PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1189 — 2007
Zastosowania metod ilościowych

Michał Trzęsiok
Akademia Ekonomiczna w Katowicach

SYMULACYJNE PORÓWNANIE JAKOŚCI MODELI OTRZYMANYCH METODĄ WEKTORÓW NOŚNYCH Z INNYMI MODELAMI REGRESJI

1. Wstęp

Wśród wielu nowych metod regresji na uwagę szczególną zasługuje metoda wektorów nośnych (SVM – *Support Vector Machines*). Choć została ona skonstruowana pierwotnie do realizowania zadań dyskryminacji [Vapnik 1998], to podobnie jak w przypadku drzew klasyfikacyjnych i regresyjnych algorytm tej metody można przeformułować tak, aby służył do budowy modeli regresji. Ponadto regresyjna metoda wektorów nośnych ma wiele pożądanych własności, które zostały przeniesione z metody dyskryminacyjnej, tj. brak konieczności sprawdzania przez użytkownika założeń o rozkładach zmiennych diagnostycznych, nieliniowość otrzymanego modelu, relatywnie małe błędy predykcji dla obserwacji ze zbioru testowego, odporność na występowanie szumu w zbiorze uczącym, bardzo liczna przestrzeń hipotez z dużą różnorodnością przeszukiwanych, potencjalnych funkcji regresji (zob. [Trzęsiok 2005]).

Podobnie jak w przypadku dyskryminacji, nie sposób nie wspomnieć o tych własnościach metody wektorów nośnych, które świadczą na jej niekorzyść. Mianowicie jest ona jedną z metod eksploracji danych, co oznacza, że jej działanie jest w dużym stopniu zautomatyzowane, podobne do „czarnej skrzynki", wyniki zaś w niewielkim stopniu poddają się interpretacji. Podsumowując, otrzymywany model, jako narzędzie służące do predykcji, charakteryzuje się dużą precyzją, lecz znajomość analitycznej postaci modelu w niewielkim stopniu pozwala badaczowi na poznanie i zrozumienie istoty relacji między zmiennymi uwzględnionymi w modelu.

Metoda wektorów nośnych uważana jest obecnie, oprócz zagregowanych drzew klasyfikacyjnych, za najdokładniejsze narzędzie w dyskryminacji. W dalszej części artykułu przedstawiona zostanie pokrótce metoda wektorów nośnych w regresji oraz próba empirycznego zweryfikowania, na ile metoda wektorów nośnych jest

konkurencyjna pod względem dokładności predykcji wobec innych znanych metod regresji, takich jak klasyczna regresja liniowa, a także wybrane metody nieparametryczne: metoda rzutowania, sieci neuronowe, drzewa regresyjne oraz modele zagregowane łączące wiele modeli drzew.

2. Metoda wektorów nośnych w regresji

Zarówno w przypadku dyskryminacji, jak i regresji w algorytmie metody wektorów nośnych wyznaczana jest funkcja o postaci liniowej. Nieliniowość metody wynika z tego, że obserwacje ze zbioru uczącego są transformowane do nowej przestrzeni o dużo większym wymiarze za pomocą nieliniowego przekształcenia φ. Zatem przy danym zbiorze uczącym $\left\{\left(\mathbf{x}^1, y^1\right), \ldots, \left(\mathbf{x}^N, y^N\right)\right\}$, gdzie $\mathbf{x}^i \in \mathbf{R}^d$ i $y^i \in \mathbf{R}$ dla $i = 1, \ldots, N$, oraz nieliniowej transformacji $\varphi: \mathbf{R}^d \rightarrow \mathbf{Z}$, zagadnienie regresji polegać będzie na wyznaczeniu w nowej przestrzeni cech $\mathbf{Z}$ funkcji liniowej o postaci:

$$y = f(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot \varphi(\mathbf{x}) + \beta_0, \tag{1}$$

gdzie $\varphi(\mathbf{x}^i) \in \mathbf{Z}$ oraz $y^i \in R$ dla $i = 1, \ldots, N$.

