PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176

Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Marcin Pełka

Uniwersytet Ekonomiczny we Wrocławiu

PODEJŚCIE WIELOMODELOWE W ANALIZIE DANYCH SYMBOLICZNYCH – METODA BAGGING

Streszczenie: W artykule przedstawiono podstawowe pojęcia związane z metodą bagging oraz metodą *k*-najbliższych sąsiadów dla danych symbolicznych. Zaprezentowano w nim także możliwość zastosowania podejścia wielomodelowego bagging w metodzie *k*-najbliższych sąsiadów dla danych symbolicznych. W części empirycznej przedstawiono zastosowanie podejścia wielomodelowego dla danych symbolicznych w przypadku przykładowych zbiorów danych wygenerowanych za pomocą funkcji `cluster.Gen` z pakietu `clusterSim` w programie **R**.

Słowa kluczowe: analiza danych symbolicznych, podejście wielomodelowe, bagging.

1. Wstęp

Podejście wielomodelowe polega, ogólnie rzecz ujmując, na łączeniu (agregacji) M modeli bazowych $D_1, ..., D_M$ w jeden model zagregowany D^* (por. [Kuncheva 2004, s. 6-7]). Celem zastosowania podejścia wielomodelowego w miejsce pojedynczego modelu jest zmniejszenie błędu predykcji. Model zagregowany jest bardziej dokładny niż jakikolwiek z pojedynczych modeli, które go tworzą (zob. np. [Gatnar 2008, s. 62]). Jedną z bardziej znanych metod agregacji modeli bazowych jest metoda agregacji bootstrapowej, zaproponowana przez Breimana w 1996 r., znana jako bagging (zob. np. [Gatnar 2008, s. 140; Breiman 1996, s. 123]).

Celem artykułu jest zaprezentowanie możliwości zastosowania metody bagging w agregacji modeli dyskryminacyjnych dla danych symbolicznych na przykładzie metody *k*-najbliższych sąsiadów dla obiektów symbolicznych. Oprócz tego przetestowana zostanie skuteczność metody bagging wykorzystującej metodę *k*-najbliższych sąsiadów w przypadku, gdy mamy do czynienia ze zmiennymi zakłócającymi (*noisy variables*) i obserwacjami odstającymi (*outliers*).

W części empirycznej przedstawiono wyniki symulacji z wykorzystaniem sztucznych zbiorów danych interwałowych wygenerowanych za pomocą pakietu `clusterSim` w programie **R**.

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 − realizacją są przedziały liczbowe rozłączne lub nierozłączne;
– wielowariantowe − realizacją zmiennej jest więcej niż jeden wariant (liczba lub kategoria);
– wielowariantowe z wagami − 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** (ustalone są *a priori* reguły decydujące o tym, czy dana zmienna opisuje dany obiekt czy nie), **taksonomiczne** (ustalone są *a priori* realizacje danej zmiennej), **logiczne** (ustalono dla nich *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** – rozumiane w sensie „klasycznym" (obiekty elementarne), np. konsument, przedsiębiorstwo, produkt, pacjent czy gospodarstwo domowe,
– **obiekty symboliczne drugiego rzędu** – utworzone w wyniku agregacji zbioru obiektów symbolicznych pierwszego rzędu, np. grupa konsumentów preferująca określony produkt, region geograficzny (jako wynik agregacji podregionów).

3. Algorytm metody *k*-najbliższych sąsiadów dla danych symbolicznych

Metoda *k*-najbliższych sąsiadów w swym klasycznym kształcie została zaproponowana przez E. Fixa i J.L. Hodgesa [1951]. Adaptację klasycznej metody *k*-najbliższych sąsiadów dla danych symbolicznych zawarł w swej pracy zespół Malerby [2003]. Algorytm tej metody dla danych symbolicznych można zapisać w następujący sposób:

1. Dla każdego obiektu $\mathbf{O}$ ze zbioru testowego obliczana jest odległość tego obiektu od obiektów ze zbioru uczącego. W pomiarze odległości w analizie danych symbolicznych wykorzystywanych jest wiele różnych miar odległości, m.in. miary Ichino-Yaguchiego, miary de Carvalho, miara wzajemnego sąsiedztwa (Gowdy--Krishny-Didaya) czy wreszcie uogólniona miara Minkowskiego, pozwalająca uwzględnić zmienne wielowariantowe z wagami (por. np. [Bock, Diday 2000, s. 182]).

2. Znajdowanych jest *k* obiektów ze zbioru uczącego najbliższych obiektowi $\mathbf{O}$. Malerba in. [2006] sugerują, by liczbę *k* sąsiadów wyznaczyć z przedziału $\left[1, \sqrt{n}\right]$, gdzie n − liczba obiektów w zbiorze uczącym (zob. [Malerba i in. 2006, s. 311]).

