PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

**Dorota Rozmus**

Akademia Ekonomiczna w Katowicach

# ZASTOSOWANIE MACIERZY WSPÓŁWYSTĄPIEŃ W METODZIE *BAGGING* W TAKSONOMII

## 1. Wstęp

W ostatnich latach podejście wielomodelowe oparte na modelach zagregowanych z powodzeniem stosowane było w klasyfikacji w celu podniesienia dokładności i stabilności metod klasyfikacji. Idea tego podejścia polega na tym, że w pierwszym etapie budowane są liczne pojedyncze i różniące się między sobą modele, które następnie różnymi operatorami są łączone w model zagregowany. W klasyfikacji najczęściej stosowane jest głosowanie majoryzacyjne, a więc wybierana jest ta klasa, która najczęściej wskazywana była przez pojedyncze modele. Do najbardziej znanych metod agregacji należą: *bagging* [Breiman 1996], oparty na losowaniu prób boostrapowych, oraz *boosting* [Freund 1990], polegający na nadawaniu wyższych wartości wag błędnie sklasyfikowanym obiektom.

W ostatnich latach idea podejścia wielomodelowego pojawiła się także w taksonomii w celu podniesienia poprawności i odporności algorytmów taksonomicznych. Zasadniczy cel tego podejścia polega na połączeniu wyników wielokrotnie przeprowadzonego grupowania. Zagadnienie łączenia w taksonomii można sformułować następująco: mając dane wyniki wielokrotnie przeprowadzonej klasyfikacji, znajdź zagregowany podział ostateczny o lepszej jakości. Liczne badania w tej dziedzinie ugruntowały nowy obszar w tradycyjnej taksonomii. Istnieje kilka możliwości zastosowania idei podejścia zagregowanego w dziedzinie uczenia bez nauczyciela:

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

Badanie niniejsze wypływa z trzech źródeł. Pierwszym jest zaproponowane przez Pekalską i Duin [2000] w dyskryminacji podejście oparte na opisie zbioru obiektów przez podobieństwo (bądź niepodobieństwo). W tradycyjnym zagadnieniu modele klasyfikacyjne budowane są na podstawie zbioru zawierającego zmienne charakteryzujące poszczególne obserwacje. Alternatywą dla tego klasycznego opisu obiektów może być zaproponowane przez wspomnianych autorów podejście oparte na macierzy podobieństwa (odległości) między obiektami. W ujęciu tym obiekty opisywane są przez pewną miarę obrazującą stopień podobieństwa między obserwacjami ze zbioru danych. Model zatem jest budowany na podstawie tej macierzy, która traktowana jest jako zbiór opisujący obserwacje.

Drugim źródłem jest zaproponowana przez Fred i Jain [2002] idea łączenia wyników wielokrotnie dokonanej klasyfikacji w celu konstrukcji macierzy współwystąpień. Biorąc jednoczesne wystąpienie pary obiektów w tej samej klasie za wskazówkę istnienia związku między nimi, pierwotny zbiór obserwacji przekształcany jest w $n \times n$-wymiarową macierz, która opisuje podobieństwo między obiektami. Ostateczne grupowanie dokonywane jest na podstawie uzyskanej macierzy współwystąpień, która traktowana jest jak zbiór danych.

Trzecim źródłem jest zaproponowane przez Dudoit i Fridlyand [2003] podejście wielomodelowe w taksonomii wykorzystujące ideę metody *bagging*.

Zasadniczym celem badań jest porównanie zdolności do odkrywania poprawnej struktury klas dla metody *bagging* w taksonomii zastosowanej dla pierwotnego zbioru danych oraz skonstruowanej macierzy współwystąpień.

## 2. Zastosowane algorytmy

W badaniu zastosowane zostało dwuetapowe podejście: najpierw konstruowana była macierz współwystąpień, która potem posłużyła jako macierz danych dla metody *bagging*. Dokładniej, etapy badania mogą zostać sformułowane następująco:

