

Joanna Trzęsiok

Akademia Ekonomiczna w Katowicach

ANALIZA WYBRANYCH WŁASNOŚCI METODY MART

1. Wstęp

Analiza regresji jest jedną z najczęściej stosowanych metod wielowymiarowej analizy statystycznej. Jej celem jest znalezienie funkcji f , która opisywałaby zależność zmiennej zależnej Y od zmiennych objaśniających $X_1, X_2, \dots, X_p$. Klasycznym podejściem do tego problemu jest metoda najmniejszych kwadratów. Zakłada ona, że postać funkcji f jest znana, poszukujemy zaś ocen parametrów tej funkcji. Metoda najmniejszych kwadratów nakłada kilka założeń na badane cechy, m.in. to, że zmienne objaśniające w modelu muszą mieć rozkład normalny, co znacznie ogranicza możliwości zastosowań tej metody. Dlatego też wprowadzone zostały metody nieparametryczne, które nie zakładają ani znajomości rozkładów badanych cech, ani postaci analitycznych związków między nimi.

Metody nieparametryczne stanowią liczną grupę zróżnicowanych i dynamicznie rozwijających się metod statystycznych. Należy do nich np. addytywna metoda drzew regresyjnych MART (*multiple additive regression trees*), która została zaproponowana przez Friedmana w 1999 r. [1999a]. Należy ona do grupy metod agregacyjnych, wykorzystujących sekwencyjne łączenie modeli składowych (*boosting*) (por. [Freund, Schapire 1997]). Funkcje składowe są w każdym kroku algorytmu systematycznie poprawiane, co w efekcie prowadzi do uzyskania modelu charakteryzującego się dużą dokładnością predykcji. Wyznaczanie parametrów funkcji wchodzących do modelu odbywa się poprzez minimalizację wartości funkcji straty. Wybór wspomnianej funkcji straty ma wpływ na jakość końcowego modelu.

W artykule główny nacisk położony zostanie na zweryfikowanie hipotezy, iż metoda MART jako jedna z metod nieparametrycznych jest odporna na występowanie w zbiorze uczącym szumu oraz wartości oddalonych.

2. Metoda MART

2.1. Postać funkcji regresji w metodzie MART

W metodach agregacyjnych sekwencyjnego łączenia modeli funkcja regresji ma postać addytywną:

$$f(\mathbf{X}) = \sum_{m=0}^M \beta_m \cdot b(\mathbf{X}, \Theta_m), \quad (1)$$

gdzie: $b(\mathbf{X}, \Theta_m)$ – funkcja składowa modelu, charakteryzowana przez zbiór parametrów oznaczonych przez Θ_m ,

β_m – tzw. współczynnik rozwoju (*expansion coefficient*) określający wagę składowej $b(\mathbf{X}, \Theta_m)$.

W omawianej metodzie MART za funkcje składowe $b(\mathbf{X}, \Theta_m)$ przyjmuje się drzewa regresyjne.

Metoda drzew regresyjnych (por. [Gatnar 2000]) polega na podziale przestrzeni zmiennych objaśniających na rozłączne obszary R_j (dla $j = 1, \dots, J$) i przypisaniu wszystkim obserwacjom z danego obszaru odpowiedniej wartości γ_j :

$$\mathbf{x} \in R_j \Rightarrow f(\mathbf{x}) = \gamma_j = \bar{y}_j \text{ dla } j = 1, \dots, J. \quad (2)$$

W zagadnieniu regresji jako γ_j przyjmujemy średnią z wszystkich realizacji zmiennej zależnej, dla których współrzędne $\mathbf{x}_i$ leżą w obszarze R_j .

Otrzymany model można zapisać w postaci:

$$T(\mathbf{x}, \Theta) = \sum_{j=1}^J \gamma_j \cdot I(\mathbf{x} \in R_j), \quad (3)$$

z parametrami $\Theta = \{R_j, \gamma_j\}_{j=1, \dots, J}$.

Zatem model zagregowany, złożony z drzew regresyjnych $T(\mathbf{x}, \Theta_m)$, możemy przedstawić jako:

$$f(\mathbf{x}) = \sum_{m=0}^M T(\mathbf{x}, \Theta_m) = \sum_{m=0}^M \sum_{j=1}^{J_m} \gamma_{jm} \cdot I(\mathbf{x} \in R_{jm}), \quad (4)$$

gdzie $\Theta_m = \{R_{jm}, \gamma_{jm}\}_{m=0, \dots, M, j=1, \dots, J_m}$.

