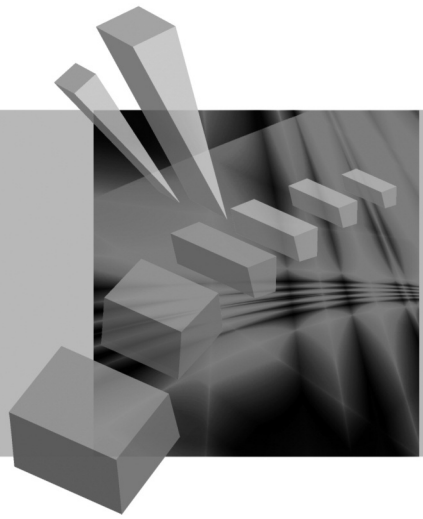
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon 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

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

PODEJŚCIE WIELOMODELOWE Z WYKORZYSTANIEM METODY *BOOSTING* W ANALIZIE DANYCH SYMBOLICZNYCH

Streszczenie: Celem artykułu jest zaprezentowanie możliwości wykorzystania metody *boosting* w agregacji modeli dla danych symbolicznych z zastosowaniem metody *k*-najbliższych sąsiadów jako klasyfikatora bazowego. W artykule przedstawiono podstawowe pojęcia z zakresu analizy danych symbolicznych, metody *k*-najbliższych sąsiadów. W części empirycznej przedstawiono zastosowanie podejścia wielomodelowego dla danych symbolicznych dla kredytów konsumpcyjnych.

Słowa kluczowe: analiza danych symbolicznych, podejście wielomodelowe, *boosting.*

1. Wstęp

Ideą podejścia wielomodelowego jest łączenie – nazywane także agregacją – wyników M modeli bazowych $(D_1, \ldots, D_M)$ w jeden model zagregowany D^*, por. [Kuncheva 2004, s. 6-7; Walesiak, Gatnar 2009, s. 261; Gatnar 2008, s. 62]. Celem zastosowania podejścia wielomodelowego, zamiast wykorzystania pojedynczego modelu, jest zmniejszenie błędu predykcji. Oznacza to, że model połączony jest modelem bardziej dokładnym niż jakikolwiek z pojedynczych modeli, które wchodzą w jego skład, zob. [Gatnar 2008, s. 62]. Metoda *boosting* jest drugą obok metody *bagging* z bardziej znanych metod stosowanych w podejściu wielomodelowym. Metoda ta została zaproponowana pod nazwą *AdaBoost* przez Freunda i Schapire w 1995 r., zob. [Gatnar 2008, s. 148; Freund, Schapire 1997, s. 119].

Celem artykułu jest zaprezentowanie możliwości zastosowania i modyfikacji metody *boosting* w agregacji modeli dla danych symbolicznych z wykorzystaniem metody *k*-najbliższych sąsiadów jako klasyfikatora bazowego. W części empirycznej przedstawiono wyniki badań z wykorzystaniem zbioru kredytów konsumpcyjnych.

2. Dane symboliczne

Obiekty symboliczne mogą być opisywane przez następujące rodzaje zmiennych symbolicznych [Bock, Diday 2000, s. 2-3]:

1) zmienne w ujęciu klasycznym, tj. ilorazowe, przedziałowe, porządkowe, nominalne;

2) zmienne symboliczne, tj. zmienne:

- interwałowe, których realizacją są przedziały liczbowe rozłączne lub nierozłączne;
- wielowariantowe, gdzie realizacją zmiennej jest więcej niż jeden wariant (liczba lub kategoria);
- wielowariantowe z wagami, gdzie realizacją zmiennej oprócz wielu wariantów są dodatkowo wagi (lub prawdopodobieństwa) dla każdego z wariantów zmiennej dla danego obiektu.

