PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Michał Trzęsiok
Akademia Ekonomiczna w Katowicach

METODA WEKTORÓW NOŚNYCH NA TLE INNYCH METOD WIELOWYMIAROWEJ ANALIZY DANYCH

1. Wstęp

Metoda wektorów nośnych (SVM – Support Vector Machines) należy do grupy dynamicznie rozwijających się metod wielowymiarowej analizy danych. Umożliwia nieliniową klasyfikację, nie dopuszczając przy tym do nadmiernego dopasowania funkcji klasyfikujących do danych ze zbioru uczącego (*overfitting*). Należy do grupy metod odpornych, tzn. pozwala, aby w zbiorze uczącym znajdowały się obserwacje błędnie sklasyfikowane. Pomimo wciąż rosnącego zainteresowania tą metodą, nadal brakuje w literaturze obszernego i kompleksowego porównania jakości klasyfikacji otrzymanej metodą wektorów nośnych z innymi metodami, zarówno klasycznymi, jak i tymi, których dynamiczny rozwój przebiega równolegle z rozwojem metody SVM.

Podstawowy cel niniejszego artykułu to przedstawienie wyników analizy porównawczej błędu klasyfikacji generowanego przez metodę wektorów nośnych oraz przez inne metody klasyfikacji wzorcowej (m.in. klasyczną liniową analizę dyskryminacyjną, metodę k najbliższych sąsiadów, drzewa klasyfikacyjne).

Porównanie metod realizowane będzie za pomocą symulacji komputerowych, przez zastosowanie metody wektorów nośnych i wybranych, alternatywnych metod dyskryminacji na tych samych zbiorach danych – zarówno rzeczywistych, jak i sztucznych – standardowo wykorzystywanych do badania własności metod wielowymiarowej analizy statystycznej.

2. Teoretyczne podstawy metody wektorów nośnych

Minimalizacja błędu klasyfikacji w zbiorze uczącym jest zazwyczaj podstawowym kryterium wyboru funkcji klasyfikującej. Taka postać kryterium wiąże się jednak z możliwością wyznaczenia bardzo złożonej funkcji dyskryminującej, z

czym związane jest występowanie nadmiernego dopasowania tej funkcji do danych ze zbioru uczącego i w konsekwencji do dużych błędów klasyfikacji w zbiorze testowym. Zachodzi więc potrzeba sformułowania kryterium kompromisowego, uwzględniającego nie tylko dobre dopasowanie funkcji dyskryminującej do obserwacji ze zbioru uczącego, ale także kontrolującego stopień złożoności otrzymywanej funkcji dyskryminującej. Jednym z takich kryteriów jest tzw. zasada minimalizacji ryzyka strukturalnego. W kolejnym podrozdziale przedstawione zostaną definicje pojęć ryzyka rzeczywistego oraz ryzyka empirycznego, niezbędne do sformułowania zasady minimalizacji ryzyka strukturalnego. W oparciu o tę zasadę skonstruowana jest metoda wektorów nośnych.

2.1. Ryzyko rzeczywiste i ryzyko empiryczne

Zakładamy, że w zadaniu dyskryminacji dany jest zbiór uczący postaci $D=\{(\mathbf{x}^1, y^1),\ldots,(\mathbf{x}^N, y^N)\}$, gdzie $\mathbf{x}^i \in \mathbf{R}^d$ oraz $y^i \in \{-1,1\}$ dla $i=1,\ldots,N$. Zakładamy ponadto, że zbiór uczący to realizacje niezależnych zmiennych losowych o jednakowym rozkładzie, określonym przez (nieznaną) funkcję gęstości:

$$p(\mathbf{x}, y) = p(\mathbf{x})p(y \mid \mathbf{x}). \tag{1}$$

Zadanie polega na znalezieniu funkcji

$$y = f(\mathbf{x}, \boldsymbol{\beta}), \tag{2}$$

gdzie $\boldsymbol{\beta}$ – wektor parametrów,

klasyfikującej obserwacje zgodnie z wartościami zmiennej Y na podstawie realizacji wektorowej zmiennej **X**, tzn. na przeszukaniu i wskazaniu jednego, najlepszego elementu w przestrzeni hipotez (tj. w zbiorze funkcji klasyfikujących obiekty ze zbioru uczącego).

