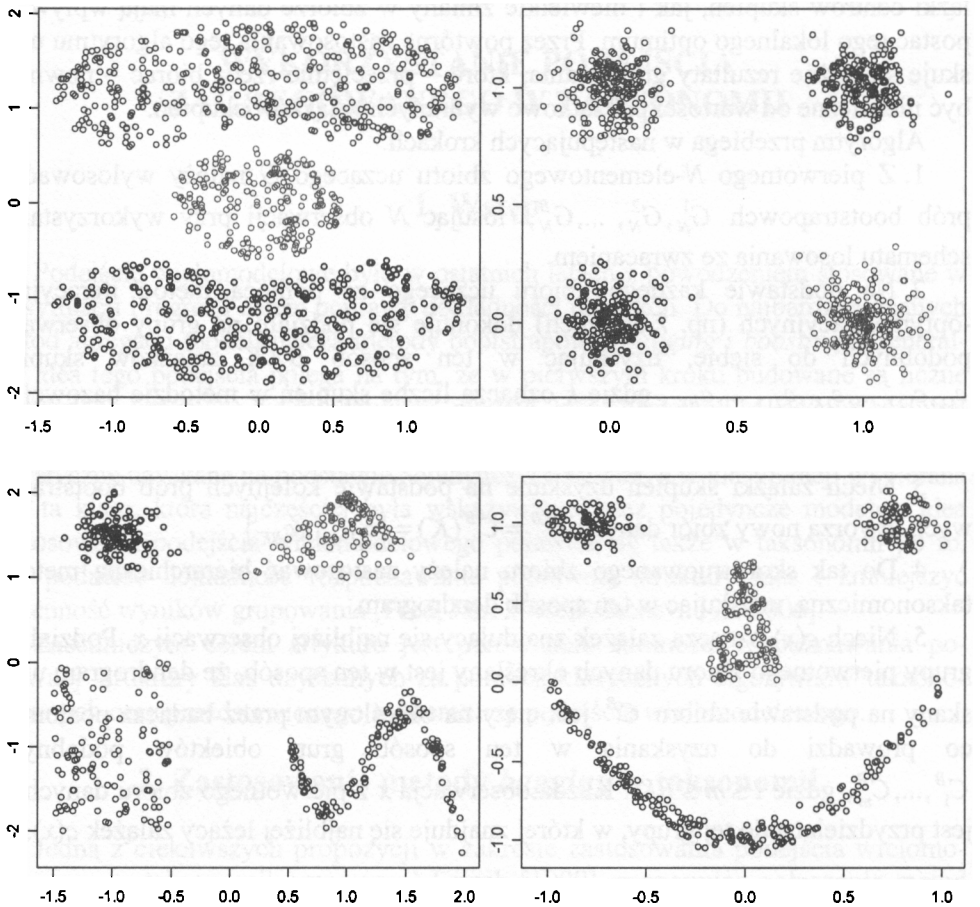
4. Do tak skonstruowanego zbioru należy zastosować hierarchiczną metodę taksonomiczną, uzyskując w ten sposób dendrogram.

5. Niech $c(x)$ oznacza zalążek znajdujący się najbliżej obserwacji $x$. 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 w ten sposób grup obiektów podobnych $C_1^B, ..., C_m^B$, gdzie $1 \le m \le BK$. Każda obserwacja $x$ z pierwotnego zbioru danych $G$ jest przydzielana do tej grupy, w której znajduje się najbliżej leżący zalążek $c(x)$.

## 3. Wyniki badań empirycznych

W badaniach postanowiono porównać zdolność do odkrywania poprawnej struktury klas omawianego podejścia wielomodelowego z następującymi klasycznymi algorytmami taksonomicznymi: metodą *k*-średnich, *c*-średnich, która jest rozmytą wersją metody *k*-średnich opracowaną przez Bezdeka [1981], oraz metodą *k*-medoidów, która jest bardziej odporną wersją metody *k*-średnich opracowaną przez Kaufmana i Rousseeuwa [1990].

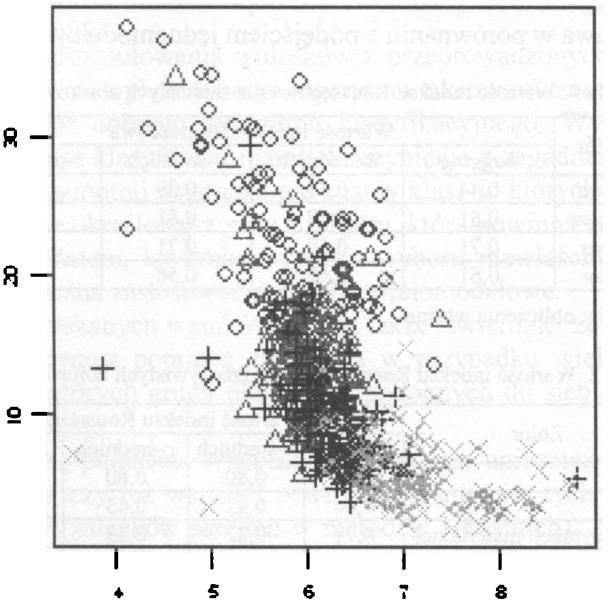
W badaniach zastosowano sztucznie wygenerowane zbiory danych, które standardowo wykorzystywane są w badaniach taksonomicznych[2]. Ich struktura przedstawiona jest na rys. 1. Wszystkie one charakteryzują się wyraźnie separowalnymi grupami wygenerowanymi w przestrzeni dwuwymiarowej. Ponadto zastosowane zostały także rzeczywiste zbiory danych, które stosowane są w klasyfikacji do budowy i oceny modelu, a więc takie, w których przynależność obiektów do klas jest znana. Informacja o liczbie klas jest traktowana jako informacja *a priori* o liczbie grup. Takie podejście jest również często stosowane przez badaczy z dziedziny taksonomii.