Równanie:

$$\boldsymbol{\beta} \cdot \varphi(\mathbf{x}) + \beta_0 = 0 \tag{2}$$

definiuje hiperpłaszczyznę, o której w dyskryminacji zakłada się, że ma być optymalną hiperpłaszczyzną rozdzielającą klasy, tj. położoną tak, aby rozdzielała klasy oraz by symetryczne otoczenie hiperpłaszczyzny, nie zawierające żadnej obserwacji, miało jak największy promień. Tak postawiony problem można sformułować w postaci zadania optymalizacji wypukłej z liniowymi ograniczeniami [Cristianini, Shawe-Taylor 2000; Trzęsiok 2004]. W regresji również wyznaczana będzie optymalna hiperpłaszczyzna, lecz nie z warunkiem rozdzielania klas, tylko zakładając, że punkty przetransformowanego zbioru uczącego leżą jak najbliżej poszukiwanej hiperpłaszczyzny. Ten geometrycznie przedstawiony warunek można zapisać analitycznie jako poszukiwanie funkcji postaci (1), która minimalizuje wartość funkcjonału mierzącego jakość dopasowania funkcji do danych ze zbioru uczącego. Twórca metody Vapnik [1998] do pomiaru dopasowania zaproponował funkcję straty, niewrażliwą na odchylenia rzędu $\varepsilon > 0$:

$$L^{\varepsilon}(y, f(\mathbf{x}, \boldsymbol{\beta})) = \begin{cases} 0, & \text{gdy } |y - f(\mathbf{x}, \boldsymbol{\beta})| \leq \varepsilon, \\ |y - f(\mathbf{x}, \boldsymbol{\beta})| - \varepsilon, & \text{gdy } |y - f(\mathbf{x}, \boldsymbol{\beta})| > \varepsilon, \end{cases} \tag{3}$$

co oznacza, że dana wartość zmiennej y^i jest dobrze aproksymowana przez funkcję f, jeśli odległość odpowiadającego jej punktu od hiperpłaszczyzny nie jest większa

niż z góry zadane $\varepsilon > 0$ (obserwacje ze zbioru uczącego znajdują się w symetrycznym otoczeniu hiperpłaszczyzny o promieniu epsylon). Uwzględniając wszystkie punkty ze zbioru uczącego, otrzymujemy kryterium jakości dopasowania i tym samym kryterium wyznaczania funkcji f:

$$R_{emp}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} L(y^i, f(\mathbf{x}^i, \boldsymbol{\beta})). \tag{4}$$

Powyższe kryterium, czyli minimalizacja funkcjonału zwanego *ryzykiem empirycznym*, oznacza jedynie jak najlepsze dopasowanie funkcji aproksymującej do obserwacji ze zbioru uczącego i może doprowadzić do nadmiernego dopasowania (*overfitting*), a w konsekwencji do dużych błędów predykcji. Dlatego w metodzie wektorów nośnych minimalizacji podlega funkcjonał będący sumą ryzyka empirycznego oraz wyrażenia $\frac{1}{2}\|\boldsymbol{\beta}\|^2$, którego mniejsze wartości implikują większą prostotę modelu i lepsze wyniki predykcji poza zbiorem uczącym. Minimalizujemy więc funkcjonał:

$$\tfrac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N} L^{\varepsilon}(y^i, f(\mathbf{x}^i, \boldsymbol{\beta})), \tag{5}$$

gdzie parametr C określa kompromis między prostotą modelu a jego dopasowaniem do obserwacji ze zbioru uczącego.

W rzeczywistych zbiorach danych bardzo często występują obserwacje nietypowe, w których wartości cech zaburzone są szumem lub różnego typu błędami. W celu uelastycznienia metody i uodpornienia jej na występowanie takiego zjawiska wprowadzone zostają zmienne $\xi_1, \ldots, \xi_N, \xi_1^*, \ldots, \xi_N^* \geq 0$. Dodatnia wartość takiej zmiennej w modelu oznacza, że odpowiadająca jej obserwacja może znajdować się poza epsylonowym otoczeniem wyznaczanej hiperpłaszczyzny. Zmniejsza to oczywiście jakość aproksymacji na zbiorze uczącym ale – szczególnie w przypadku obserwacji obarczonych błędami – może mieć pozytywny wpływ na jakość predykcji dla nowych obiektów ze zbioru rozpoznawanego.

Omówione zadanie minimalizacji z liniowymi ograniczeniami można zapisać w postaci analitycznej:

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum\limits_{i=1}^{N}(\xi_i + \xi_i^*), \\ y^i - \left(\boldsymbol{\beta}\cdot\varphi(\mathbf{x}^i) + \beta_0\right) \leq \varepsilon + \xi_i, \qquad i = 1,\ldots,N, \\ -y^i + \left(\boldsymbol{\beta}\cdot\varphi(\mathbf{x}^i) + \beta_0\right) \leq \varepsilon + \xi_i^*, \quad i = 1,\ldots,N, \\ \xi_i, \xi_i^* \geq 0. \end{cases} \tag{6}$$

Zadanie to można rozwiązać metodą mnożników Lagrange'a. Z warunku Kuhna-Tuckera, który spełnia rozwiązanie optymalne, wynika, że wiele współczynników

Lagrange'a jest równych zeru. Te obserwacje, którym odpowiadają niezerowe współczynniki Lagrange'a, nazywamy wektorami nośnymi. Z geometrycznego punktu widzenia są to obserwacje leżące na brzegu epsylonowego otoczenia wyznaczonej hiperpłaszczyzny, a także te obiekty, które leżą poza tym otoczeniem. Jedynie wektory nośne mają wpływ na położenie hiperpłaszczyzny, a tym samym na funkcję regresji [Trzęsiok 2005].

