PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Eugeniusz Gatnar
Akademia Ekonomiczna w Katowicach

WYKORZYSTANIE METOD TAKSONOMICZNYCH DO BUDOWY MODELI ZAGREGOWANYCH

1. Wstęp

Sukces podejścia polegającego na agregacji modeli w analizie dyskryminacyjnej zależy od zróżnicowania modeli składowych (lokalnych) $C_1(\mathbf{x}),\ldots,C_K(\mathbf{x})$.

Jak udowodnili Tumer i Ghosh [1996], błąd klasyfikacji modelu zagregowanego $C^*(\mathbf{x})$ maleje wraz ze spadkiem stopnia podobieństwa („korelacji") modeli składowych. Celowe staje się więc łączenie tylko takich modeli, które jak najbardziej różnią się od siebie.

Jednym z rozwiązań mogących zapewnić to, że modele składowe będą maksymalnie niepodobne jest wykorzystanie metod taksonomicznych. W artykule zaproponowano metodę budowy modeli zagregowanych, która polega na wygenerowaniu dużej liczby modeli, np. $K = 300$, i pogrupowaniu ich w klasy, a następnie połączeniu reprezentantów klas w jeden model $C^*(\mathbf{x})$.

W związku z tym, że brak jest głębszych badań nad omawianym zagadnieniem, kolejnym celem artykułu stała się analiza porównawcza metod taksonomicznych, które mogą być wykorzystane w proponowanym podejściu. Pojawił się także problem konstrukcji macierzy niepodobieństwa modeli, ponieważ istnieje kilka konkurencyjnych miar mogących służyć do tego celu. W dalszej części pracy zbadano więc także to, jak wielkość błędu klasyfikacji zależy od wyboru odpowiedniej miary zróżnicowania.

2. Metody łączenia modeli

Agregacja modeli polega na tym, że jeśli mamy zbiór uczący $U = \{(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_N, y_N)\}$, zawierający obiekty o znanej przynależności do klas

(y), to możemy na jego podstawie utworzyć próby uczące $U_1,\ldots,U_K$. Próby te służą do budowy modeli indywidualnych $C_1(\mathbf{x}),\ldots,C_K(\mathbf{x})$, które są następnie łączone (np. na zasadzie majoryzacji), dając w rezultacie model zagregowany:

$$C^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{k=1}^{K} I(C_k(\mathbf{x}) = y) \right\}, \quad (1)$$

gdzie I jest funkcją wskaźnikową, tj. $I(a)=1$, gdy a jest prawdziwe, zaś $I(a)=0$ w przeciwnym wypadku.

W literaturze można znaleźć wiele metod łączenia modeli składowych. Różnią się one albo sposobem, w jaki tworzone są modele składowe, tj. sposobem konstrukcji prób uczących, albo sposobem łączenia wyników klasyfikacji dla obiektów ze zbioru testowego.

Ogólnie rzecz biorąc, wszystkie metody budowy modeli składowych możemy podzielić na trzy grupy:

- wykorzystujące obiekty ze zbioru uczącego, np. *bagging* [Breiman 1996], *boosting* [Freund, Shapire 1997] oraz *arcing* [Breiman 1998],
- wykorzystujące zmienne charakteryzujące obiekty w zbiorze uczącym, np. *random subspaces* [Ho 1998] oraz *random forests* [Breiman 2001],
- manipulujące wartościami zmiennej objaśnianej (y), np. *error-correcting output coding* [Dietterich, Bakiri 1995].

3. Proponowana metoda budowy modeli zagregowanych

Zagadnienie wykorzystania metod taksonomicznych do budowy modeli zagregowanych nie było poruszane w literaturze przedmiotu zbyt często, a jeśli już, to w nieco innym znaczeniu niż przedstawiona w tym artykule propozycja.

Partridge i Yates [1996] zaproponowali metodę polegającą na utworzeniu wielu modeli, z których wybierane są te, które najbardziej różnią się od siebie (*overproduce and choose*). Wspomniany wybór był sterowany za pomocą statystyki kappa (5).

Z kolei Giacinto, Roli i Fumera [2000] wykorzystali miarę (4) do wyznaczenia odległości między modelami oraz metodę najbliższego sąsiada do łączenia modeli składowych w klasy.