Niezależnie od typu zmiennej w analizie danych symbolicznych możemy mieć do czynienia ze zmiennymi strukturalnymi [Bock, Diday 2000, s. 2-3; 33-37]. Do tego typu zmiennych zalicza się **zmienne hierarchiczne** – w których *a priori* ustalone są reguły decydujące o tym, czy dana zmienna opisuje dany obiekt czy nie; **zmienne taksonomiczne** – w których ustalone są *a priori* realizacje danej zmiennej; **zmienne logiczne** – tj. takie, dla których ustalono *a priori* reguły logiczne lub funkcyjne, które decydują o wartościach zmiennej.

W analizie danych symbolicznych wyróżnia się dwa typy obiektów symbolicznych:

- **obiekty symboliczne pierwszego rzędu** – obiekty rozumiane w sensie „klasycznym" (obiekty elementarne), np. konsument, przedsiębiorstwo, produkt, pacjent czy gospodarstwo domowe,
- **obiekty symboliczne drugiego rzędu** – obiekty utworzone w wyniku agregacji zbioru obiektów symbolicznych pierwszego rzędu, np. grupa konsumentów preferująca określony produkt, region geograficzny (jako wynik agregacji podregionów).

3. Idea metody *boosting*

Drugą, obok metody *bagging*, popularną metodą łączenia modeli bazowych jest metoda *boosting,* zob. [Gartnar 2008, s. 145-154; Polikar 2006, s. 28-32; Kuncheva 2004, s. 212-222]. Metoda ta realizuje w swej konstrukcji architekturę szeregową modeli zagregowanych (zob. rys. 1). Oznacza to, że wyniki kolejnych modeli zależą od wyników modeli poprzednich.

Rys. 1. Architektura szeregowa

Źródło: [Gatnar 2008, s. 69].

Metoda *boosting* polega na poprawianiu (inaczej wzmacnianiu) dokładności predykcji modelu zagregowanego D^* w rezultacie modyfikacji kolejnych modeli ba-

zowych $D_1, \ldots, D_M$, por. [Gatnar 2008, s. 145]. Poprawę dokładności predykcji uzyskuje się poprzez zastosowanie podwójnego systemu wag. Pierwszy dotyczy obserwacji i polega na tym, że obserwacje, które błędnie sklasyfikował *i*-ty model D_i, otrzymują wyższe wagi. Drugi system wag polega na przydzieleniu każdemu z modeli wag proporcjonalnych do błędu jego predykcji. Obserwacje są losowane do każdego z M modeli bazowych zgodnie z przypisanymi im wagami, których suma dla obiektu musi wynosić jeden, zob. [Gatnar 2008, s. 145; Kuncheva 2004, s. 216; Polikar 2006, s. 29].

Algorytm metody *boosting* zostanie omówiony na przykładzie algorytmu ***AdaBoost*** (nazwa pochodzi od ***Ada**ptive **Boost**ing*) [por. Gatnar 2008, s. 146]. Algorytm ten składa się z czterech kroków [Gatnar 2008, s. 146; Polikar 2006, s. 29-30; Kuncheva 2004, s. 216]:

1. Ustalenie liczby modeli bazowych M.

2. Ustalenie początkowych wag obserwacji ze zbioru uczącego U:

a) wagi mogą być odwrotnie proporcjonalne do liczby obiektów w zbiorze uczącym:

$$\forall_{i=1,\ldots,N} w_i^{(1)} = \frac{1}{N}, \tag{1}$$

b) wagi mogą być zależne od potencjału opisowego obiektów symbolicznych – propozycję takiego rozwiązania zawarł w swojej pracy A. Dudek, zob. [2009, s. 33-40]:

$$\forall_{i=1,\ldots,N} w_i^{(1)} = \frac{\pi(O_i)}{\pi(O_R)}, \tag{2}$$

gdzie: O_i – *i*-ty obiekt symboliczny ze zbioru uczącego,

O_E – syntetyczny obiekt symboliczny opisujący wszystkie obiekty ze zbioru uczącego,

π – potencjał obiektu symbolicznego liczony zgodnie ze wzorem [Bock, Diday 2000, s. 176]:

$$\pi(O_i) = \prod_{j=1}^{P} \mu(v_{ij}), \tag{3}$$

gdzie: $j = 1, \ldots, P$ – numer zmiennej symbolicznej,

$\mu(v_{ij})$ – długość przedziału dla zmiennych interwałowych, dla zmiennych wielowariantowych – liczba elementów (wariantów) j-tej zmiennej dla *i*-tego obiektu symbolicznego.