O jakości modelu regresji, otrzymanego metodą wektorów nośnych, decyduje wybór wartości parametru C oraz przede wszystkim wybór odpowiedniej transformacji nieliniowej φ. Okazuje się, że zarówno w zadaniu optymalizacyjnym, po przejściu do jego postaci dualnej, jak i w postaci analitycznej rozwiązania, przekształcenie φ występuje jedynie związane iloczynem skalarnym [Gunn 1997]. Wobec tej uwagi nie jest konieczne definiowanie φ wprost. Wystarczy bowiem wybrać funkcję, która będzie definiowała iloczyn skalarny w $\mathbf{Z}$, co znacznie redukuje problemy analityczne i numeryczne metody. W metodzie wektorów nośnych, zamiast definiować transformację φ, użytkownik decyduje o wyborze sposobu liczenia iloczynu skalarnego w pewnej przestrzeni o większym wymiarze. Standardowo wykorzystuje się do tego celu funkcje z rodziny funkcji jądrowych postaci $\mathrm{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}) \cdot \varphi(\mathbf{v})$, w wypadku których udowodniono, że definiują iloczyny skalarne w pewnej przestrzeni. Najczęściej stosowane funkcje jądrowe to:

1) funkcja Gaussa: $\mathrm{K}(\mathbf{u},\mathbf{v}) = \exp(-\gamma\|\mathbf{u}-\mathbf{v}\|^2)$,
2) funkcja wielomianowa: $\mathrm{K}(\mathbf{u},\mathbf{v}) = \gamma(\mathbf{u}\cdot\mathbf{v}+\delta)^d$, $d = 1, 2, \ldots$,
3) funkcja sigmoidalna: $\mathrm{K}(\mathbf{u},\mathbf{v}) = \tanh(\gamma\mathbf{u}\cdot\mathbf{v}+\delta)$.

Warto zaznaczyć, że wybór samej funkcji jądrowej oraz wybór wartości parametrów tej funkcji ma kluczowe znaczenie dla jakości otrzymanego modelu. Niestety, brak jest jednoznacznej odpowiedzi na pytanie, jak dobierać rodzaj funkcji do przeprowadzanej analizy i jak identyfikować optymalne wartości parametrów. W tym celu użytkownik zazwyczaj musi przeprowadzić wstępną analizę symulacyjną dla kilku funkcji jądrowych i odpowiednio dużego zakresu wartości parametrów metody z wykorzystaniem np. sprawdzania krzyżowego. Badania empiryczne wskazują jednak na pewną przewagę wielomianowych funkcji jądrowych oraz funkcji Gaussa z odpowiednio dobranymi parametrami (zob. [Trzęsiok 2004]).

Przy zadanej funkcji jądrowej definiującej iloczyn skalarny w $\mathbf{Z}$ szukaną funkcję regresji możemy zapisać w postaci:

$$f(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*)\mathrm{K}(\mathbf{x}^i, \mathbf{x}) + \hat{\beta}_0, \tag{7}$$

gdzie α_i, α_i^* są różnymi od zera współczynnikami Lagrange'a z rozwiązania optymalnego (odpowiadającymi wektorom nośnym) [Cristianini, Shawe-Taylor 2000].

3. Porównanie modeli pod względem dokładności predykcji

3.1. Kryterium porównawcze

Wysoka pozycja metody wektorów nośnych w grupie metod dyskryminacji, pod względem zarówno małych błędów klasyfikacji, jak i możliwości zastosowań do różnych typów danych, wydaje się niepodważalna (zob. [Trzęsiok 2004]). Wobec faktu, że wiele własności metody SVM w dyskryminacji w sposób naturalny przenosi się na jej odpowiednik w regresji, zachodzi pytanie, czy w regresji również zauważalna będzie dominująca pozycja metody wektorów nośnych. Odpowiedź na to pytanie zostanie sformułowana po przeprowadzeniu empirycznego porównania jakości różnych modeli regresji zbudowanych na zbiorach danych standardowo, wykorzystywanych do badania własności i porównywania metod wielowymiarowej analizy statystycznej.