Giacinto i Roli [2001] swoje podejście do budowy modeli zagregowanych oparli na iteracyjnym łączeniu kolejnych modeli w jeden model zagregowany (aż do momentu, gdy wszystkie modele zostaną połączone) i wyborze w każdym kroku reprezentanta istniejącego już zbioru modeli. Modele te w ostatnim kroku procedury były łączone w jeden model przy użyciu formuły (1).

W proponowanej metodzie M modeli składowych zostaje podzielonych na K rozłącznych podzbiorów $\{S_1,\ldots,S_K\}$ za pomocą pewnej metody taksonomicznej. Następnie z każdego skupienia S_k jest wybierany model, który ma najmniejszy błąd klasyfikacji. Modele te następnie podlegają agregacji za pomocą formuły (1).

Algorytm tej metody składa się z następujących kroków:

1. Zbuduj M pojedynczych modeli na podstawie zbioru uczącego.
2. Utwórz macierz niepodobieństwa $D=[d_{ij}]$, wykorzystując jedną z miar służących do oceny stopnia zróżnicowania modeli.
3. Podziel zbiór modeli $C_1(\mathbf{x}),\ldots,C_M(\mathbf{x})$ na K podzbiorów, gdzie $K=2,\ldots,M-1$, za pomocą jednej z hierarchicznych metod analizy skupień.
4. Znajdź optymalną liczbę klas K.
5. Wybierz z każdej klasy jeden model $C_k(\mathbf{x})$, np. ten, który ma najmniejszy błąd klasyfikacji.
6. Połącz K wybranych w jeden model zagregowany za pomocą formuły (1).

4. Konstrukcja macierzy niepodobieństwa

Aby móc wykorzystać metody hierarchicznej analizy skupień, należy zbudować macierz niepodobieństwa dla wszystkich par modeli. W tym celu można wykorzystać jedną z miar zróżnicowania modeli, przedstawionych w pracy Gatnara [2006].

Ponieważ różni autorzy rekomendowali różne miary, należało dokonać porównania tych miar i wybrać najlepszą. W eksperymentach obliczeniowych przedstawionych w niniejszej pracy wykorzystano sześć z nich. Wszystkie te miary wykorzystują macierz wartości zero-jedynkowych (*oracle labels*):

$$O_k(\mathbf{x}_i)=\begin{cases}1 & C_k(\mathbf{x}_i)=y_i\\ 0 & C_k(\mathbf{x}_i)\neq y_i\end{cases}, \tag{2}$$

gdzie wartość $O_k(\mathbf{x}_i)=1$ oznacza, że model $C_k(\mathbf{x})$ poprawnie rozpoznał klasę obiektu $\mathbf{x}_i$, zaś $O_k(\mathbf{x}_i)=0$ oznacza to, że się pomylił. Wtedy zgodność predykcji pary modeli $C_k(\mathbf{x})$ i $C_m(\mathbf{x})$ można analizować za pomocą tablicy kontyngencji o wymiarach 2×2 (tab. 1).

Tabela 1. Tablica kontyngencji dla wyników klasyfikacji

	O_m=1	O_m=0
O_k=1	*a*	*b*
O_k=0	*c*	*d*

Źródło: opracowanie własne.

Najbardziej znaną miarą zgodności jest zero-jedynkowa wersja współczynnika korelacji Pearsona:

$$p(i,j)=\frac{ad-bc}{(a+b)(c+d)(a+c)(b+d)}. \tag{3}$$

Można ją wykorzystać do budowy macierzy **D** po przekształceniu na niepodobieństwo.

Z kolei Giacinto i Roli [2001] zaproponowali miarę, którą nazwali miarą podwójnego błędu (*double fault*):

$$d(i,j)=\frac{d}{a+b+c+d}, \tag{4}$$

ponieważ jest frakcją obiektów błędnie sklasyfikowanych przez oba modele.

Partridge i Yates [1996] oraz Margineantu i Dietterich [1997] przedstawili miarę nazwaną zróżnicowaniem wewnątrzgrupowym (*within-set generalization diversity*). Ta miara jest po prostu statystyką κ (kappa), która mierzy poziom zgodności dwóch modeli klasyfikacyjnych:

$$p(i,j)=\frac{2(ac-bd)}{(a+b)(c+d)+(a+c)(b+d)}. \tag{5}$$

Skalak [1996] wykorzystał miarę niezgodności (*disagreement measure*), aby ocenić zróżnicowanie dwóch modeli:

$$d(i,j)=\frac{b+c}{a+b+c+d}. \tag{6}$$

Jest to frakcja liczby obiektów, które zostały błędnie sklasyfikowane przez przynajmniej jeden z modeli.

