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

Joanna Trzęsiok
Akademia Ekonomiczna w Katowicach

PORÓWNANIE METODY GRADIENTOWEJ WYKORZYSTUJĄCEJ TRANSFORMACJĘ LOGISTYCZNĄ Z WYBRANYMI METODAMI DYSKRYMINACJI

1. Wstęp

Agregacyjna metoda gradientowa została zaproponowana przez Friedmana w roku 1999 (zob. [Friedman 2001]). Jest ona przykładem nieparametrycznej metody wykorzystującej sekwencyjne łączenie modeli składowych, tworzonych i systematycznie poprawianych w kolejnych krokach algorytmu (*boosting method*) (por. [Freund, Schapire 1997; Hastie, Tibshirani, Friedman 2001]). W efekcie wykorzystania tej procedury poprawiona zostaje również jakość modelu końcowego.

Metoda gradientowa pierwotnie znalazła zastosowanie w analizie regresji. Jej algorytm wykorzystywany w regresji nosi nazwę MART – agregacyjnej metody drzew regresyjnych (*multiple additive regression trees*). Jednak Friedman przedstawił również modyfikację algorytmu MART, w której zastosował metodę gradientową, opartą na transformacji logistycznej, do rozwiązania problemu dyskryminacji K klas.

Nieparametryczna metoda MART, stosowana w regresji, charakteryzuje się bardzo dobrymi własnościami statystycznymi. Jest odporna na występowanie w zbiorze uczącym szumu oraz wartości oddalonych (zob. [Trzęsiok 2006]). Głównym atutem modelu skonstruowanego poprzez algorytm MART jest dokładność predykcji. W porównaniu z innymi nieklasycznymi metodami regresji modele otrzymane za pomocą tej metody charakteryzują się jednymi z najwyższych wartości miar jakości (zob. [Meyer, Leisch, Hornik 2002]).

Powstaje więc pytanie, czy metoda gradientowa wykorzystująca transformację logistyczną, jako modyfikacja algorytmu MART, jest dobrym narzędziem do rozwiązania problemu dyskryminacji K klas. Interesuje nas, czy na tle innych metod

dyskryminacji omawiana metoda gradientowa charakteryzuje się tak dobrą dokładnością uzyskiwanych modeli jak metoda MART w porównaniu z innymi metodami regresji.

2. Metoda gradientowa wykorzystująca transformację logistyczną

2.1. Sekwencyjne łączenie modeli składowych

Przedstawiona w tym artykule metoda gradientowa wykorzystująca transformację logistyczną jest przykładem sekwencyjnego łączenia modeli składowych, gdzie model końcowy wyznaczony zostaje iteracyjnie w trakcie wykonywania algorytmu.

W pierwszym kroku tego algorytmu, na podstawie zbioru uczącego, skonstruowana zostaje funkcja dyskryminacyjna $F_1(\mathbf{X})$. Za pomocą pewnej funkcji straty badamy jakość klasyfikacji wyznaczonego modelu dyskryminacji, a następnie zbiór uczący U zostaje odpowiednio przekształcony. Dla przetransformowanych obserwacji ze zbioru U_1 ponownie szukamy reguły dyskryminującej $F_2(\mathbf{X})$, po czym badamy jakość tej klasyfikacji i po raz drugi przekształcamy obserwacje w zbiorze uczącym. Procedurę powtarzamy aż do momentu, kiedy jakość modelu otrzymanego w kolejnym kroku nie różni się istotnie od jakości modelu z kroku poprzedniego. Z ciągu uzyskanych reguł $F_m(\mathbf{X})$ tworzymy zagregowany model końcowy.

2.2. Postać modelu w metodzie gradientowej

Zagregowany model dyskryminacyjny w metodzie gradientowej można zapisać równaniem:

$$F(\mathbf{X}) = \sum_{m=0}^{M} F_m(\mathbf{X}) = \sum_{m=0}^{M} T(\mathbf{X}, \Theta_m) = \sum_{m=0}^{M} \sum_{j=1}^{J_m} \gamma_{jm} \cdot I(\mathbf{X} \in R_{jm}). \tag{1}$$

W modelu tym funkcje składowe mają postać **drzew klasyfikacyjnych**:

$$F_m(\mathbf{X}) = T(\mathbf{X}, \Theta_m) = \sum_{j=1}^{J_m} \gamma_{jm} \cdot I(\mathbf{X} \in R_{jm}) \tag{2}$$

z parametrami $\Theta_m = \{R_{jm}, \gamma_{jm}\}_{j=1,\ldots,J_m}$, gdzie R_{jm} są rozłącznymi obszarami otrzymanymi poprzez podział przestrzeni zmiennych objaśniających, natomiast γ_{jm} to numery klas przypisane obserwacjom z danego obszaru (zob. [Gatnar 2000]).