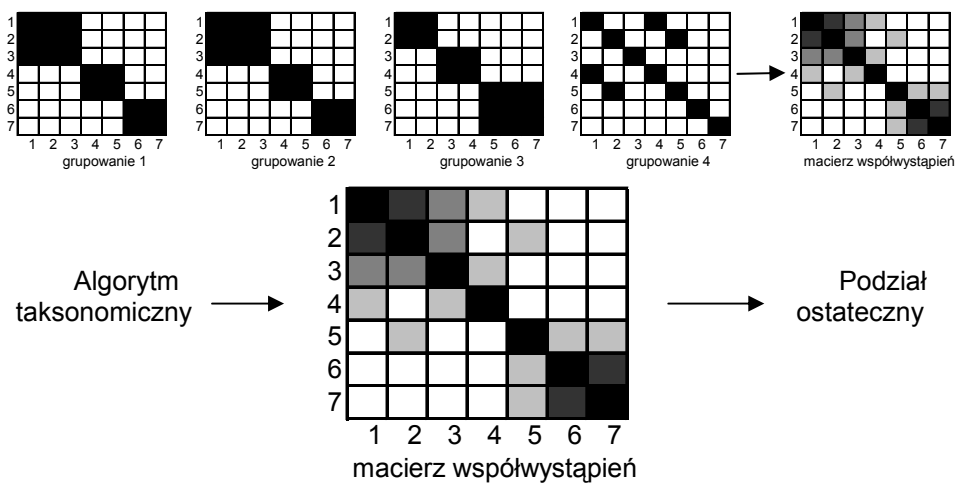
- **Wielokrotna klasyfikacja.** Dla założonej liczby składowych $C$ macierzy współwystąpień należy dokonać grupowania obiektów np. za pomocą metody $k$-średnich, uzyskując różniące się między sobą rezultaty dzięki losowo wybranym zalążkom skupień.
- **Agregacja.** U podstaw tego podejścia leży założenie, że obiekty należące do tej samej grupy najprawdopodobniej będą lokowane w tej samej klasie wśród tych $C$ składowych podziałów. Biorąc zatem współwystąpienie pary obiektów w tej samej grupie za wskazówkę istnienia związku między nimi, wyniki klasyfikacji uzyskane dzięki wielokrotnie zastosowanej metodzie $k$-średnich można przekształcić w macierz współwystąpień o wymiarach $n \times n$:

$$co\_assoc(a,b) = votes_{ab}, \tag{1}$$

gdzie $votes_{ab}$ zlicza, ile razy para obiektów *a* i *b* zaliczona została do tej samej grupy wśród tych *C* składowych klasyfikacji.

- **Ostateczny podział.** W celu określenia podziału ostatecznego należy zastosować dowolny algorytm taksonomiczny do skonstruowanej wcześniej macierzy współwystąpień, traktując ją jak macierz danych.

  Kroki służące do konstrukcji macierzy współwystąpień przedstawione są na rys. 1.




Rys. 1. Konstrukcja macierzy współwystąpień i ostateczna klasyfikacja
Źródło: opracowanie własne.

W badaniu natomiast do tak przygotowanego zbioru danych zastosowana została jedna z zaproponowanych procedur podejścia wielomodelowego w taksonomii, a mianowicie metoda *bagging* [Dudoit, Fridlyand 2003]. Kroki algorytmu można ująć następująco:

Dla założonej liczby klas *K* i założonej liczby prób bootstrapowych *B*:

1. Dokonaj klasyfikacji obserwacji z pierwotnego zbioru danych *S* za pomocą iteracyjno-optymalizacyjnego algorytmu taksonomicznego *A*, uzyskując w ten sposób etykiety klas $A(x_i, S) = \hat{y}_i$ dla każdej obserwacji $x_i$, $i = 1, \ldots, n$.

2. Skonstruuj *b*-tą próbę bootstrapową $S^b = (x_1^b, ..., x_n^b)$.

3. Dokonaj klasyfikacji obserwacji w skonstruowanej próbie bootstrapowej $S^b$ za pomocą algorytmu taksonomicznego *A*, uzyskując podział na klasy: $A(x_i^b, S^b)$ dla każdej obserwacji w zbiorze $S^b$.

4. Dokonaj permutacji etykiet klas przyznanych obiektom w próbie bootstrapowej $S^b$ tak, by zachodziła jak największa zbieżność z etykietami klas przyznanych obserwacjom z oryginalnego zbioru danych *S*. Niech $P_K$ oznacza zbiór

wszystkich permutacji zbioru liczb całkowitych 1,…, *K*. Znajdź permutację $\tau^b \in P_K$ maksymalizującą:

$$\sum_{i=1}^{n} I(\tau(A(x_i^b, S^b)) = A(x_i^b, S)), \tag{2}$$

gdzie $I(\cdot)$ to funkcja wskaźnikowa równa 1, gdy stwierdzenie jest prawdziwe, 0 w przypadku przeciwnym.

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

$$\arg\max_{1\le k\le K} \sum_{b:x_i\in S^b} I(\tau^b(A(x_i, S^b)) = k). \tag{3}$$

## 3. Badania empiryczne

Jak już na początku zostało wspomniane, celem badań empirycznych jest porównanie zdolności do odkrywania poprawnej struktury klas dla:

- metody *bagging* zastosowanej do oryginalnego zbioru danych,
- metody *bagging* opartej na macierzy współwystąpień.

Wśród wyników empirycznych zawarto także wyniki dla klasycznej metody *k*-średnich, by pokazać, że podejście wielomodelowe jest w stanie zapewnić rezultaty lepsze niż klasyczne podejście bazujące na jednokrotnym zastosowaniu algorytmu taksonomicznego.