Oznaczmy przez $H = \{f(\mathbf{x}, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbf{R}^d\}$ przestrzeń hipotez. Jakość klasyfikacji jest mierzona za pomocą funkcji straty $L(y, f(\mathbf{x}, \boldsymbol{\beta}))$, zdefiniowanej wzorem:

$$L(y, f(\mathbf{x}, \boldsymbol{\beta})) = \begin{cases} 1, & gdy\ f(\mathbf{x}, \boldsymbol{\beta}) \neq y, \\ 0, & gdy\ \ f(\mathbf{x}, \boldsymbol{\beta}) = y. \end{cases} \tag{3}$$

Przy tak przyjętych oznaczeniach najlepsza funkcja dyskryminująca to funkcja minimalizująca ryzyko rzeczywiste, tj. funkcjonał dany wzorem:

$$R(\boldsymbol{\beta}) = \int L(y, f(\mathbf{x}, \boldsymbol{\beta}))p(\mathbf{x}, y)d\mathbf{x}dy. \tag{4}$$

Minimalizacja ryzyka rzeczywistego nie może jednak stanowić kryterium wyboru funkcji dyskryminującej, gdyż wartości tego funkcjonału nie są znane, ponieważ nieznany jest rozkład $p(\mathbf{x}, y)$. Mając dany zbiór uczący D, możemy oszacować wartość $R(\boldsymbol{\beta})$. Minimalizacji podlegać zatem będzie ryzyko empiryczne:

$$R_{emp}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} L(y^i, f(\mathbf{x}^i, \boldsymbol{\beta})). \tag{5}$$

Znajdując $\hat{\boldsymbol{\beta}}$ minimalizujące ryzyko empiryczne, znajdziemy również odpowiadającą mu funkcję dyskryminującą $f(\mathbf{x},\hat{\boldsymbol{\beta}})$. Ze względu na to, że ryzyko empiryczne jest jedynie składową ryzyka rzeczywistego, funkcja $f(\mathbf{x},\hat{\boldsymbol{\beta}})$ nie musi minimalizować ryzyka rzeczywistego (4). Oznacza to, że najmniejszy z możliwych, a nawet zerowy błąd klasyfikacji w zbiorze uczącym nie gwarantuje uzyskania równie małych błędów w zbiorze testowym.

2.2. Zasada minimalizacji ryzyka strukturalnego

Koncepcja Vapnika [1998] poszukiwania najlepszej funkcji dyskryminującej polega na tym, aby w pierwszym etapie znaleźć jak najbliższe ograniczenie z góry wartości ryzyka rzeczywistego (4) przez wyrażenia, których wartości można wyliczyć przy danym zbiorze uczącym i danej przestrzeni hipotez, a następnie – skonstruować algorytm, który wskazywałby tę funkcję z przestrzeni hipotez, która minimalizuje wartość wyrażenia będącego ograniczeniem górnym ryzyka rzeczywistego. W latach siedemdziesiątych XX w. Vapnik i Chervonenkis udowodnili wiele twierdzeń pokazujących ograniczenia z góry wartości ryzyka empirycznego. Bodaj najważniejsze z nich zostało sformułowane za pomocą pojęcia wymiaru Vapnika–Chervonenkisa przestrzeni funkcji, który definiuje się jako największą liczbę h punktów $\mathbf{x}^1,\ldots,\mathbf{x}^h$, która może zostać rozdzielona na dwie klasy, na wszystkie 2^h możliwe sposoby, przez funkcje ze zbioru H. Jeśli dla każdego $n \in \mathbf{N}$ istnieje zbiór n wektorów, który można rozdzielić na dwie klasy za pomocą funkcji ze zbioru H, to przyjmujemy, że wymiar Vapnika–Chervonenkisa tej przestrzeni funkcji jest równy $+\infty$. Wymiar Vapnika–Chervonenkisa jest zatem pewną miarą złożoności potencjalnych funkcji dyskryminujących. Wspomniane wyżej twierdzenie ma następującą postać:

Twierdzenie. Niech dana będzie przestrzeń hipotez H o wymiarze Vapnika–Chervonenkisa równym h oraz zbiór uczący D o liczebności N. Wtedy z prawdopodobieństwem $1-\eta$ spełniona jest nierówność:

$$R(\boldsymbol{\beta}) \leq R_{emp}(\boldsymbol{\beta}) + \sqrt{\frac{h(\log\frac{2N}{h}+1)-\log\frac{\eta}{4}}{N}}. \tag{6}$$