Kuncheva i inni w pracy [Kuncheva i in. 2000] zaproponowali statystykę Q Yule'a do oceny stopnia zróżnicowania modeli klasyfikacyjnych:

$$p(i,j)=\frac{ad-bc}{ad+bc}. \tag{7}$$

Jest to miara symetryczna i przyjmuje wartości pomiędzy -1 i 1, gdzie wartość 0 oznacza statystyczną niezależność dwóch modeli klasyfikacyjnych. Ponieważ ma ona jednak dosyć poważne wady, Gatnar [2006] zaproponował wykorzystanie współczynnika Hamanna do oceny zróżnicowania modeli składowych:

$$p(i,j)=\frac{(a+d)-(b+c)}{a+b+c+d}. \tag{8}$$

Jej wartość znajduje się w przedziale [-1,1], przy czym wartość 0 wskazuje na równą liczbę klasyfikacji zgodnych i niezgodnych, –1 oznacza idealną niezgodność, zaś 1 – idealną zgodność. Oczywiście, miary (5), (7) i (8) muszą zostać przekształcone na miary odległości (niepodobieństwa), aby mogły utworzyć macierz **D**.

5. Metody taksonomiczne i określenie liczby klas

Jak już wspomniano, w kroku 3 proponowanego algorytmu modele składowe są dzielone na klasy na podstawie macierzy niepodobieństwa **D**. W prowadzonych eksperymentach wykorzystano cztery znane hierarchiczne metody taksonomiczne:

- najbliższego sąsiedztwa,
- najdalszego sąsiedztwa,

- średniego połączenia między skupieniami,
- Warda.

Ich komputerowe implementacje znajdują się w bibliotece `cluster` programu R.

Z kolei w celu wyboru optymalnej liczby klas został wykorzystany wskaźnik kształtu (*silhouette index*), zaproponowany przez Rousseuwa [1987], również zaimplementowany w bibliotece `cluster`. Dla każdej obserwacji $\mathbf{x}$ jest liczona wielkość indeksu:

$$s(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{\max\{a(\mathbf{x}), b(\mathbf{x})\}}, \tag{9}$$

gdzie $a(\mathbf{x})$ to średnia wielkość odległości pomiędzy $\mathbf{x}$ a innymi obserwacjami w tym samym skupieniu, zaś $b(\mathbf{x})$ to odległość między $\mathbf{x}$ a najbliższym innym skupieniem. Optymalna liczba klas to ta, dla której zachodzi:

$$K = \arg\max_{k=2,\ldots,M-1} \left\{ \sum_{n=1}^{N} s(\mathbf{x}) \right\}. \tag{10}$$

6. Wybór reprezentantów skupień i wyniki eksperymentów

Kiedy struktura klas jest już znana, z każdego skupienia wybierany jest jeden model, który będzie podlegał łączeniu. Można rozważać kilka różnych strategii selekcji reprezentantów klas, jednak w proponowanym algorytmie są wybierane modele o najmniejszym błędzie predykcji dla zbioru testowego.

Aby zdecydować, którą z metod taksonomicznych należy zastosować w trzecim kroku proponowanej metody, przeprowadzono wiele eksperymentów obliczeniowych. Ich celem było także sprawdzenie przydatności przedstawionych wyżej miar zróżnicowania modeli. Wykorzystano 10 znanych zbiorów danych, które są standardowo wykorzystywane w analizach porównawczych metod dyskryminacji.