Estymatory parametrów Θ_m w modelu zagregowanym (1) otrzymujemy poprzez minimalizację funkcji straty w zbiorze uczącym:

$$\hat{\Theta}_m = \underset{(\Theta_m)_{m=0,\ldots,M}}{\arg\min} \sum_{i=1}^{N} L\left(y_i, \sum_{m=0}^{M} T(\mathbf{x}_i, \Theta_m) \right). \tag{3}$$

Zagadnienie minimalizacji (3) jest złożone obliczeniowo, dlatego do jego rozwiązania stosuje się często **metodę największego spadku** (*steepest descent algorithm*) (zob. [Hastie, Tibshirani, Friedman 2001]).

W metodzie tej, aby uzyskać minimum funkcji straty, czyli inaczej największą redukcję wartości funkcji L, będziemy przesuwać się w kierunku przeciwnym do jej gradientu:

$$-\rho_m \cdot \mathbf{g}_m = -\rho_m \cdot \left[g_m(\mathbf{x}_i)\right]_{i=1,\ldots,N} . \tag{4}$$

Parametr $\rho_m \in \mathbf{R}$ nazywany jest długością kroku, zaś $\mathbf{g}_m \in \mathbf{R}^N$ jest gradientem z funkcji straty L o składowych:

$$g_m(\mathbf{x}_i) = \left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x}_i)=f_{m-1}(\mathbf{x}_i)} . \tag{5}$$

Metoda gradientowa stosowana zarówno w dyskryminacji, jak i regresji zawdzięcza swą nazwę właśnie metodzie największego spadku wykorzystującej gradient funkcji L.

2.3. Postać funkcji straty w metodzie gradientowej wykorzystującej transformację logistyczną

Rozpatrywana w tym artykule metoda gradientowa dotyczy przypadku dyskryminacji K klas. Załóżmy więc, że:

$$y_k = \begin{cases} 1, & \text{gdy } \mathbf{x}_k \text{ znajduje się w klasie } k \\ 0, & \text{w przeciwnym wypadku} \end{cases} \tag{6}$$

Oznaczmy przez $p_k(\mathbf{x})$ prawdopodobieństwo tego, że obiekt $\mathbf{x}$ należy do klasy k:

$$p_k(\mathbf{x}) = \Pr(y_k = 1 \,|\, \mathbf{x}) . \tag{7}$$

Przyjmujemy, że reguła dyskryminacyjna $F_k(\mathbf{X})$ dla klasy o numerze k oparta jest na **transformacji logistycznej** i można zapisać ją w postaci:

$$F_k(\mathbf{X}) = \ln p_k(\mathbf{X}) - \frac{1}{K}\sum_{l=1}^{K} \ln p_l(\mathbf{X}) . \tag{8}$$

Po przekształceniu formuły (8), określającej funkcję $F_k(\mathbf{X})$, otrzymujemy równanie wyrażające prawdopodobieństwo $p_k(\mathbf{X})$:

$$p_k(\mathbf{X}) = \frac{\exp(F_k(\mathbf{X}))}{\sum_{l=1}^{K} \exp(F_l(\mathbf{X}))} . \tag{9}$$

W rozpatrywanym zagadnieniu dyskryminacji K klas wykorzystamy **logistyczną funkcję straty** L:

$$L((y_k, F_k(\mathbf{X}))_{k=1,\ldots,K}) = -\sum_{k=1}^{K} y_k \ln p_k(\mathbf{X}) , \tag{10}$$

gdzie y_k oraz prawdopodobieństwo $p_k(\mathbf{X})$ opisane są wzorami (6), (9).