Posługując się nierównością (6), można łatwo wyjaśnić fenomen nadmiernego dopasowania, tzn. tego, że bezwzględna minimalizacja ryzyka empirycznego (błędu klasyfikacji w zbiorze uczącym) może prowadzić do wygenerowania modelu dającego bardzo duże błędy w zbiorze testowym. Otóż w przypadku zbiorów liniowo nieseparowalnych niektóre metody w celu rozdzielenia klas przeszukują przestrzenie funkcji o bardzo złożonych postaciach analitycznych. I choć uzyskuje się przez to bardzo małą lub nawet zerową wartość ryzyka empirycznego, druga ze składowych, występująca po prawej stronie nierówności (6), ma bardzo dużą wartość, gdyż duży jest wymiar Vapnika–Chervonenkisa h, występujący w liczniku

wyrażenia. Można też zauważyć, że zgodnie z oczekiwaniami, gdy liczebność zbioru uczącego zmierza do nieskończoności, wartość ryzyka empirycznego zmierza do wartości ryzyka rzeczywistego.

Zasada minimalizacji ryzyka strukturalnego polega na utworzeniu łańcucha przestrzeni hipotez (ciągu zbiorów będących ze sobą w relacji inkluzji), na znalezieniu funkcji minimalizującej wartość ryzyka empirycznego w każdej z przestrzeni tworzącej łańcuch oraz na wyborze spośród nich tej funkcji, która minimalizuje wartość wyrażenia występującego po prawej stronie nierówności (6) (funkcji minimalizującej ograniczenie górne ryzyka rzeczywistego).

Algorytm metody wektorów nośnych skonstruowany jest tak, aby realizował zasadę minimalizacji ryzyka strukturalnego. Metoda wykorzystuje hiperpłaszczyzny i związane z nimi pewne funkcje liniowe (funkcje o prostej postaci analitycznej) do rozdzielania klas w nowej przestrzeni cech. Związane jest to z tym, że rodzina funkcji liniowych ma niewielki wymiar Vapnika–Chervonenkisa, co zgodnie z wcześniejszymi uwagami, jest cechą pożądaną.

3. Porównanie metod

Wobec przedstawionych w rozdziale drugim formalnych podstaw teoretycznych (matematycznych) metody wektorów nośnych zachodzi naturalna potrzeba empirycznego sprawdzenia jakości modeli klasyfikacji otrzymanych metodą wektorów nośnych, w szczególności zaś porównania jakości tych modeli z jakością klasyfikacji otrzymanej przez zastosowanie innych znanych metod wielowymiarowej analizy statystycznej. Jako miarę jakości klasyfikacji wybrano błąd klasyfikacji liczony w zbiorze testowym, stanowiącym 33% wszystkich obserwacji badanego zbioru.

3.1. Porównywane wybrane metody dyskryminacji

W przeprowadzonej analizie zastosowano na tych samych zbiorach danych sześć metod dyskryminacji. Należą do nich:

1) SVM – metoda wektorów nośnych,
2) LDA – liniowa analiza dyskryminacyjna,
3) KNN – metoda k najbliższych sąsiadów,
4) NNET – sieć neuronowa,
5) RPART – drzewa klasyfikacyjne,
6) RFOREST – zagregowane drzewa klasyfikacyjne Breimana.

Symbole, którymi oznaczone zostały poszczególne metody, zostały zaczerpnięte z nazw funkcji realizujących te metody w pakiecie statystycznym R, który wraz z dodatkowymi bibliotekami został wykorzystany do przeprowadzenia analizy.

Większość badanych metod wymaga od użytkownika ustalenia wartości pewnych parametrów. Do porównania wybierany był ten model (taki układ wartości parametrów), który dawał najmniejszy błąd klasyfikacji, liczony zazwyczaj metodą sprawdzania krzyżowego z podziałem zbioru uczącego na 10 części. Zatem w metodzie wektorów nośnych użyto wielomianowej funkcji jądrowej, zmieniając stopień wielomianu od 2 do 5 (por. [Trzęsiok 2004]), wartość parametru C od 10^{-2} do 10^{2}, wybierając model minimalizujący błąd klasyfikacji liczony metodą sprawdzania krzyżowego. W metodzie k najbliższych sąsiadów wartość parametru k zmieniała się od 1 do 10. Ze względu na możliwości użytego oprogramowania zastosowano najbardziej popularną architekturę sieci neuronowej z jedną ukrytą warstwą, liczbą obserwacji w warstwie ukrytej zmieniającą się od 1 do $\ln(N)$, stosując jak poprzednio sprawdzanie krzyżowe do wyboru sieci o najmniejszym błędzie klasyfikacji. W przypadku drzew klasyfikacyjnych wymaganą minimalną liczbę obserwacji w węźle, aby nastąpił dalszy podział, ustalano na poziomie od 3 do 10, kryterium zaś minimalnej poprawy jakości modelu (co jest związane z przycinaniem drzewa) – na poziomie od 1 do 3%. Tu również stosowano sprawdzanie krzyżowe. W metodzie zagregowanych drzew klasyfikacyjnych 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 (por. [Rozmus 2004]), minimalną zaś liczbę obserwacji w liściu: 1, 5, 10.