3. Obliczane jest prawdopodobieństwo przydzielenia obiektu $\mathbf{O}$ do poszczególnych klas zbioru uczącego.

Prawdopodobieństwo przydzielenia obiektu $\mathbf{O}$ do klasy jest obliczane na trzy sposoby (por. [Malerba i in. 2004, s. 25, 2006, s. 310-311]):

a) jeżeli odległość między obiektem $\mathbf{O}$ a wszystkimi jego sąsiadami jest równa zeru, i dodatkowo wszyscy sąsiedzi należą do tej samej klasy, wówczas prawdopodobieństwo przydzielenia obiektu $\mathbf{O}$ do tej klasy wynosi 1;

b) jeżeli odległość między obiektem $\mathbf{O}$ a wszystkimi jego sąsiadami jest równa zeru, a sąsiedzi ci należą do różnych klas, prawdopodobieństwo przydzielenia obiektu $\mathbf{O}$ do poszczególnych klas wyraża się wzorem:

$$P(\mathbf{O} \mid C_j) = \frac{K_j}{K}, \forall j = 1, \ldots, J, \tag{1}$$

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

K – liczba sąsiadów,

$j = 1, ..., J$ – liczba klas w zbiorze uczącym;

c) jeżeli odległość między obiektem $\mathbf{O}$ a wszystkimi jego sąsiadami jest różna od zera, wówczas prawdopodobieństwo przydzielenia obiektu $\mathbf{O}$ do poszczególnych klas wyraża się wzorem:

$$P(\mathbf{O} \mid C_j) = \frac{\frac{K_j}{K} \cdot \Omega_j}{\sum_{j=1}^{J} \frac{K_j}{K} \cdot \Omega_j}, \forall j = 1, ..., J, \tag{2}$$

gdzie:

$$\Omega_j = w_i \cdot \delta(C_j, C_k), \tag{3}$$

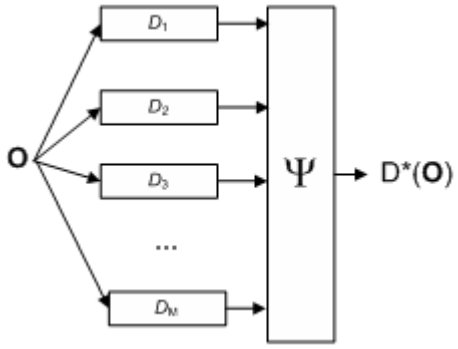
$$w_i = \frac{1}{d(\mathbf{O}, \mathbf{X}_i)} - \text{wagi}, \tag{4}$$

$\delta(C_j, C_k) = 1$, jeżeli klasa, do której należy *k*-ty sąsiad, jest taka sama, jak klasa, do której przyporządkowujemy obiekt $\mathbf{O}$; $\delta(C_j, C_k) = 0$, jeżeli klasa, do której należy *k*-ty sąsiad jest różna od klasy, do której przyporządkowujemy obiekt $\mathbf{O}$, pozostałe oznaczenia jak we wzorze 1.

W metodzie *k*-najbliższych sąsiadów dla danych symbolicznych „ważniejszy jest głos" sąsiadów bliższych obiektowi klasyfikowanemu niż pozostałych sąsiadów.

4. Idea metody bagging

Jedną z bardziej znanych metod agregacji modeli bazowych jest metoda agregacji bootstrapowej, znana jako bagging (por. [Gatnar 2008, s. 140; Breinman 1996, s. 123; Kuncheva 2004, s. 203]). Metoda bagging realizuje w swej konstrukcji architekturę równoległą modeli zagregowanych (zob. rys. 1).



Rys. 1. Architektura równoległa łączenia modeli bazowych

Źródło: [Gatnar 2008, s. 68].

Architektura równoległa łączenia modeli bazowych zakłada niezależne działanie każdego z M modeli bazowych za pomocą funkcji

$$\hat{D}^*(\mathbf{O}) = \Psi\left(\hat{D}_1^*(\mathbf{O}), ..., \hat{D}_M^*(\mathbf{O})\right),$$

dając w rezultacie model zagregowany D^* (zob. [Gatnar 2008, s. 63, 68]).

Metoda bagging polega na zbudowaniu M modeli bazowych na podstawie prób uczących $U_1, ..., U_M$ losowanych ze zwracaniem ze zbioru uczącego *U*. Próby te nazywane są próbami bootstrapowymi [Gatnar 2008, s. 140; Polikar 2007, s. 60,

2006, s. 29]. Zwykle około 37% zbioru uczącego nie trafia do żadnej z prób uczących. Tworzą one dodatkowy zbiór, tzw. OOB (*Out-Of-Bag*), który jest często wykorzystywany jako dodatkowy zbiór testowy [Gatnar 2008, s. 140].