Wyznaczenie modelu dyskryminacyjnego (1) sprowadza się do wyznaczenia parametrów tego modelu poprzez minimalizację funkcję straty za pomocą strategii wspinaczki. W tym celu konieczne jest wyznaczenie składowych gradientu funkcji L:

$$\left[\frac{\partial L((y_{il}, F_l(\mathbf{x}_i))_{l=1,\ldots,K})}{\partial F_k(\mathbf{x}_i)}\right]_{(F_l(\mathbf{x})=f_{l,m-1}(\mathbf{x}))_{l=1,\ldots,K}} \tag{11}$$

Po podstawieniu do równania (10) prawdopodobieństwa $p_k(\mathbf{X})$ danego wzorem (9) i obliczeniu gradientu funkcji L otrzymujemy następujące składowe gradientu funkcji straty:

$$-\left[\frac{\partial L((y_{il}, F_l(\mathbf{x}_i))_{l=1,\ldots,K})}{\partial F_k(\mathbf{x}_i)}\right]_{(F_l(\mathbf{x})=f_{l,m-1}(\mathbf{x}))_{l=1,\ldots,K}} = y_{ik} - p_{k,m-1}(\mathbf{x}_i), \tag{12}$$

gdzie $p_{k,m-1}(\mathbf{X})$ jest pochodną funkcji $F_{k,m-1}(\mathbf{X})$.

W algorytmie metody gradientowej składowe zapisane wzorem (12) nazywać będziemy **pseudoresztami**.

2.4. Algorytm metody gradientowej wykorzystującej transformację logistyczną

Algorytm metody gradientowej, oparty na logistycznej funkcji straty i wykorzystujący metodę największego spadku, można zapisać w następujących krokach [Friedman 2001]:

1. Przyjmujemy początkowe funkcje dyskryminujące:
$$F_{k0}(\mathbf{X}) = 0, \text{ dla } k = 1,\ldots,K. \tag{13}$$
2. Dla $m = 1,\ldots,M$ wykonujemy kroki:

   A. Zgodnie z wzorem (9) obliczamy prawdopodobieństwa $p_k(\mathbf{X})$, dla $k = 1,\ldots,K$.

   B. Dla $k = 1,\ldots,K$:

   - Zgodnie ze wzorem (12) obliczamy pseudo-reszty:
   $$\tilde{y}_{ik} = y_{ik} - p_k(\mathbf{x}_i), \text{ dla } i = 1,\ldots,N. \tag{14}$$
   - Szukamy obszarów R_{jkm} (dla $j = 1,\ldots,J$), posługując się przekształconym zbiorem uczącym:
   $$\{(\mathbf{x}_i, \tilde{y}_{ik}) : i = 1,\ldots,N\}. \tag{15}$$
   - Na podstawie algorytmu Newtona-Raphsona wyznaczamy parametry modelu ze wzoru:
   $$\gamma_{jkm} = \frac{K-1}{K} \cdot \frac{\sum_{\mathbf{x}_i \in R_{jkm}} \tilde{y}_{ik}}{\sum_{\mathbf{x}_i \in R_{jkm}} |\tilde{y}_{ik}|(1-|\tilde{y}_{ik}|)}, \text{ dla } j = 1,\ldots,J. \tag{16}$$

- Przyjmujemy:

$$F_{km}(\mathbf{X}) = F_{k,m-1}(\mathbf{X}) + \sum_{j=1}^{J} \gamma_{jkm} \cdot I(\mathbf{X} \in R_{jkm}) . \quad (17)$$

3. Końcowe reguły dyskryminujące mają postać:

$$F_k(\mathbf{X}) = F_{kM}(\mathbf{X}), \text{ dla } k = 1,\ldots,K . \quad (18)$$

W pierwszym kroku tego algorytmu ustalamy początkowe reguły dyskryminujące (13), a następnie wyznaczamy prawdopodobieństwa $p_k(\mathbf{X})$ zgodnie ze wzorem (9). W kolejnym etapie obliczamy pseudoreszty $\tilde{y}_{ik}$, wyznaczające największy spadek funkcji straty, do których dopasowujemy drzewo dyskryminacyjne. Parametry γ_{jkm} modelu należy wyznaczyć poprzez minimalizację funkcji straty w zbiorze uczącym. Okazuje się jednak, że jest to zagadnienie złożone obliczeniowo, dlatego do jego rozwiązania wykorzystujemy algorytm Newtona-Raphsona (zob. [Friedman 2001]), za pomocą którego otrzymujemy oszacowania szukanych parametrów w postaci zapisanej wzorem (16). W kolejnym etapie algorytmu poprawiamy modele składowe i powtarzamy procedurę. Modele końcowe (18) uzyskujemy w ostatnim kroku algorytmu.