W części empirycznej artykułu wykorzystane zostaną zarówno wagi zależne od potencjału opisowego, jak i wagi odwrotnie proporcjonalne do liczby obiektów.

3. Wykonanie dla każdego $m = 1, \ldots, M$ następujących czynności:

a) wylosowanie ze zbioru uczącego U do próby uczącej U_m obiektów zgodnie z rozkładem ich wag,

b) zbudowanie modelu bazowego D_m na podstawie próby uczącej U_m i obliczenie błędu predykcji (jako błędu resubstytucji):

$$e(D_m) = \sum_{i=1}^{N} w_i^{(m)} I\left(\hat{D}_m(O_i) \neq y_i\right), \tag{4}$$

c) jeżeli $e(D_m) = 0$ lub $e(D_m) \geq 0{,}5$, należy przerwać działanie algorytmu,

d) w przeciwnym razie obliczana jest waga dla modelu bazowego D_m:

$$\beta_m = \frac{e(D_m)}{1 - e(D_m)}, \tag{5}$$

e) zmodyfikowanie wag obserwacji zgodnie ze wzorem:

$$w_i^{(m+1)} = \frac{w_i^{(m)} \beta_m^{I(\hat{D}_m(O_i) = y_i)}}{\sum_{k=1}^{N} w_k^{(m)} \beta_m^{I(\hat{D}_m(O_k) = y_k)}}, \tag{6}$$

f) powrót do podpunktu 3a.

4. Dokonanie predykcji modelu zagregowanego dla obserwacji O_i za pomocą modeli bazowych z wykorzystaniem ważonego głosowania:

$$\hat{D}^*(O_i) = \arg\max_j \left\{ \sum_{m=1}^{M} \ln\left(\frac{1}{\beta_m}\right) I\left(\hat{D}_m(O_i) = C_j\right) \right\}. \tag{7}$$

Jak wspomniano we wprowadzeniu, klasyfikatorem bazowym w przykładzie empirycznym jest metoda *k*-najbliższych sąsiadów dla danych symbolicznych. Algorytm tej metody można wyrazić za pomocą następujących kroków (zob. [Malerba, D'Amato, Esposito, Monopoli 2003; Malerba, Esposito, D'Amato, Appice 2004; Malerba, Esposito, D'Amato, Appice 2006; Pełka 2010]):

1. Wybór liczby sąsiadów branych pod uwagę w dalszej części algorytmu (*k*).

2. Obliczenie odległości między obiektami symbolicznymi (ze zbioru uczącego i testowego).

3. Wybór *k* obiektów ze zbioru uczącego najbliższych *i*-temu obiektowi ze zbioru testowego.

4. Obliczenie prawdopodobieństw *a posteriori* przydzielenia obiektu ze zbioru testowego do każdej z klas zbioru uczącego. Prawdopodobieństwo to obliczane jest zgodnie ze wzorem:

$$P\left(O_i \mid C_j\right) = \frac{\frac{K_j}{K} \cdot \Omega_j}{\sum_{j=1}^{J} \frac{K_j}{K} \cdot \Omega_j}, \quad \forall j = 1, \ldots, J, \tag{8}$$

gdzie: $\Omega_j = \sum_{j=1}^{J} w_i \cdot \delta\left(C_j, C_k\right)$,

$w_i = \frac{1}{d\left(O_i, O_k\right)}$ – wagi, które są odwrotnością odległości między *i*-tym obiektem ze zbioru testowego a *k*-tym sąsiadem ze zbioru uczącego,

$\delta\left(C_j, C_k\right) = 1$ – jeżeli klasa, do której należy *k*-ty sąsiad, jest taka sama jak klasa, do której przyporządkowujemy *i*-ty obiekt,

$\delta\left(C_j, C_k\right) = 0$ – jeżeli klasa, do której należy *k*-ty sąsiad, jest inna niż klasa, do której przyporządkowujemy *i*-ty obiekt,

K_j – liczba sąsiadów należących do *j*-tej klasy,

$j = 1, \ldots, J$ – numer klasy.