Algorytm metody bagging można przedstawić za pomocą następujących kroków [Gatnar 2008, s. 140; Polikar 2007, s. 60-61, 2006, s. 29; Kuncheva 2004, s. 204]:

1. Ustalenie liczby modeli bazowych M.

2. Następnie dla każdego z modeli bazowych $(m = 1, ..., M)$:

a) wylosowanie próby bootstrapowej ze zbioru uczącego,

b) zbudowanie modelu bazowego na podstawie próby bootstrapowej (metoda *k*-najbliższych sąsiadów dla danych symbolicznych).

3. Dokonanie predykcji modelu zagregowanego za pomocą modeli bazowych − stosując metodę głosowania większościowego w przypadku dyskryminacji lub uśrednia wyniki w przypadku regresji.

Metoda głosowania większościowego (*majority voting*) polega na przydzieleniu obserwacji do klasy, którą wskazała największa liczba modeli bazowych [Gatnar 2008, s. 114]:

$$\hat{D}^{*}(\mathbf{O}) = \arg\max_{j} \sum_{m=1}^{M} I\left(\hat{D}_{m}(\mathbf{O}) = C_{j}\right).$$

5. Wyniki badań symulacyjnych

Zbiór pierwszy (model I) to 1000 obiektów podzielonych na pięć niezbyt dobrze separowalnych klas, opisywanych przez dwie zmienne symboliczne interwałowe. Zbiór uczący stanowiło 800 obiektów, a zbiór testowy 200 obiektów. Zmienne w tym zbiorze są losowane z dwuwymiarowego rozkładu normalnego o średnich (5, 5), (–3, 3), (3, –3), (0, 0), (–5, –5) oraz macierzy kowariancji $\sum\left(\sigma_{jj} = 1, \sigma_{jl} = 0{,}9\right)$. Zbiór ten nie zawiera zmiennych zakłócających czy obserwacji odstających.

Zbiór drugi (model II) to 200 obiektów podzielonych na dwie klasy o wydłużonym kształcie, opisywane prze dwie zmienne symboliczne interwałowe. Zmienne w tym zbiorze są losowane z dwuwymiarowego rozkładu normalnego o średnich (0, 0), (1, 5) i macierzy kowariancji $\sum\left(\sigma_{jj} = 1, \sigma_{jl} = -0{,}9\right)$. Zbiór nie zawiera zmiennych zakłócających i obserwacji odstających. Elementy zbioru podzielono na zbiór uczący (120 obiektów) i zbiór testowy (80 obiektów).

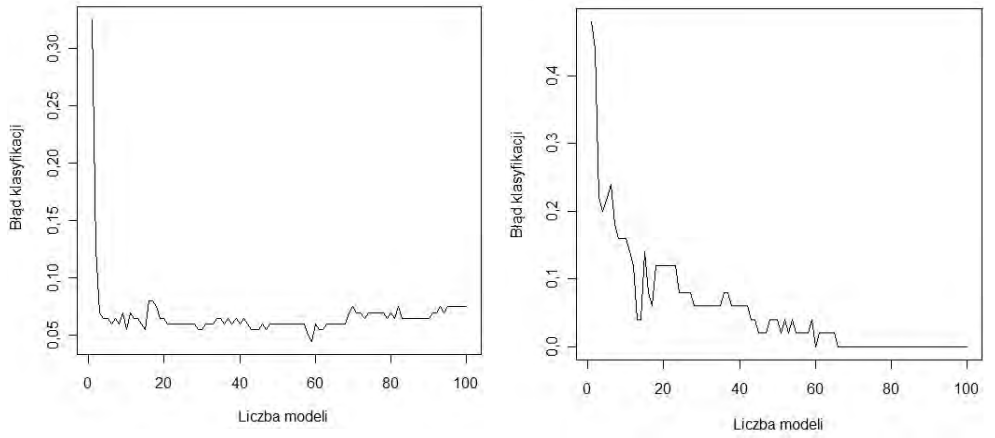
Zbiór trzeci (model III) to 800 obiektów podzielonych na trzy klasy o wydłużonym kształcie, opisywane przez trzy zmienne symboliczne interwałowe. Zbiór uczący zawiera 640 obiektów, a zbiór testowy − 160. Zmienne opisujące te obiekty są losowane z dwuwymiarowego rozkładu normalnego o średnich (0, 0), (1,5, 7),

(3, 14) oraz macierzy kowariancji $\sum\left(\sigma_{jj}=1,\sigma_{jl}=-0{,}9\right)$. Zbiór zawiera jedną zmienną zakłócającą, natomiast nie zawiera obserwacji odstających.