Estymatory parametrów Θ_m w modelu zagregowanym (4) otrzymujemy poprzez minimalizację funkcji straty $L(y, f(\mathbf{x}))$ w zbiorze uczącym U :

$$\hat{\Theta}_m = \arg \min_{(\Theta_m)_{m=0, \dots, M}} \sum_{i=1}^N L\left(y_i, \sum_{m=0}^M T(\mathbf{x}_i, \Theta_m)\right). \quad (5)$$

W zagadnieniu minimalizacji (5) możemy wykorzystywać różne funkcje straty. Najczęściej są to jednak:

- kwadratowa funkcja straty:

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2, \quad (6)$$

- absolutna funkcja straty:

$$L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|, \quad (7)$$

- funkcja straty Hubera:

$$L(y, f(\mathbf{x})) = \begin{cases} (y - f(\mathbf{x}))^2, & \text{gdy } |y - f(\mathbf{x})| < \delta \\ \delta(2|y - f(\mathbf{x})| - \delta), & \text{gdy } |y - f(\mathbf{x})| \geq \delta \end{cases}. \quad (8)$$

W dalszej analizie wybranych własności metody MART zostanie zbadane, która z powyższych funkcji straty daje modele charakteryzujące się największą dokładnością predykcji.

2.2. Wyznaczanie parametrów modelu – metoda największego spadku

Zagadnienie minimalizacji funkcji straty (5) jest złożone obliczeniowo, dlatego do jego rozwiązania wykorzystujemy najczęściej **strategię wspinaczki** (por. [Hastie, Tibshirani, Friedman 2001]), polegającą na tym, że w każdym kroku (dla $m = 1, \dots, M$) szukane jest rozwiązanie optymalne jedynie w sensie lokalnym:

$$\hat{\Theta}_m = \arg \min_{\Theta_m} \sum_{i=1}^N L(y_i, f_{m-1}(\mathbf{x}_i) + T(\mathbf{x}_i, \Theta_m)). \quad (9)$$

Po oszacowaniu estymatora $\hat{\Theta}_m$ przyjmujemy, że składowa f_m jest sumą funkcji uzyskanej w poprzedniej iteracji oraz drzewa $T(\mathbf{x}, \hat{\Theta}_m)$:

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + T(\mathbf{x}, \hat{\Theta}_m). \quad (10)$$

Model końcowy jest funkcją uzyskaną w ostatnim kroku algorytmu:

$$f(\mathbf{x}) = f_M(\mathbf{x}) = \sum_{m=0}^M T(\mathbf{x}, \hat{\Theta}_m). \quad (11)$$

Strategię wspinaczki cechuje prostota obliczeniowa, a dla dobrze określonego kryterium uzyskujemy model końcowy bliski globalnemu rozwiązaniu optymalnemu zagadnienia minimalizacji (5).

Do wyznaczenia parametrów modelu (4) w metodzie MART wykorzystana została koncepcja strategii wspinaczki połączona z **metodą największego spadku** (*steepest descent algorithm*) [Hastie, Tibshirani, Friedman 2001].

W metodzie największego spadku rozważamy zagadnienie minimalizacji funkcji straty na zbiorze uczącym. Szukamy więc funkcji f postaci (4), spełniającej warunek:

$$\hat{\mathbf{f}} = \arg \min_{\mathbf{f}} \sum_{i=1}^N L(y_i, f(\mathbf{x}_i)), \quad (12)$$

gdzie:

$$\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_N)). \quad (13)$$

Aby uzyskać największą redukcję wartości funkcji straty, przesuwamy się przeciwnie do jej gradientu. Zatem przyjmujemy:

$$\mathbf{f} = -\rho_m \cdot \mathbf{g}_m, \quad (14)$$

gdzie $\rho_m \in \mathbf{R}$ jest parametrem nazywanym długością kroku, a $\mathbf{g}_m \in \mathbf{R}^N$ – gradientem funkcji straty L o składowych:

$$g_{im} = \left[\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)} \right]_{f(\mathbf{x}_i)=f_{m-1}(\mathbf{x}_i)} \quad \text{dla } i = 1, \dots, N. \quad (15)$$

Algorytm metody największego spadku możemy przedstawić w następujących krokach:

1. Przyjmujemy model początkowy:

$$f_0(\mathbf{x}) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma). \quad (16)$$

2. Dla $m = 1, \dots, M$ wykonujemy następujące kroki:

- a) obliczamy składowe gradientu funkcji straty według wzoru (15),
- b) wyznaczamy wartość parametru ρ_m :

$$\rho_m = \arg \min_{\rho} \sum_{i=1}^N L(y_i, f_{m-1}(\mathbf{x}_i) - \rho \cdot \mathbf{g}_{im}), \quad (17)$$

c) przyjmujemy:

$$\mathbf{f}_m = \mathbf{f}_{m-1} - \rho_m \cdot \mathbf{g}_m. \quad (18)$$

3. Model końcowy otrzymujemy w ostatnim kroku:

$$f(\mathbf{x}) = f_M(\mathbf{x}). \quad (19)$$

Pełną postać algorytmu metody MART, wykorzystującą przedstawioną strategię wspinaczki oraz metodę największego spadku, zaprezentowano w podrozdziale 2.4 niniejszej pracy.

2.3. Podejście stochastyczne do metody MART

W 1999 r. Friedman zaproponował pewne udoskonalenie metody MART [Friedman 1999b], polegające na tym, że w każdym kroku algorytmu modele składowe tworzone są na podstawie prób uczących losowo wybranych z całego zbioru danych $U = \{(\mathbf{x}_i, y_i) : i = 1, 2, \dots, N\}$.

Oznaczmy przez $\{\pi(i)\}_{i=1, \dots, N}$ dowolnie ustaloną permutację zbioru indeksów $I = \{1, 2, \dots, N\}$. W każdej iteracji elementy zbioru I zostają spermutowane za pomocą $\{\pi(i)\}_{i=1, \dots, N}$, przez co zmieniona zostaje kolejność obiektów w zbiorze uczącym U . Do budowy modelu składowego wybranych zostaje tylko $\tilde{N} < N$ pierwszych obiektów zbioru U . Zatem w każdym kroku algorytmu modele składowe powstają na podstawie zbiorów:

$$\{(\mathbf{x}_{\pi(i)}, y_{\pi(i)}) : i = 1, \dots, \tilde{N}\}. \quad (20)$$

Zostało pokazane [Friedman 1999b], iż modele zbudowane w oparciu o gradientową metodę *boosting* w ujęciu stochastycznym charakteryzują się najlepszymi wartościami współczynnika determinacji R^2 wtedy, gdy w kolejnych etapach algorytmu przyjmuje się frakcję obserwacji wylosowanych do prób na poziomie 0,4-0,6.

2.4. Algorytm MART

1. Przyjmujemy model początkowy:

$$f_0(\mathbf{x}) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, f(\mathbf{x}_i)). \quad (21)$$

2. Dla $m = 1, \dots, M$ wykonujemy kroki:

a) permutujemy za pomocą $\{\pi(i)\}_{i=1, \dots, N}$ obserwacje w zbiorze uczącym U ,

b) dla $n = 1, \dots, \tilde{N}$ obliczamy składowe gradientu funkcji straty L :

$$g_{\pi(i)m} = - \left[\frac{\partial L(y_{\pi(i)}, f(\mathbf{x}_{\pi(i)}))}{\partial f(\mathbf{x}_{\pi(i)})} \right]_{f=f_{m-1}}, \quad (22)$$

c) szukamy obszarów R_{jm} (dla $j = 1, \dots, J_m$), posługując się zbiorem danych:

$$\{(\mathbf{x}_{\pi(i)}, g_{\pi(i)m}) : i = 1, \dots, \tilde{N}\}, \quad (23)$$

d) dla $j = 1, \dots, J_m$ obliczamy:

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{\mathbf{x}_{\pi(i)} \in R_{jm}} L(y_{\pi(i)}, f_{m-1}(\mathbf{x}_{\pi(i)}) + \gamma), \quad (24)$$

e) przyjmujemy:

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \sum_{j=1}^{J_m} \gamma_{jm} \cdot I(\mathbf{x} \in R_{jm}). \quad (25)$$

3. Model końcowy ma postać:

$$f(\mathbf{x}) = f_M(\mathbf{x}) = \sum_{j=1}^M \gamma_{jM} \cdot I(\mathbf{x} \in R_{jM}). \quad (26)$$