4. Przykład empiryczny

Bank BGŻ SA osiągnął w 2004 r. dziesiąte miejsce pod względem sumy udzielonych kredytów, a jedenaste pod względem funduszy własnych, por. [Adomski 2005, s. 6]. Jednocześnie bank ten w rankingu 50 największych banków w Polsce został uznany w 2004 r. za drugi bank, po BPH SA, w kategorii banków uniwersalnych, por. [Adomski 2005, s. 22].

Zadłużenie z tytułu kredytów konsumpcyjnych według danych na koniec 2004 r. stanowiło 69% całego bankowego zadłużenia gospodarstw domowych, zob. [Penczar i in. 2005, s. 19].

W przykładzie empirycznym wykorzystano dwa zbiory danych dotyczący kredytów konsumpcyjnych udzielonych w 2004 r. przez Bank Gospodarki Żywnościowej SA w Kłodzku. Jako metodę doboru próby do badania wybrano dobór nielosowy (wybór kwotowy), zob. [Szreder 2004, s. 53-60]. Kwotami w tym przypadku były poszczególne rodzaje kredytów konsumpcyjnych udzielone przez BGŻ SA. Oddział w Kłodzku. Zbiór danych zawiera 100 obiektów podzielonych na dwie klasy (72 decyzje o udzieleniu kredytu – klasa 1, 28 decyzji o odrzuceniu wniosku kredytowego – klasa 2). Obydwie klasy opisuje trzynaście zmiennych:

1. Średnie wpływy na rachunek bieżący – zmienna interwałowa.
2. Staż pracy kredytobiorcy – zmienna interwałowa.
3. Czas trwania kredytu w miesiącach – zmienna interwałowa.
4. Dochody kredytobiorcy – zmienna interwałowa.
5. Wnioskowana kwota kredytu – zmienna interwałowa.
6. Historia kredytowa – zmienna wielowariantowa.
7. Staż klienta w banku BGŻ SA – zmienna interwałowa.

8. Wskazanie poręczyciela – zmienna wielowariantowa.
9. Ocena poręczyciela – zmienna wielowariantowa.
10. Inne proponowane zabezpieczenia – zmienna wielowariantowa.
11. Ocena klienta w BGŻ SA – zmienna wielowariantowa.
12. Lojalność klienta wobec BGŻ SA – zmienna wielowariantowa.
13. Udzielona informacja o sytuacji kredytowej – zmienna wielowariantowa.

Zbiór danych podzielono na dwa podzbiory – zbiór uczący stanowiło 75 obiektów, a zbiór testowy 25 obiektów.

Wyniki otrzymane przy zastosowaniu podejścia wielomodelowego *boosting* z zastosowaniem wag odwrotnie proporcjonalnych do liczby obiektów zestawiono w tab. 1. Wyniki otrzymane przy zastosowaniu wag zależnych od potencjału opisowego obiektów symbolicznych zestawiono w tab. 2.

Tabela 1. Wyniki obliczeń dla wag odwrotnie proporcjonalnych do liczby obiektów.

		Klasyfikacja rzeczywista	
		klasa 1	klasa 2
Decyzja KNN	klasa 1	12	2
	klasa 2	1	10

Źródło: obliczenia własne w programie R.

Najmniejszy błąd klasyfikacji (12%) otrzymano dla 38 modeli bazowych. W tym przypadku błędnie sklasyfikowano 3 spośród 25 obiektów w zbiorze testowym.

Tabela 2. Wyniki obliczeń dla wag zależnych od potencjału opisowego obiektów

		Klasyfikacja rzeczywista	
		klasa 1	klasa 2
Decyzja KNN	klasa 1	12	0
	klasa 2	1	12

Źródło: obliczenia własne w programie R.