3.2. Zbiory danych

Wybrane metody dyskryminacji zastosowano na sześciu zbiorach danych, standardowo wykorzystywanych do badania własności metod klasyfikacji. Pierwsze cztery z nich, to zbiory danych rzeczywistych, zaczerpnięte z bazy zlokalizowanej w Uniwersytecie Kalifornijskim[1]. Kolejne dwa to zbiory sztuczne wygenerowane za pomocą biblioteki `mlbench` pakietu statystycznego R (por. [Leisch 2004]). Zbiory zostały tak dobrane, aby różniły się m.in. pod względem liczebności, liczby zmiennych, liczby klas. Te wybrane charakterystyki analizowanych zbiorów przedstawione są w tab. 1.

Tabela 1. Zbiory danych wykorzystane do analizy

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.

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

3.3. Wyniki analizy

Procedura badawcza przebiegała według następującego schematu:

1. Losowy podział zbioru danych na zbiór uczący D i testowy T w proporcji 2:1.
2. Wyznaczenie funkcji dyskryminującej zbiór uczący dla różnych wartości parametrów metody.
3. Wybór układu parametrów minimalizującego błąd klasyfikacji, liczony metodą sprawdzania krzyżowego.
4. Wyznaczenie błędu klasyfikacji na zbiorze testowym.
5. Zestawienie i porównanie otrzymanych wyników.

Otrzymane błędy klasyfikacji w procentach dla różnych metod dyskryminacji, wyliczone na zbiorach testowych przedstawia tab. 2.

Tabela 2. Błędy klasyfikacji dla różnych metod dyskryminacji (w %)

Zbiory	SVM	LDA	KNN	NNET	RPART	RFOREST
P.I.Diabetes	**21,0**	21,8	28,6	29,8	25,6	22,5
Sonar	**11,3**	26,8	14,1	22,5	31,0	15,5
Satellite	11,1	16,4	11,8	15,2	19,2	**9,9**
Glass	34,3	34,3	32,9	39,7	31,5	**24,7**
Circle	**2,9**	51,0	55,9	5,2	7,8	6,9
Spirals	**3,7**	46,3	33,1	6,8	5,2	4,2

Źródło: opracowanie własne.

Pogrubioną czcionką zaznaczone zostały najmniejsze błędy klasyfikacji.

4. Podsumowanie

W przypadku czterech zbiorów najlepsze wyniki otrzymano dla metody wektorów nośnych. Tylko w jednym przypadku metoda wygenerowała błąd klasyfikacji zdecydowanie gorszy od najlepszego wyniku. Można zatem stwierdzić, że metoda wektorów nośnych jest jedną z najdokładniejszych metod klasyfikacji w tym sensie, że generuje na ogół najmniejszy błąd klasyfikacji na zbiorach testowych.

Literatura

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

Leisch F., Dimitriadou E. (2004), *The* `mlbench` *Package – a Collection for Artificial and Real–World Machine Learning Benchmarking Problems,* R package, Version 1.0-0, http://cran.R-project.org.

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

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

Vapnik V. (1998), *Statistical Learning Theory*, John Wiley & Sons, New York.

SUPPORT VECTOR MACHINES IN COMPARISON TO OTHER METHODS OF MULTIVARIATE STATISTICAL ANALYSIS

Summary

Support vector machines (SVM) belong to a group of rapidly developing methods of multivariate statistical analysis. The main goal of this article is to compare the SVMs to other classification methods (e.g. Linear Discriminant Analysis, *k*–Nearest Neighbour Classification, Classification Trees) in terms of classification error generated by the final model. The analysis was conducted on some real and artificial benchmarking data sets.