Michał Trzęsiok

ZARYS TEORETYCZNYCH PODSTAW METODY DYSKRYMINACJI WYKORZYSTUJĄCEJ WEKTORY NOŚNE

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 klasyfikującej o niewielkiej zdolności objaśniania (uogólnienia). W opracowaniu przedstawiono inne kryterium, tzw. zasadę minimalizacji ryzyka strukturalnego, która oprócz jakości dyskryminacji uwzględnia również stopień uogólnienia wyznaczanego modelu. Następnie przedstawiono zarys pewnej metody dyskryminacji, skonstruowanej na podstawie zasady minimalizacji ryzyka strukturalnego, zwanej metodą wektorów nośnych.

1. Ryzyko rzeczywiste i ryzyko empiryczne

Zakładamy, że w zadaniu dyskryminacji jest dany zbiór uczący w postaci $\left\{\left(\mathbf{x^1}, y^1\right),\ldots,\left(\mathbf{x^N}, y^N\right)\right\}$, gdzie $\mathbf{x^i} \in \mathbf{R}^d$ oraz $y^i \in \{0,1\}$ dla $i = 1,\ldots,N$. Zakładamy ponadto, że zbiór uczący to realizacje niezależnych zmiennych losowych o jednakowym rozkładzie określonym przez nieznaną funkcję gęstości:

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

Zadanie polega na znalezieniu reguły klasyfikującej obserwacje zgodnie z wartościami zmiennej binarnej Y na podstawie realizacji wektorowej zmiennej $\mathbf{X}$, tzn. na przeszukaniu i wskazaniu jednego, najlepszego elementu

w przestrzeni hipotez (w zbiorze funkcji klasyfikujących obiekty ze zbioru uczącego). Ponieważ $y^i \in \{0,1\}$, więc można przyjąć, że przestrzeń hipotez składa się z funkcji charakterystycznych podzbiorów przestrzeni $\mathbf{R}^d$.

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

$$L\left(y, f(\boldsymbol{x}, \boldsymbol{\alpha})\right) = \begin{cases} 1, & gdy\ f(\boldsymbol{x}, \boldsymbol{\alpha}) \neq y \\ 0, & gdy\ \ f(\boldsymbol{x}, \boldsymbol{\alpha}) = y \end{cases} \tag{1}$$

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

$$R(\boldsymbol{\alpha}) = \int L(y, f(\boldsymbol{x}, \boldsymbol{\alpha})) p(\boldsymbol{x}, y) d\boldsymbol{x} dy \tag{2}$$

Wartości tego funkcjonału są nieznane, jako że nieznany jest rozkład $p(\mathbf{x}, y)$. Możemy zatem jedynie oszacować wartości $R(\boldsymbol{\alpha})$ na podstawie zbioru uczącego. Minimalizacji będzie wtedy podlegać **ryzyko empiryczne**:

$$R_{emp}(\boldsymbol{\alpha}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i, \boldsymbol{\alpha})) \tag{3}$$

Znajdując $\boldsymbol{\alpha}^*$ minimalizujące ryzyko empiryczne, znajdziemy również odpowiadającą mu funkcję dyskryminującą $f(\mathbf{x}, \boldsymbol{\alpha}^*)$. Zauważmy jednak, że $\boldsymbol{\alpha}^*$ nie musi minimalizować ryzyka (2). Oznacza to, że najmniejszy z możliwych, a nawet zerowy błąd klasyfikacji w zbiorze uczącym, nie gwarantuje uzyskania równie małych błędów w zbiorze testowym. Żeby wybrana funkcja f dawała dobre uogólnienie modelu, żądamy, aby zmienna $R_{emp}(\boldsymbol{\alpha})$ była **zgodnym** estymatorem nieznanej wartości parametru $\min R(\boldsymbol{\alpha})$, tzn. aby spełnione były następujące warunki:

$$\begin{aligned} \lim_{N \to +\infty} R_{emp}(\boldsymbol{\alpha}^*) &= \min R(\boldsymbol{\alpha}) \\ \lim_{N \to +\infty} R(\boldsymbol{\alpha}^*) &= \min R(\boldsymbol{\alpha}) \end{aligned} \tag{4}$$

gdzie powyższa zbieżność jest rozumiana w sensie zbieżności względem prawdopodobieństwa. Własność zgodności estymatora $R_{emp}(\boldsymbol{\alpha})$ zapewnia to, że w zagadnieniu minimalizacji ryzyka empirycznego otrzymamy ciąg modeli, dla którego obydwa rodzaje ryzyka są zbieżne do nieznanej, minimalnej wartości funkcjonału (2).

Wobec powyższych uwag, kluczowym twierdzeniem w tej teorii jest twierdzenie podające warunek konieczny i wystarczający na to, aby estymator (3) był zgodny. Takie twierdzenie zostało udowodnione przez Vapnika i Chervonenkisa (1989).