3. Porównanie metody gradientowej wykorzystującej transformację logistyczną z innymi metodami dyskryminacji

Nieparametryczna metoda MART wykorzystywana w regresji charakteryzuje się bardzo dobrymi własnościami statystycznymi: generuje modele dokładne (o niskich błędach predykcji) oraz jest odporna na występowanie w zbiorze uczącym szumu i wartości oddalonych (zob. [Trzęsiok 2006]). Zachodzi pytanie, czy również metoda gradientowa wykorzystująca transformację logistyczną, która jest modyfikacją metody MART, ma tak dobre własności statystyczne. W tym celu przeprowadzona została analiza porównawcza omawianej metody gradientowej z innymi metodami dyskryminacji. W przeprowadzonej analizie zastosowano:

- BOOST_LN – agregacyjną metodę gradientową wykorzystującą transformację logistyczną,
- LDA – liniową analizę dyskryminacyjną,
- KNN – metodę *k*-najbliższych sąsiadów,
- NNET – sieć neuronową,
- RFOREST – zagregowane drzewa klasyfikacyjne Breimana,
- SVM – metodę wektorów nośnych.

Symbole, którymi oznaczone zostały poszczególne metody, zostały zaczerpnięte z nazw funkcji realizujących te metody w pakiecie statystycznym „R". Pakiet ten był pomocny przy wyznaczeniu większości modeli dyskryminacyjnych oraz w przeprowa-

dzeniu wszystkich wymaganych obliczeń. Jedynie do wygenerowania zagregowanego modelu opartego na metodzie gradientowej wykorzystano program „TreeNet".

Analizę przeprowadzono na sześciu zbiorach danych standardowo wykorzystywanych do porównywania własności różnych metod dyskryminacji. Podstawowe charakterystyki tych zbiorów przedstawione zostały w tab. 1.

Pierwsze cztery zbiory są zbiorami danych rzeczywistych. Natomiast dwa pozostałe – „Circle" i „Spirals" – to sztuczne zbiory danych, które wygenerować można za pomocą biblioteki `mlbench` pakietu „R".

Tabela 1. Charakterystyki zbiorów danych wykorzystywanych w analizie

Nazwa	Liczebność	Liczba zmiennych	Liczba klas
P.I.Diabetes	768	9	2
Sonar	208	61	2
Satellite	6435	37	6
Glass	214	10	7
Circle	300	2	2
Spirals	400	2	2

Źródło: opracowanie własne.

Procedura badawcza obejmowała następujące etapy:

1. Każdy ze zbiorów danych został losowo podzielony na część uczącą i testową w proporcji 2:1.

2. Ponieważ większość porównywanych metod wymaga ustalenia parametrów na odpowiednim poziomie, symulacyjnie, na zbiorach uczących skonstruowanych zostało wiele modeli dyskryminacji. Do analizy porównawczej wybierany był ten model, który dawał najmniejszy błąd klasyfikacji liczony metodą sprawdzania krzyżowego (z podziałem zbioru uczącego na 10 części):

   a) w metodzie gradientowej wykorzystano transformację logistyczną i zbudowano model zagregowany dla liczby drzew równej: 200, 500 i 1000;

   b) w metodzie *k*-najbliższych sąsiadów wartość parametru *k* zmieniała się od 1 do 10;

   c) w metodzie sieci neuronowych zastosowano najbardziej popularną architekturę z jedną ukrytą warstwą oraz liczbą obserwacji w tej warstwie zmieniającą się od 1 do $\ln(N)$;

   d) w metodzie zagregowanych drzew Breimana liczbę zmiennych losowanych przy każdym podziale ustalano na poziomie $\frac{\sqrt{d}}{2}, \sqrt{d}, oraz\ 2\sqrt{d}$, liczbę drzew równą 100, 150 oraz 200, zaś minimalną liczbę obserwacji w liściu: 1, 5, 10 (zob. [Rozmus 2004]);

   e) w metodzie wektorów nośnych użyto wielomianowej funkcji jądrowej, zmieniając stopień wielomianu od 2 do 5, zaś wartość parametru C od 10^{-2} do 10^{2} (zob. [Trzęsiok 2004]).

3. Dla zbudowanych modeli optymalnych (dających najmniejszy błąd na zbiorze uczącym) na zbiorach testowych obliczono wartości błędu klasyfikacji. Otrzymane wyniki przedstawione zostały w tab. 2.