W pierwszym kroku algorytmu przyjmujemy model początkowy, a następnie permutujemy obserwacje w zbiorze uczącym. Do budowy modeli składowych będziemy wykorzystywać jedynie pewną frakcję zbioru danych U . W kolejnej iteracji wyznaczamy składowe gradientu funkcji straty, według wzoru (22), które nazywamy pseudoresztami. Wykorzystując metodę drzew regresyjnych, szukamy rozłącznych obszarów R_{jm} , na podstawie zbioru danych $\{(\mathbf{x}_{\pi(i)}, g_{\pi(i)m}) : i = 1, \dots, \tilde{N}\}$, w których wartości zmiennej zależnej zastąpiono obliczonymi w kroku 2b pseudoresztami. Następnie, poprzez minimalizację funkcji straty, wyznaczone zostają estymatory parametrów γ_{jm} . W kolejnym kroku poprawiamy składową $f_m(\mathbf{x})$ i powtarzamy powyższą procedurę. Model końcowy jest modelem uzyskanym w ostatnim etapie algorytmu.

3. Analiza wybranych własności metody MART

3.1. Zbiory danych

Aby zweryfikować hipotezę o odporności metody MART na występowanie w zbiorze uczącym szumu oraz wartości oddalonych, wykonana została analiza błędów predykcji dla modeli regresji zbudowanych na zbiorach *Friedman 1*, *Friedman 2* i *Friedman 3*. Są to sztuczne zbiory danych, wygenerowane komputerowo za pomocą funkcji dostępnych w pakiecie `mlbench` programu statystycznego R, zgodnie z koncepcją Friedmana [1991]. Zbiory te są standardowo wykorzystywane do porównywania własności różnych funkcji regresji. Tabela 1 przedstawia ich wybrane charakterystyki:

Tabela 1. Charakterystyka zbiorów danych wykorzystanych w analizie

Zbiory	Liczebność	L. zmiennych	Postać funkcji
<i>Friedman 1</i>	500	10	$f(\mathbf{x}) = 10 \sin(\pi x_1 x_2) + 20(x_1 - 0,5)^2 + 10x_4 + 5x_5 + e$
<i>Friedman 2</i>	500	4	$f(\mathbf{x}) = \sqrt{x_1^2 + \left(x_2 x_3 - \frac{1}{x_1 x_4}\right)^2} + e$
<i>Friedman 3</i>	500	4	$f(\mathbf{x}) = \arctg \frac{x_2 x_3 - \frac{1}{x_1 x_4}}{x_1} + e$

Źródło: opracowanie własne.

Składnik losowy e jest szumem gaussowskim o rozkładzie $N(0, s)$, gdzie s jest parametrem zadawanym przez użytkownika.

Do obliczeń wykorzystany został pakiet statystyczny R oraz program TreeNet.

3.2. Procedura badawcza i wyniki analizy

W pierwszej grupie eksperymentów testowano odporność metody MART na występowanie w zbiorze uczącym szumu.

- Do każdego ze zbiorów uczących dodany został szum na poziomie od 5 do 40% zmienności mierzonej wariancją. Poziom szumu był regulowany przez dobieranie odpowiednich wartości odchylenia standardowego s składnika losowego w funkcji generującej dane.
- Modele regresyjne zbudowane zostały dla trzech różnych funkcji straty przy różnych poziomach szumu.
- Modele składowe zbudowano w oparciu o próby losowe, stanowiące 50% zbioru uczącego.
- Zbiór testowy w każdym z przypadków składał się z 200 elementów i nie był zaburzony szumem.

Dla każdego zbudowanego modelu regresyjnego obliczono na zbiorze testowym współczynnik determinacji:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} . \quad (27)$$

Wyniki analizy przedstawione zostały w tabelach 2-4.

Tabela 2. Wartości współczynnika determinacji obliczonego na zbiorze testowym *Friedman 1*, dla różnych funkcji straty, przy różnych poziomach szumu

Funkcje straty	5%	10%	20%	30%	40%
F. Hubera	0,923	0,905	0,905	0,888	0,869
F. absolutna	0,914	0,857	0,882	0,860	0,832
F. kwadratowa	0,923	0,909	0,909	0,889	0,871

Źródło: obliczenia własne.