Jeżeli chodzi o metodę *k*-średnich, to parametr *k* był równy liczbie klas. Zastosowana procedura badawcza w podejściu wielomodelowym była następująca: w pierwszym kroku konstruowana była macierz współwystąpień. W tym celu dokonano 10-krotnej klasyfikacji zbioru danych z zastosowaniem metody *k*-średnich. W drugim kroku do określenia ostatecznego podziału stosowana była zaproponowana przez Dudoit i Fridlyand metoda *bagging*, przy czym jako zbiór danych raz stosowany był oryginalny zbiór pierwotnych zmiennych opisujących obiekty, a raz była to skonstruowana w kroku pierwszym macierz współwystąpień. Liczba składowych (parametr *B*) w metodzie *bagging* była równa 50.

Większość obliczeń została wykonana w programie **R** z zastosowaniem algorytmu `kmeans` z biblioteki *stats*, natomiast metodę *bagging* oprogramowano w bibliotece *clue* pod nazwą procedury `cl_bag`. Do konstrukcji macierzy współwystąpień wykorzystywany był natomiast program autorski.

Do pomiaru poprawności odkrytej struktury klas zastosowano indeks Randa [1971].

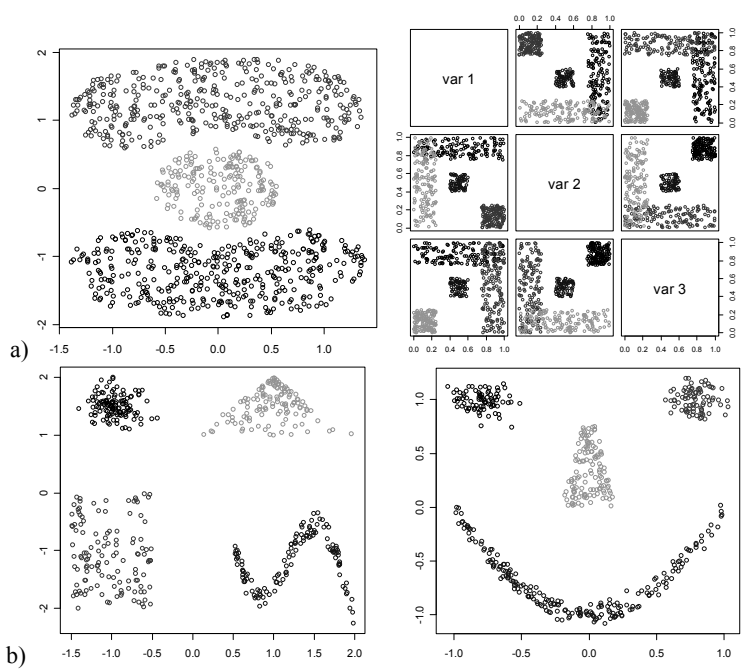
W badaniach zastosowano rzeczywiste oraz sztucznie generowane zbiory danych (ich krótka charakterystyka znajduje się w tab. 1). Wśród rzeczywistych zbiorów zastosowano dane standardowo wykorzystywane w dyskryminacji w celu bu-

dowy i oceny dokładności modelu[1]. Są to zbiory, w których przynależność obiektów do klas jest znana. Ta informacja jest traktowana jako informacja *a priori* o liczbie grup; takie podejście jest bardzo często stosowane przez badaczy z dziedziny taksonomii. Sztuczne zbiory natomiast to również standardowo wykorzystywane dane w badaniach porównawczych w taksonomii. *Cassini*, *Shape* i *Smiley* mają wyraźnie separowalne klasy wygenerowane w przestrzeni dwuwymiarowej, *Cuboids* natomiast jest problemem czteroklasowym wygenerowanym w trzech wymiarach. Ich charakterystyka znajduje się na rys. 2.

Tabela 1. Charakterystyka zastosowanych zbiorów danych

| Zbiór danych | Liczba obiektów | Liczba cech | Liczba klas |
|---|---|---|---|
| *Boston* | 506 | 13 | 4 |
| *Ecoli* | 336 | 8 | 8 |
| *Glass* | 214 | 10 | 6 |
| *Cassini* | 500 | 2 | 3 |
| *Cuboids* | 500 | 3 | 4 |
| *Shapes* | 500 | 2 | 4 |
| *Smiley* | 500 | 2 | 4 |

Źródło: opracowanie własne.




Rys. 2. Sztuczne zbiory danych; a) *Cassini* i *Cuboids*, b) *Shapes* i *Smiley*
Źródło: opracowanie własne.

[1] Udostępniane są przez Uniwersytet Kalifornijski [Blake, Keogh, Merz 1988].