Rys. 1. Zastosowane sztucznie wygenerowane zbiory danych,
u góry – *Cassini* i *Corners*; na dole – *Shapes* i *Smiley*
Źródło: opracowanie własne.

[2] Zbiory te znajdują się w bibliotece *mlbench* w programie R.

Dwa spośród zastosowanych rzeczywistych zbiorów[3] – *Spam* oraz *Boston* – to zbiory standardowo wykorzystywane w badaniach porównawczych, udostępniane przez Uniwersytet Kalifornijski [Blake, Keogh, Merz 1988]. Natomiast dwa kolejne powstały na podstawie badań budżetów gospodarstw domowych i służą klasyfikacji obserwacji ze względu na ocenę sytuacji materialnej gospodarstwa domowego oraz ocenę dochodów gospodarstwa domowego. W przeciwieństwie do sztucznie generowanych, rzeczywiste zbiory danych należą do zbiorów wielowymiarowych, w których obserwacje nie są już tak wyraźnie separowalne. Przykładowy podział na grupy obserwacji podobnych do siebie w przestrzeni dwóch zmiennych, które są najsilniej związane ze zmienną objaśnianą dla zbioru *Boston*, ilustruje rys. 2.




Rys. 2. Przynależność obserwacji do grup obiektów podobnych do siebie w przestrzeni dwuwymiarowej (zbiór *Boston*)

Źródło: opracowanie własne.

Wszystkie obliczenia zostały wykonane w programie R, przy wykorzystaniu odpowiednio algorytmów: `kmeans` z biblioteki *stats*, `cmeans` z biblioteki *e1071* oraz `pam` z biblioteki *cluster*. Omawiane podejście wielomodelowe, wykorzystujące metodę *bagging*, znajduje się w bibliotece *e1071* pod nazwą `bclust`. W podejściu tym losowano $B = 10$ prób bootstrapowych, a jako algorytm bazowy przyjęto metodę

[3] Zbiór *Spam* służy do klasyfikacji poczty elektronicznej na listy i pocztę niechcianą. Zbiór *Boston* służy określaniu wartości nieruchomości i standardowo wykorzystywany jest w regresji. Na potrzeby niniejszego badania przekodowano wartość pierwotnej zmiennej zależnej, korzystając z kwartyli, oraz usunięto z niego jedną zmienną niemetryczną.

*k*-średnich z parametrem $k = 50$. W efekcie uzyskano zbiór danych zawierający 500 obserwacji, który następnie poddano grupowaniu metodą środka ciężkości.

Do pomiaru poprawności odkrytej struktury klas zastosowano indeks Rousseeuwa (*silhouette index*). Im wyższą wartość przyjmie ten indeks, tym silniejsza jest struktura klas odkrywana przez algorytm.

Na podstawie wyników uzyskanych dla zbiorów sztucznych (tab. 1) można stwierdzić, że generalnie wartości indeksu Rousseeuwa dla podejścia wielomodelowego wykazują wyższe wartości niż w przypadku podejścia jednomodelowego. Wyjątkiem są tylko zbiory *Shapes* i *Corners*, gdzie podejście wielomodelowe daje takie same wyniki jak metoda *c*-średnich i *k*-medoidów. Natomiast w żadnym przypadku podejście wielomodelowe nie doprowadziło do obniżenia wartości indeksu Rousseeuwa w porównaniu z podejściem jednomodelowym.

Tabela 1. Wartość indeksu Rousseeuwa dla sztucznych zbiorów danych

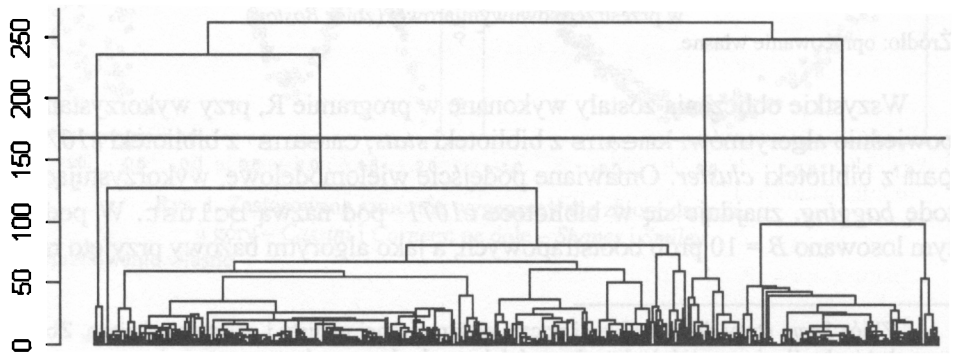
| Zbiór | Wartość indeksu Rousseeuwa | | | |
|---|---|---|---|---|
| | *bagging* | *k*-średnich | *c*-średnich | *k*-medoidów |
| *Cassini* | 0,44 | 0,41 | 0,39 | 0,36 |
| *Corners* | 0,81 | 0,67 | 0,81 | 0,81 |
| *Shapes* | 0,71 | 0,59 | 0,71 | 0,71 |
| *Smiley* | 0,61 | 0,55 | 0,56 | 0,59 |