Za miarę jakości modelu przyjęto błąd średniokwadratowy. Ponieważ podstawowym celem regresji jest predykcja na nowych obiektach, spoza zbioru uczestniczącego w procesie budowy modelu, więc wartości błędów średniokwadratowych obliczane będą na zbiorach testowych, stanowiących wyodrębnioną część 33% całego analizowanego zbioru. Pozostałe 66% tworzy zbiór uczący.

3.2. Porównywane metody

Badana grupa metod obejmuje:

1) SVM – metodę wektorów nośnych,
2) LM – klasyczną regresję liniową,
3) PPR – metodę rzutowania,
4) NNET – sieć neuronową,
5) RPART – drzewa regresyjne,
6) RFOREST – zagregowane drzewa regresyjne Breimana,
7) BAGG – metodę łączenia równoległego drzew przez rozszerzanie (bagging)

Poszczególne metody zostały oznaczone symbolami zgodnymi z nazwami funkcji realizujących je w pakiecie statystycznym R. Program ten wraz z dodatkowymi bibliotekami został wykorzystany do przeprowadzenia analizy.

3.3. Zbiory danych

W analizie wykorzystano pięć zbiorów danych. Trzy z nich: Friedman1 (300 obserwacji, 11 zmiennych), Friedman2 (300 obserwacji, 5 zmiennych), Friedman3 (300 obserwacji, 5 zmiennych), to zbiory sztuczne, tj. takie, których obserwacje generowane są komputerowo. Zostały one specjalnie zaprojektowane przez Friedmana [1991] tak, by zawierały wiele elementów wymagających od metody odpowiedniego modelowania (nieliniowość, szum, zmienne diagnostyczne, które w ogóle nie biorą udziału w generowaniu wartości zmiennej objaśnianej itp.). Zbiory

te są powszechnie stosowane do porównań, podobnie jak dwa kolejne: Boston (506 obserwacji, 14 zmiennych) i Servo (167 obserwacji i 5 zmiennych). Są to zbiory danych rzeczywistych dostępne w internetowej bazie „UCI Repository Of Machine Learning Databases" zlokalizowanej w Uniwersytecie Kalifornijskim[1]. Wszystkie badane zbiory i geneatory zbiorów sztucznych dostępne są także w pakiecie mlbench programu statystycznego R.

3.4. Procedura badawcza i porównanie modeli

W pierwszym etapie z badanego zbioru w sposób losowy wybierano 33% obserwacji do zbioru testowego. Obserwacje te nie uczestniczyły w procesie budowania modelu regresji. Wykorzystywane były jedynie do wyliczenia błędu średniokwadratowego stanowiącego kryterium porównania.

Większość badanych metod wymaga od użytkownika ustalenia wartości pewnych parametrów. Symulacyjnie na zbiorze uczącym budowano wiele modeli dla różnych układów tychże parametrów. Do porównania wybierany był ten model (taki układ wartości parametrów), który dawał najmniejszy błąd klasyfikacji liczony metodą sprawdzania krzyżowego z podziałem zbioru uczącego na 10 części. Przeszukiwane zakresy parametrów dla poszczególnych metod są następujące:

a) w metodzie wektorów nośnych użyto wielomianowej funkcji jądrowej, zmieniając stopień wielomianu od 3 do 5 (zob. [Trzęsiok 2004]), wartość parametru C od 10^{-2} do 10^{2}, epsylon równe 0,1 oraz 0,5,

b) w metodzie rzutowania wartość parametru opisującego początkową liczbę funkcji składowych modelu przyjmowano na poziomie 10, 15, 20, 25, końcowa zaś liczba tychże funkcji w modelu zmieniała się od 1 do 10,

c) sieć neuronowa z jedną ukrytą warstwą, z liczbą obserwacji w warstwie ukrytej zmieniającą się od 1 do $\ln(N)$,

d) dla drzew regresyjnych wymaganą minimalną liczbę obserwacji w węźle, aby nastąpił dalszy podział, ustalano na poziomie od 3 do 10, a kryterium minimalnej poprawy jakości modelu (przycinanie drzewa) na poziomie od 1 do 3%,

e) w metodzie zagregowanych drzew regresyjnych Breimana liczbę zmiennych losowanych przy każdym podziale ustalano na poziomie $\frac{\sqrt{d}}{2}$, $\sqrt{d}$, *oraz* $2\sqrt{d}$ (d – liczba zmiennych), liczbę drzew równą 100 oraz 200, minimalną zaś liczbę obserwacji w liściu: 1, 5, 10.