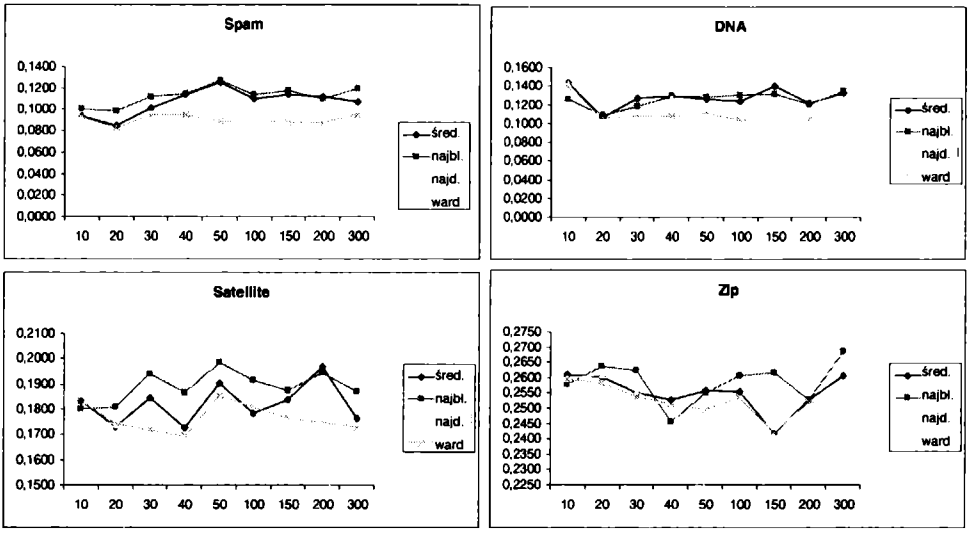
Charakterystykę zbiorów oraz ich podział na część uczącą i testową pokazano w tab. 2.

Tabela 2. Zbiory danych wykorzystywane w analizie dyskryminacyjnej

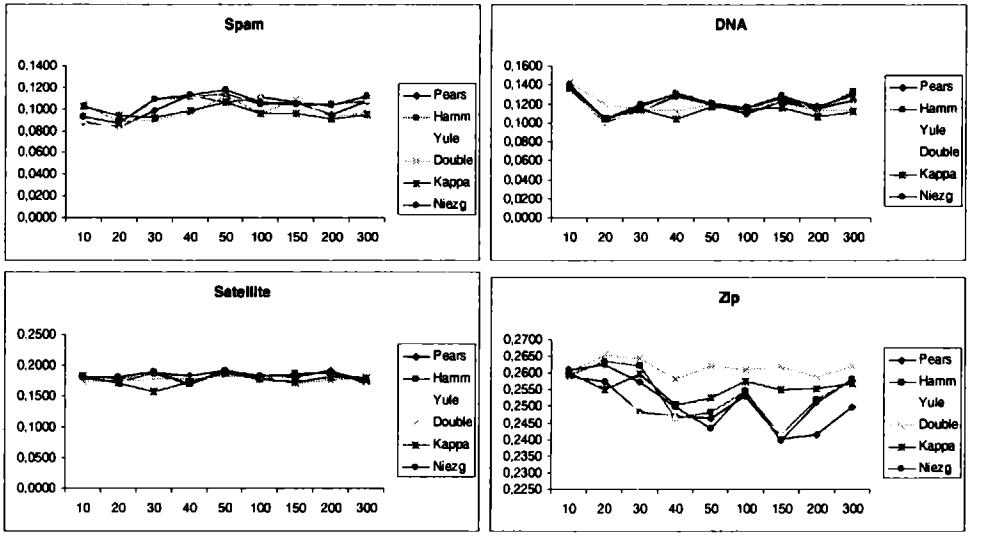
Zbiór danych	Liczba obserwacji w zbiorze uczącym	Liczba obserwacji w zbiorze testowym	Liczba zmiennych	Liczba klas
DNA	2 124	1 062	180	3
Letter	16 000	4 000	16	26
Satellite	4 290	2 145	36	6
Iris	100	50	4	3
Spam	3 000	1 601	57	2
Diabetes	512	256	8	2
Sonar	138	70	60	2
Vehicle	564	282	18	4
Soybean	455	228	34	19
Zip	7 291	2 007	256	10

Źródło: opracowanie własne.

Zbiory te znajdują się na stronie www.ics.uci.edu/~mlearn/MLRepository.html oraz w bibliotece mlbench dołączonej do programu R.



Rys. 1. Błędy klasyfikacji dla wybranych metod taksonomicznych
Źródło: obliczenia własne.



Rys. 2. Błędy klasyfikacji dla wybranych miar niepodobieństwa modeli
Źródło: obliczenia własne.

Dla każdego zbioru uczącego generowano 14 różnych zestawów modeli (na podstawie prób boostrapowych), liczących kolejno 10, 20, 30, 40, 50, 60, 70, 80,

90, 100, 150, 200, 250, 300 modeli indywidualnych. Wybraną klasą modeli były drzewa klasyfikacyjne, tworzone za pomocą procedury `rpart` w programie R [Therneau, Atkinson 1997].