Źródło: obliczenia własne.

Tabela 2. Wartość indeksu Rousseeuwa dla rzeczywistych zbiorów danych

| Zbiór | Wartość indeksu Rousseeuwa | | | |
|---|---|---|---|---|
| | *bagging* | *k*-średnich | *c*-średnich | *k*-medoidów |
| *Spam* | 0,85 | 0,80 | 0,80 | 0,74 |
| *Boston* | 0,66 | 0,42 | 0,43 | 0,56 |
| Ocena sytuacji materialnej | 0,71 | 0,27 | 0,23 | 0,25 |
| Ocena dochodów | 0,68 | 0,3 | 0,21 | 0,23 |

Źródło: obliczenia własne.




Rys. 3. Dendrogram będący podstawą grupowania obiektów dla zbioru *Boston*
Źródło: opracowanie własne.

Natomiast wyniki dla rzeczywistych zbiorów danych (tab. 2) ujawniają już wyraźną przewagę podejścia wielomodelowego w porównaniu z jednomodelowym. Za każdym razem indeks Rousseeuwa wykazuje wyższe wartości dla proponowanego podejścia wielomodelowego niż dla klasycznych algorytmów taksonomicznych.

Na rys. 3 pokazany jest przykładowy dendrogram będący podstawą grupowania obiektów z pierwotnego zbioru danych dla zbioru *Boston*. W przypadku tego zbioru zakładano, że obserwacje należą do jednej z czterech możliwych grup obserwacji i na podstawie tego dendrogramu można stwierdzić, że grupy te są wyraźnie widoczne.

## 4. Wnioski

Przechodząc do sformułowania wniosków z przeprowadzonych analiz, należy na wstępie zauważyć, że wybór dobrego algorytmu taksonomicznego jest znacznie trudniejszy niż wybór dobrego algorytmu klasyfikacyjnego. Wynika to przede wszystkim z tego, że w klasyfikacji mamy do czynienia z zagadnieniem uczenia z nauczycielem. W taksonomii natomiast nie znamy klas, do których należą obiekty, a tym samym brakuje określonej z góry struktury, która powinna zostać rozpoznana przez algorytm. Zatem, by ominąć ryzyko wyboru niewłaściwego algorytmu taksonomicznego, można zastosować podejście wielomodelowe.

Na podstawie uzyskanych wyników można także stwierdzić, że agregacja efektów grupowania przynosi poprawę zwłaszcza w przypadku wielowymiarowych zbiorów danych, w których grupy obserwacji podobnych do siebie nie są już tak łatwo separowalne.

Dodatkową zaletą podejścia wielomodelowego jest uniezależnienie wyników od wybranej metody czy też wartości pewnych parametrów tych metod (np. początkowo wybranych zalążków skupień w metodzie *k*-średnich). Agregacja wyników zatem pozwala na stabilizację rezultatów grupowania.

## Literatura


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

Blake C., Keogh E., Merz C.J. (1988), *UCI Repository of Machine Learning Databases,* Department of Information and Computer Science, University of California, Irvine.

Breiman L. (1996), *Bagging Predictors*, „Machine Learning, 26(2), s. 123-140.

Fred N.L., Jain A.K. (2005), *Combining Multiple Clusterings Using Evidence Accumulation*, „IEEE Transactions on PAMI", 27(6), s. 835-850.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Kaufman L., Rousseeuw P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Kuncheva L.I., Hadjitodorov S.T., Todorova L.P. (2006), *Experimental Comparison of Cluster Ensemble Methods*, „Proc FUSION 2006", Florence, Italy.
Leisch F. (1999), *Bagged Clustering*, Adaptive Information Systems and Modeling in Economics and Management Science, Working Paper 51, SFB.

## CLUSTER ENSEMBLE

### Summary

Ensemble methods are used in classification and regression to achieve better prediction accuracy. Recent research reveals that ensemble methods can be used also in taxonomy in order to gain better and more robust objects' classification [Fred, Jain 2005; Kuncheva et al. 2006]. Moreover aggregated approach decreases the risk of gaining a wrong classification because of choosing an unsuitable algorithm.

The main aim of the article is to show the possibility of applying one of the most popular ensemble methods, which is bagging [Breiman 1996] in taxonomy. We also show the results of research that main aim was to compare the results of classification with using both classical and ensemble methods with the existing class structure.