Zbiór czwarty (model IV) to 660 obiektów podzielonych na dwie klasy o wydłużonym kształcie, opisywane przez dwie zmienne symboliczne interwałowe. Zbiór uczący stanowiło 528 obserwacji, a zbiór testowy 132 obserwacje. Zmienne opisujące te zbiory są losowane z dwuwymiarowego rozkładu normalnego o średnich (0, 0), (1, 5) oraz macierzy kowariancji $\sum\left(\sigma_{jj}=1,\sigma_{jl}=-0{,}9\right)$. Zbiór zawiera 10% obserwacji odstających, ale nie zawiera zmiennych zakłócających.

Modele I, II, III oraz IV wygenerowano z wykorzystaniem funkcji `cluster.Gen` z pakietu `clusterSim` w programie **R**.

Dla każdego ze zbiorów zastosowano metodę bagging dla metody *k*-najbliższych sąsiadów dla danych symbolicznych, przyjmując liczbę modeli bazowych od jeden do stu, a liczbę *k* sąsiadów na poziomie trzy razy pierwiastek z liczby obserwacji w zbiorze uczącym. Wyniki symulacji dla modelu I i II przedstawia rys. 2.



a) model I b) model II

Rys. 2. Wyniki symulacji dla modelu I i II

Źródło: obliczenia własne w programie **R**.

Najmniejszy błąd klasyfikacji, równy 4,5% w przypadku modelu I, otrzymano dla 59 modeli bazowych. W przypadku modelu II najmniejszy błąd klasyfikacji, równy zero, otrzymano dla 60 modeli bazowych.

Dla modelu III, który zawierał jedną zmienną zakłócającą, błąd klasyfikacji wyniósł aż 62,5% (dla 14 modeli bazowych). W przypadku modelu IV, którego zbiór zawierał obiekty odstające, błąd klasyfikacji wyniósł 10% (klasyfikację obiektów ze zbioru uczącego dla 50 modeli bazowych zawarto w tab. 1).

Tabela 1. Klasyfikacja obiektów dla modelu **IV**

Faktyczna przynależność do klas	Predykcja przynależności do klas	
	klasa 1	klasa 2
Klasa 1	56	0
Klasa 2	0	64
Obserwacje odstające	7	5

Źródło: obliczenia własne w programie **R**.

W przypadku modelu IV obiekty z klasy pierwszej i drugiej są klasyfikowane poprawnie. Błędna klasyfikacja obiektów pojawia się natomiast w przypadku obserwacji odstających.

6. Podsumowanie

Metoda bagging, w której klasyfikatorem bazowym jest metoda *k*-najbliższych sąsiadów, może znaleźć zastosowanie w klasyfikacji różnych zbiorów danych symbolicznych. Podejście wielomodelowe połączone z KNN dla danych symbolicznych pozwala na precyzyjne zidentyfikowanie klas słabo separowalnych oraz klas o wydłużonym kształcie.

W przypadku podejścia wielomodelowego w klasyfikacji obiektów symbolicznych z zastosowaniem metody bagging dla metody *k*-najbliższych sąsiadów najlepsze wyniki otrzymano dla zbiorów danych niezawierających zmiennych zakłócających. Gdy mieliśmy do czynienia ze zmiennymi zakłócającymi, błąd klasyfikacji wyniósł aż 62,5%, co może wynikać z faktu, że zmienne te wpływają znacząco na pomiar odległości, a to z kolei przenosi się na prawdopodobieństwa *a posteriori* przydzielenia obiektów do klas.

Etapem dalszych prac będzie próba zaadaptowania innych metod podejścia wielomodelowego dla danych symbolicznych.

Literatura

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

Breiman L., *Bagging predictors*, „Machine Learning" 1996, vol. 24, s. 123-140.

Fix E., Hodges J.L., *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties.* Report 4. Project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas 1951.

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 Statis-

tica „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:] P. Brito, M. Noirhomme-Fraiture (red.), *Symbolic and spatial data analysis: mining complex data structures*, University of Pisa, Pisa 2004, s. 19-30.

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, s. 301-324.

Polikar R., *Ensemble based systems in decision making*, „IEEE Circuits and Systems Magazine" 2006, vol. 6, no. 3, s. 21-45.

Polikar R., *Bootstrap inspired techniques in computational intelligence*, „IEEE Signal Processing Magazine" 2007, vol. 24, no. 4, s. 56-72.

ENSEMBLE LEARNING FOR SYMBOLIC DATA WITH APPLICATION OF *BAGGING*

Summary: The paper presents the most important basic terms of *bagging* and *k*-nearest neighbour for symbolic data. The article also presents an application of *bagging* ensemble for *k*-nearest neighbour for symbolic data.In the empirical part the application of ensemble learning for symbolic data is presented for some data sets generated by function `cluster.Gen` from `clusterSim` package of **R** software.