Tabela 2. Błąd klasyfikacji liczony na zbiorach testowych, otrzymany dla różnych metod dyskryminacji

Zbiór danych	BOOST_LN	LDA	KNN	NNET	RFOREST	SVM
P.I.Diabetes	0,278	0,218	0,286	*0,298*	0,225	**0,210**
Sonar	0,164	*0,268*	0,141	0,225	0,155	**0,113**
Satellite	0,108	*0,164*	0,118	0,152	**0,099**	0,111
Glass	*0,431*	0,343	0,329	0,397	**0,247**	0,343
Circle	0,063	0,510	*0,559*	0,052	0,069	**0,029**
Spirals	0,054	*0,463*	0,331	0,068	0,042	**0,037**

Źródło: opracowanie własne.

Najniższe wartości błędu klasyfikacji zostały zaznaczone pogrubioną czcionką, natomiast wartości najgorsze – kursywą.

Możemy zauważyć, iż najmniejsze błędy klasyfikacji otrzymujemy dla metody wektorów nośnych SVM lub dla zagregowanych drzew Breimana (RFOREST). Zdecydowanie najsłabiej z prawidłowym wyodrębnieniem klas radzi sobie klasyczna, liniowa analiza dyskryminacyjna.

Błędy klasyfikacji otrzymane dla modeli uzyskanych za pomocą metody gradientowej opartej na transformacji logistycznej są relatywnie małe. Jednak w porównaniu z innymi nieparametrycznymi metodami dyskryminacji, takimi jak SVM i RFOREST, dla omawianej metody otrzymujemy zawsze wartości błędów klasyfikacji o kilka procent wyższe.

4. Podsumowanie

Metoda MART, stosowana w regresji, generuje modele charakteryzujące się niskimi wartościami błędów predykcji. Na ogół jest ona najdokładniejsza w badanej grupie metod. Dla jej modyfikacji – omawianej metody gradientowej, otrzymaliśmy relatywnie dobre wartości błędów klasyfikacji, obliczonych na zbiorach testowych. Jednak w porównaniu z innymi metodami regresji otrzymujemy zawsze błędy klasyfikacji o co najmniej kilka procent wyższe niż np. dla modeli skonstruowanych na podstawie metody wektorów nośnych lub zagregowanych drzew Breimana.

Warto podkreślić, że atutem omawianej metody gradientowej jest jej przynależność do klasy metod nieparametrycznych. Nie wymaga ona znajomości rozkładów badanych zmiennych ani analitycznych postaci związków między nimi. Ponadto, ponieważ jest oparta na drzewach klasyfikacyjnych, pozwala na wprowadzanie do konstruowanego modelu zmiennych zarówno metrycznych, jak i niemetrycznych.

Literatura

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

Friedman J.H. (2001), *Greedy Function Approximation: a Gradient Boosting Machine*, Technical report, Dept. of Statistics, Stanford University.

Friedman J.H., Hastie T., Tibshirani R. (1999), *Additive Logistic Regression: a Statistical View of Boosting*, Technical report, Dept. of Statistics, Stanford University.

Gatnar E. (2000), *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Hastie T., Tibshirani R., Friedman J.H. (2001), *The Elements of Statistical Learning*, Springer-Verlag, New York.

Meyer D., Leisch F., Hornik K. (2002), *Benchmarking Support Vector Machines*, Report no. 78, Vienna University of Economics and Business Administration, http://www.wu-wien.ac.at/am/Download/report78.pdf.

Rozmus D. (2004), *Random forest jako metoda agregacji modeli dyskryminacyjnych*, [w:] K. Jajuga, M. Walesiak (red.), *Taksonomia 11, Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe AE we Wrocławiu nr 1022, AE, Wrocław, s. 441-448.

Trzęsiok M. (2004), *Analiza wybranych własności metody dyskryminacji wykorzystującej wektory nośne*, [w:] A.S. Barczak (red.), *Postępy ekonometrii*, AE, Katowice, s. 331-342.

Trzęsiok J. (2006), *Analiza wybranych własności metody MART*, [w:] K. Jajuga, M. Walesiak (red.), *Taksonomia 13, Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe AE we Wrocławiu nr 1126, AE, Wrocław, s. 510-518.

COMPARISON OF GRADIENT BOOSTING METHOD USING THE MULTIPLE LOGISTIC TRANSFORM WITH OTHER CLASSIFICATION METHODS

Summary

The main subject of this paper is related to gradient boosting method using the multiple logistic transform. This is a method for classification, which is based on MART – the multiple additive regression trees.

The goal of the article is to compare the presented method with other classification methods in terms of classification error generated by the final model. The analysis was conducted on some real and artificial benchmarking data sets.