Tabela 3. Wartości współczynnika determinacji obliczonego na zbiorze testowym *Friedman 2*, dla różnych funkcji straty, przy różnych poziomach szumu

Funkcje straty	5%	10%	20%	30%	40%
F. Hubera	0,988	0,989	0,975	0,979	0,951
F. absolutna	0,980	0,988	0,962	0,973	0,936
F. kwadratowa	0,993	0,990	0,981	0,979	0,959

Źródło: obliczenia własne.

Tabela 4. Wartości współczynnika determinacji obliczonego na zbiorze testowym *Friedman 3*, dla różnych funkcji straty, przy różnych poziomach szumu

Funkcje straty	5%	10%	20%	30%	40%
F. Hubera	0,948	0,898	0,904	0,934	0,884
F. absolutna	0,932	0,868	0,874	0,902	0,869
F. kwadratowa	0,947	0,917	0,917	0,939	0,887

Źródło: obliczenia własne.

W drugiej grupie eksperymentów do zbioru uczącego wprowadzone zostały wartości oddalone.

a) z każdego ze zbiorów uczących wylosowano 2% obserwacji, które zamieniono na wygenerowane komputerowo wartości oddalone,

b) zgodnie z ustaleniami zawartymi w funkcjach pakietu statystycznego R przyjęto, iż z wartościami oddalonymi y_i mamy do czynienia, gdy:

$$y_i \notin (Q_1 - 3Q, Q_3 + 3Q),$$

gdzie Q_1 to kwartył pierwszy, Q_3 – kwartył trzeci, a Q – odchylenie ćwiartkowe,

c) modele regresyjne zbudowano dla trzech różnych funkcji straty,

d) podobnie jak poprzednio, do budowy modeli składowych wykorzystywano połowę obserwacji wylosowanych ze zbioru uczącego,

e) zbiór testowy nie był zaburzony ani szumem, ani wartościami oddalonymi i zawierał 200 elementów.

Tabela 5. Wartości współczynnika determinacji obliczonego na zbiorach testowych, w przypadku występowania wartości oddalonych, dla różnych funkcji straty

Funkcje straty	<i>Friedman 1</i>	<i>Friedman 2</i>	<i>Friedman 3</i>
F. Hubera	0,912	0,990	0,931
F. absolutna	0,892	0,979	0,893
F. kwadratowa	0,760	0,938	0,912

Źródło: obliczenia własne.

Ponownie dla każdego modelu obliczony został na zbiorze testowym współczynnik determinacji. Tabela 5 przedstawia otrzymane wyniki.

4. Podsumowanie

Na podstawie wyników przedstawionych w tab. 2-5, możemy zauważyć, że nawet w przypadku zaburzenia obserwacji szumem na poziomie 40% otrzymujemy bardzo wysokie wartości współczynnika determinacji obliczonego na zbiorze testowym. Do analogicznych wniosków możemy dojść, analizując wyniki w przypadku modeli zbudowanych w oparciu o zbiory uczące, w których występują wartości oddalone.

Możemy więc powiedzieć, że metoda MART jest odporna na występowanie w zbiorze uczącym szumu oraz wartości oddalonych.

Największe wartości współczynnika R^2 dla modeli zbudowanych na zbiorach danych zaburzonych szumem otrzymujemy dla funkcji straty kwadratowej oraz Hubera. Natomiast w przypadku występowania wartości oddalonych najlepsze wyniki otrzymujemy przy zastosowaniu funkcji Hubera. Nasuwa się więc wniosek praktyczny dla użytkownika programu TreeNet, iż do budowy modelu regresyjnego metodą MART, na rzeczywistych zbiorach danych, w których nieznany jest poziom szumu ani informacja o występowaniu wartości oddalonych, najbezpieczniej jest wykorzystywać właśnie funkcję straty Hubera.

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” nr 55, s. 119-139.
- Friedman J.H. (1991), *Multivariate Adaptive Regression Splines*, „Annals of Statistics” nr 19, s.1-141.
- Friedman J.H. (1999a), *Greedy Function Approximation: a Gradient Boosting Machine*, Technical report, Dept. of Statistics, Stanford University.
- Friedman J.H. (1999b), *Stochastic Gradient 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.

THE ANALYSIS OF SOME PROPERTHIES OF MART

Summary

In the paper MART – the multiple additive regression trees are presented. This is one of the boosting methods. It seeks for a linear combination of base functions which are regression trees. The final model in the stagewise approach is obtained by improving the performance of the components in each iteration and has a high accuracy.

The primary goal of this paper was to show that MART is resistant against noise and outliers.