Najmniejszy błąd klasyfikacji (4%) uzyskano dla 26 modeli bazowych. W tym przypadku jedynie jeden obiekt (o pozytywnej decyzji kredytowej) został sklasyfikowany jako obiekt, który nie powinien otrzymać kredytu.

5. Podsumowanie

Metoda *boosting* może znaleźć zastosowanie w klasyfikacji różnych zbiorów danych symbolicznych. Podejście wielomodelowe analizy danych symbolicznych, podobnie jak podejście wielomodelowe dla danych klasycznych, pozwala osiągnąć mniejszy błąd klasyfikacji niż zastosowanie pojedynczego modelu.

Na potrzeby badań empirycznych opracowano w programie R skrypt realizujący algorytm metody *boosting* z zastosowaniem metody *k*-najbliższych sąsiadów jako klasyfikatora bazowego.

W metodzie *boosting* można zastosować dwa sposoby ważenia obiektów symbolicznych – zależny od liczby obiektów oraz zależny od potencjału opisowego. W obydwu przypadkach nieco lepsze wyniki uzyskano, stosując ważenie obiektów symbolicznych zależne od ich potencjału opisowego.

Etapem dalszych prac będzie porównanie wyników otrzymywanych przy różnych sposobach ważenia obiektów symbolicznych oraz porównanie metody *boosting* z innymi metodami podejścia wielomodelowego (np. *bagging*).

Literatura

Adomski G., *Każdemu według potrzeb*, „Bank" 2005, nr 4(150).

Bock H.-H., Diday E (red.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin 2000.

Dudek A., *Tworzenie zagregowanych modeli dyskryminacyjnych dla obiektów symbolicznych – wybrane problemy,* [w:] J. Pociecha, *Współczesne problemy statystyki, ekonometrii i matematyki stosowanej*, Studia i Prace Uniwersytetu Ekonomicznego w Krakowie, Kraków 2009.

Freund Y., Schapire R.E., *A decision-theoretic generalization of on-line learning and an application to boosting*, „Journal of Computer and System Sciences" 1997, vol. 55, no 1.

Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2008.

Kuncheva L.I., *Combining Pattern Classifiers. Methods and Algorithms*, Wiley, New Jersey 2004.

Malerba D., D'Amato C., Esposito F., Monopoli M., *Extending the K-Nearest Neighbour Classification Algorithm to Symbolic Objects,* Atti del Convegno Intermedio della Società Italiana di Statistica „Analisi Statistica Multivariata per le scienze economico-sociali, le scienze naturali e la tecnologia", Napoli 2003.

Malerba D., Esposito F., D'Amato C., Appice A., *K-nearest Neighbor Classification for Symbolic Objects*, [w:] *Symbolic and Spatial Data Analysis: Mining Complex Data Structures*, P. Brito, M. Noirhomme-Fraiture (red.), University of Pisa, Pisa 2004.

Malerba D., Esposito F., D'Amato C., Appice A., *Classification of symbolic objects: A lazy learning approach*, „Intelligent Data Analysis" 2006, vol. 10, no 4.

Polikar R., *Ensemble based systems in decision making*, „IEEE Circuits and Systems Magazine" 2006, vol. 6, no 3.

Pełka M., *K-nearest neighbour classification for symbolic data*, „Acta Universitatis Lodziensis. Folia Oeconomica" 2010, nr 235.

Penczar M., Lepczyński B., Gostomski E. (red.), *Zadłużenie konsumentów w bankach i instytucjach finansowych*, Instytut Badań nad Gospodarką Rynkową, Gdańsk.

Szreder M., *Metody i techniki sondażowych badań opinii*, PWE, Warszawa 2004.

Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

ENSEMBLE LEARNING WITH THE APPLICATION OF *BOOSTING* IN SYMBOLIC DATA ANALYSIS

Summary: The aim of this paper is to present the application of boosting method in ensemble learning for symbolic data with the application of *k*-nearest neighbour method as the base classifier. The article presents basic terms of symbolic data, k-nearest neighbour classification rule for symbolic data. In the empirical part the results of application of ensemble learning for symbolic data applied for credit data set are presented.

Keywords: symbolic data analysis, ensemble learning, *boosting*.