Jeżeli chodzi o analizę porównawczą metod grupowania, to na rys. 1 pokazano błędy klasyfikacji dla 4 wybranych zbiorów danych.

Podobne analizy symulacyjne prowadzono dla przedstawionych 6 miar zróżnicowania, służących do budowy macierzy **D**. Wyniki porównań pokazano na rys. 2, przy czym należy dodać, iż wykorzystano tutaj metodę Warda.

7. Podsumowanie

Wyniki przeprowadzonej analizy porównawczej wskazują, że w zaproponowanym algorytmie budowy modeli zagregowanych najlepiej wykorzystać taksonomiczną metodę najdalszego sąsiada lub metodę Warda. Obie te metody zwyciężyły w porównaniach.

Z kolei jeśli chodzi o wybór miary zróżnicowania, to wyniki nie są jednoznaczne. W trzech przypadkach najbardziej dokładne modele dyskryminacyjne uzyskano, gdy macierz niepodobieństwa została utworzona za pomocą statystyki kappa (5) lub miary podwójnego błędu (4). Jeżeli zaś chodzi o zbiór ZIP, to najmniejszy błąd uzyskano, gdy do pomiaru odległości zastosowano miarę niezgodności (6) oraz współczynnik korelacji Pearsona (3). Należy jednak dodać, że miara Hamanna (8) także dawała dobre wyniki.

Literatura

Breiman L. (1998), *Arcing Classifiers*, „Annals of Statistics" 26, s. 801-849.

Breiman L. (1996), *Bagging Predictors*, „Machine learning" 24, s. 123-140.

Breiman L. (2001), *Random Forests*, „Machine Learning" 45, s. 5-32.

Dietterich T., Bakiri G. (1995), *Solving Multiclass Learning Problem via Error-Correcting Output Codes*, „Journal of Artificial Intelligence Research" 2, s. 263-286.

Freund Y., Schapire R.E. (1997), *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, „Journal of Computer and System Sciences" 55, s. 119-139.

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

Gatnar E. (2006), *Wykorzystanie miary Hamanna do oceny podobieństwa modeli w podejściu wielomodelowym, Taksonomia 13*, Prace Naukowe AE we Wrocławiu nr 1126, AE, Wrocław, s. 56-64.

Giacinto G., Roli F. (2001), *Design of Effective Neural Network Ensembles for Image Classification Processes*, „Image Vision and Computing Journal" 19, s. 699-707.

Giacinto G., Roli F., Fumera G. (2000), *Design of Effective Multiple Classifier Systems by Clustering of Classifiers*, Proc. of the Int. Conference on Pattern Recognition, ICPR'00, IEEE.

Ho T.K. (1998), *The Random Subspace Method for Constructing Decision Forests*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" 20, s. 832-844.
Kuncheva L., Whitaker C., Shipp D., Duin R. (2000), *Is Independence Good for Combining Classifiers*, Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, s. 168-171.
Margineantu M.M., Dietterich T.G. (1997), *Pruning Adaptive Boosting*, Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufmann, San Mateo, s. 211-218.
Partridge D., Yates W.B. (1996), *Engineering Multiversion Neural-Net Systems*, „Neural Computation" 8, s. 869-893.
Skalak D.B. (1996), *The Sources of Increased Accuracy for Two Proposed Boosting Algorithms*, Proceedings of the American Association for Artificial Intelligence AAAI-96, Morgan Kaufmann, San Mateo.
Therneau T.M., Atkinson E.J. (1997), *An Introduction to Recursive Partitioning Using the RPART Routines*, Mayo Foundation, Rochester.
Tumer K., Ghosh J. (1996), *Analysis of Decision Boundaries in Linearly Combined Neural Classifiers*, „Pattern Recognition" 29, s. 341-348.
Wolpert D. (1992), *Stacked Generalization*, „Neural Networks" 5, s. 241-259.

THE APPLICATION OF CLUSTER ANALYSIS IN AGGREGATION OF CLASSIFIERS

Summary

Combining multiple classifiers into an ensemble has proved to be very successful in the past decade. The key issue is the diversity of the component classifiers, because the most unrelated members the most accurate is the ensemble. In this paper we propose a new method of combining classifiers, that is based on clustering of the single models. We also compare different hierarchical clustering methods and diversity measures used to create the dissimilarity matrix.