Tabela 2. Wartości indeksu Randa

| Zbiór danych | INDEKS RANDA | | |
|---|---|---|---|
| | metoda *k*-średnich | *bagging* + oryginalny zbiór danych | *bagging* + macierz współwystąpień |
| *Boston* | 0,615 | 0,615 | 0,665 |
| *Ecoli* | 0,811 | 0,830 | 0,842 |
| *Glass* | 0,680 | 0,684 | 0,678 |
| *Cassini* | 0,801 | 0,806 | 0,815 |
| *Cuboids* | 0,945 | 0,996 | 0,998 |
| *Shapes* | 0,836 | 0,874 | 0,874 |
| *Smiley* | 0,829 | 0,864 | 0,884 |

Źródło: obliczenia własne.

Porównując wyniki empiryczne dla metody *bagging* opartej na klasycznym zbiorze danych i metody *bagging* opartej na macierzy współwystąpień, można zauważyć, że generalnie macierz współwystąpień pozwala uzyskać wyższe wartości indeksu Randa w porównaniu z podejściem opartym na tradycyjnym zbiorze danych. Natomiast porównując podejście wielomodelowe z jednokrotnym zastosowaniem metody *k*-średnich, widzimy, że niemalże za każdym razem metoda *bagging* zapewnia lepsze rezultaty. Wyjątkiem jest jedynie zbiór *Boston*, który w przypadku metody *bagging* opartej na oryginalnym zbiorze danych daje taki sam rezultat jak metoda *k*-średnich.

## 4. Wnioski

Należy zauważyć, że wybór dobrego algorytmu taksonomicznego jest znacznie trudniejszy niż wybór dobrego algorytmu klasyfikacyjnego. Wynika to przede wszystkim z faktu, że w klasyfikacji 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. Zatem by ominąć ryzyko wyboru niewłaściwego algorytmu taksonomicznego, można zastosować podejście wielomodelowe w celu połączenia wyników klasyfikacji różnych algorytmów. Każdy pojedynczy algorytm ma swoje mocne i słabe strony, ale wydaje się, że ich łączne zastosowanie przyniesie pozytywny efekt kompensacji.

Drugą zaletą podejścia wielomodelowego jest uniezależnienie wyników od jednej 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.

Kolejną zaletą podejścia wielomodelowego jest zwiększenie odporności algorytmów taksonomicznych, a więc ich mniejsza wrażliwość na szum i obserwacje oddalone oraz zmienność związaną np. z zastosowaniem metod losowania bootstrapowego.

Ostatni wniosek, który sformułować można na podstawie wyników empirycznych, jest taki, że generalnie metoda *bagging* pozwala na poprawę dokładności klasyfikacji w porównaniu z podejściem tradycyjnym; poprawę można dostrzec zwłaszcza wtedy, gdy jako zbiór danych zastosujemy macierz współwystąpień.

Chociaż metoda ta została zilustrowana jedynie z zastosowaniem metody *k*-średnich do konstrukcji macierzy współwystąpień, w dalszych badaniach do jej budowy warto zastosować także inne metody, np. metodę *c*-średnich. Interesującym kierunkiem badań byłoby zastosowanie również i innych schematów losowania, podobnych do losowania adaptacyjnego w metodzie *boosting* [Freund 1999].

## Literatura

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" no 26(2), s. 123-140.

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

Fred N.L., Jain A.K. (2002), *Combining multiple clusterings using evidence accumulation*, „IEEE Transactions on PAMI" no 27(6), s. 835-850.

Freund Y. (1990), *Boosting a weak learning algorithm by majority*, „Proceedings of the Third Annual Workshop on Computational Learning Theory", s. 202-216.

Freund Y. (1999), *An adaptive version of the boost by majority algorithm*, „Proceedings of the Twelfth Annual Conference on Computational Learning Theory", s. 293-318.

Pekalska E., Duin R.P.W. (2000), *Classifiers for dissimilarity-based pattern recognition*, [w:] A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquezar, A.K. Jain, J. Kittler (red.), „Proceedings Fifteenth International Conference on Pattern Recognition", IEEE Computer Society, Press, Los Alamitos, s. 12-16.

Rand W.M. (1971), *Objective criteria for the evaluation of clustering methods*, „Journal of the American Statistical Association" no 336, s. 846-850.

## APPLYING OF CO-OCCURRENCE MATRIX IN BAGGING IN TAXONOMY

### Summary

Ensemble approach has been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. Recently, analogous techniques for cluster analysis have been suggested.

The main aim of this article is to introduce one of the ensemble techniques in taxonomy [Dudoit, Fridlyand 2003] and to compare the accuracy of classification with traditional algorithms.

The novelty of this article flows from the fact of joining proposed by Dudoid and Fridlyand [2003] *bagging* technique in taxonomy with the concept of describing the objects by co-occurrence (co-association) matrix [Fred, Jain 2002]. In this matrix objects are described by some distance measure showing the similarity between objects.