3.5. Wyniki analizy

Po wyznaczeniu modelu optymalnego dla każdej z metod i każdego z badanych zbiorów według kryterium błędu średniokwadratowego liczonego metodą sprawdzania krzyżowego na zbiorze uczącym obliczano błąd średniokwadratowy na

[1] Dostępne pod adresem: ftp://ftp.ics.uci.edu/pub/machine-learning-databases, http://www.ics.uci.edu/~mlearn/MLRepository.html.

zbiorze testowym. Wyniki zestawiono w tab. 1 (wyniki najlepsze zaznaczono pogrubioną czcionką).

Tabela 1. Błąd średniokwadratowy liczony na zbiorze testowym, dla różnych modeli regresji

Metoda	Friedman1	Friedman2	Friedman3	Boston	Servo
SVM	**4,16**	19206,9	0,070	15,21	21,38
LM	8,00	38762,8	0,104	25,27	64,02
PPR	7,58	22844,2	0,026	20,04	37,96
NNET	6,69	21705,1	0,047	17,45	33,18
RPART	11,17	31007,4	0,047	18,68	21,42
RFOREST	7,19	21114,9	0,027	**10,96**	**11,80**
BAGG	7,64	**17170,1**	**0,025**	12,72	16,86

Źródło: opracowanie własne.

Tylko w przypadku jednego zbioru danych metoda wektorów nośnych dała najmniejszy błąd predykcji. W pozostałych wypadkach zajmowała odpowiednio miejsca drugie, piąte i dwukrotnie trzecie. Wnioskujemy stąd, że w przypadku regresji o metodzie wektorów nośnych nie można powiedzieć, że daje na ogół najlepsze rezultaty. Zdecydowanie wyniki świadczą o tym, że predykcja za pomocą modelu otrzymanego metodą SVM jest znacznie lepsza niż klasycznego modelu liniowego, lecz jednocześnie konkurencyjne dla SVM wydają się metody zagregowane wykorzystujące drzewa regresyjne.

4. Podsumowanie

Dyskryminacja z wykorzystaniem metody wektorów nośnych na ogół daje mniejsze błędy klasyfikacji niż metody alternatywne. Otrzymane za jej pomocą modele są nieliniowe, przestrzeń hipotez jest bardzo liczna, ale jednocześnie w metodzie zaimplementowany jest mechanizm regularyzacji, przeciwdziałający nadmiernemu dopasowaniu modelu do danych ze zbioru uczącego, gdyż jest to częstą przyczyną wystąpienia dużych błędów predykcji.

Istnieje naturalny sposób przeformułowania metody wektorów nośnych tak, aby realizowała zadania regresji. Wiele pożądanych własności dyskryminacyjnej metody SVM przenosi się na jej odpowiednik regresyjny, lecz w porównaniu z innymi modelami regresji wydaje się tracić pozycję metody najdokładniejszej na rzecz metod wykorzystujących drzewa regresyjne. Ponadto metody zagregowanych drzew regresyjnych mają przewagę w większej prostocie i możliwościach interpretowania modeli składowych oraz możliwości pozyskiwania z nich wiedzy o badanym zjawisku. Ich algorytm jest prostszy i efektywniejszy pod względem numerycznym niż w metodzie SVM.

Na uwagę zasługuje fakt, potwierdzony również przyspieszonym rozwojem badań w tym obszarze, że na ogół najlepsze rezultaty otrzymujemy, gdy budujemy nie jeden model, lecz wiele modeli składowych, które agregujemy, otrzymując model końcowy.

Literatura

Cristianini N., Shawe-Taylor J., *An Introduction To Support Vector Machines (and Other Kernel–based Learning Methods)*, Cambridge University Press, 2000.

Friedman J., *Multivariate Adaptive Regression Splines*, „The Annals of Statistics" 1991, 19(1).

Gunn S.R., *Support Vector Machines for Classification and Regression*, Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer Verlag, N.Y. 2001.

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

Trzęsiok M., *Metoda wektorów nośnych w konstrukcji nieparametrycznych modeli regresji*, [w:] K. Jajuga, M. Walesiak (red.), Taksonomia 12, *Klasyfikacja i analiza danych*, AE, Wrocław 2005.

Vapnik V., *Statistical Learning Theory*, John Wiley & Sons, N.Y. 1998.

BENCHMARKING SUPPORT VECTOR MACHINES AGAINST OTHER REGRESSION METHODS

Summary

For nonlinear regression problem, support vector machines (SVM) map the input space into a high-dimensional feature space first, and then perform linear regression in the high-dimensional feature space. The paper presents the comparison of SVM and other regression methods (linear regression, Projection Pursuit Regression, Neural Networks, Regression Trees, Random Forest, Bagging) by the means of mean squared test set error.