Twierdzenie 1

Niech $H = \left\{ f(\mathbf{x}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathbf{R}^d \right\}$ będzie zbiorem funkcji spełniających warunek:

$$A \leq \int L(y, f(\boldsymbol{x}, \boldsymbol{\alpha})) p(\boldsymbol{x}, y) d\boldsymbol{x} dy \leq B \tag{5}$$

Estymator $R_{emp}(\boldsymbol{\alpha})$ jest zgodny wtedy i tylko wtedy, gdy (3) jest zbieżne jednostajnie względem prawdopodobieństwa do ryzyka $R(\boldsymbol{\alpha})$, tj. jeśli:

$$\underset{\varepsilon>0}{\forall} \lim_{n \to +\infty} P\left\{ \sup_{\alpha} \left(R(\boldsymbol{\alpha}) - R_{emp}(\boldsymbol{\alpha}) \right) > \varepsilon \right\} = 0 \tag{6}$$

2. Górne ograniczenia dla ryzyka

Zadanie znalezienia funkcji dyskryminującej, minimalizującej ryzyko rzeczywiste (2), jest transformowane w ten sposób, że z góry znajduje się jak najlepsze ograniczenie wartości funkcjonału (2), a następnie poszukuje się funkcji *f* minimalizującej to ograniczenie. W teorii Vapnika-Chervonenkisa dowodzi się prawdziwości pewnych ograniczeń dla ryzyka, które wskazują na zbiory funkcji spełniających warunek jednostajnej zbieżności. Zgodnie z koncepcją Vapnika, przeszukiwany zbiór hipotez (*H*) zostaje zacieśniony do zbioru funkcji, których **zdolność objaśniania**, czyli zdolność do bezbłędnego dyskryminowania zbiorów uczących, odpowiada wielkości zbioru uczącego. Jako miarę zdolności objaśniania zbioru funkcji *F* zaproponowano tzw. **wymiar Vapnika-Chervonenkisa (wymiar VC)**.

Definicja 1

Niech dany będzie zbiór funkcji charakterystycznych *H*. Wymiarem VC nazywamy największą liczbę *h* wektorów $\mathbf{z}^1, \ldots, \mathbf{z}^h$, która może zostać rozdzielona na dwie klasy, na wszystkie 2^h możliwe sposoby, przez funkcje ze zbioru *F*. Jeśli dla każdego $n \in \mathbf{N}$ istnieje zbiór *n* wektorów, który można rozdzielić na dwie klasy za pomocą funkcji ze zbioru *F*, to przyjmujemy, że wymiar VC tego zbioru funkcji jest równy $+\infty$ (Vapnik, Chervonenkis, 1968).

Na przykład jeśli rozważymy zbiór funkcji charakterystycznych półpłaszczyzn, wyznaczonych przez różne proste na płaszczyźnie $\mathbf{R}^2$, to wymiar VC tego zbioru funkcji jest równy trzy, gdyż każde dwa rozłączne podzbiory zbioru składającego się z trzech punktów można rozdzielić prostą. Natomiast łatwo można zaznaczyć na płaszczyźnie cztery punkty ułożone tak, że dwa z nich tworzą podzbiór, którego nie można oddzielić linią prostą od podzbioru złożonego z pozostałych dwóch punktów.

Korzystając z pojęcia wymiaru VC dla zbioru funkcji charakterystycznych, Vapnik i Chervonenkis (1974) wykazali, iż z prawdopodobieństwem $1-\eta$ spełnione są m.in. następujące nierówności:

$$R(\boldsymbol{\alpha}) \le R_{emp}(\boldsymbol{\alpha}) + \sqrt{\frac{h(\log\frac{2N}{h}+1)-\log\frac{\eta}{4}}{N}} \tag{7}$$

$$R(\boldsymbol{\alpha}^*) - \min R(\boldsymbol{\alpha}) \le \sqrt{\frac{-\log\frac{\eta}{2}}{2N}} + \frac{\varepsilon}{2}\left(1+\sqrt{1+\frac{4}{\varepsilon}}\right) \tag{8}$$

gdzie:

$$\varepsilon = \frac{h(\log\frac{2N}{h}+1)-\log\frac{\eta}{8}}{N}$$

Pierwsza z powyższych nierówności podaje górne ograniczenie różnicy między ryzykiem rzeczywistym a ryzykiem empirycznym dla ustalonego α (tzn. ustalonej funkcji dyskryminującej). Druga zaś przedstawia ograniczenie różnicy między minimalnym ryzykiem empirycznym a minimalnym ryzykiem rzeczywistym. Dokładniejsze oszacowania uzyskano dzięki wykorzystaniu innych miar zdolności objaśniania zbioru funkcji, jak np. **entropia Vapnika-Chervonenkisa**. Jednak ze względu na skomplikowaną definicję tej miary oraz duże kłopoty związane z wyznaczaniem jej wartości, podstawową rolę w tej teorii odgrywa wymiar VC.

3. Minimalizacja ograniczenia górnego dla ryzyka – zasada minimalizacji ryzyka strukturalnego

Podstawowym celem rozważanego zadania jest wskazanie optymalnej, ze względu na wartość ryzyka, funkcji klasyfikującej, należącej do pewnego zbioru tworzącego przestrzeń hipotez. Oznacza to jednoczesne minimalizowanie wyrażenia występującego po prawej stronie nierówności (7) ze względu na wartości obydwu antagonistycznie zachowujących się składowych. Zauważmy jednak, że pierwsza ze składowych – ryzyko empiryczne – zależy od jednej, wybranej funkcji dyskryminującej. Zaś druga składowa – związana z wymiarem VC – dotyczy całej rodziny funkcji. W związku z tym, minimalizując obydwie składowe jednocześnie, potraktujemy wymiar VC jako zmienną kontrolną.

Powyższe rozważania prowadzą do sformułowania następującej zasady wyznaczania optymalnej funkcji klasyfikującej:

Niech dany będzie zbiór funkcji charakterystycznych H tworzący dostatecznie dużą przestrzeń hipotez, w której została określona klasa zbiorów o strukturze łańcuchowej, tj. zbiory funkcji $H_i \subset H$ takie, że:

$$H_1 \subset H_2 \subset \ldots \subset H_n \subset \ldots \tag{9}$$

Przez h_i oznaczmy wymiar VC zbioru H_i. Załóżmy, że wartość h_i jest skończona dla każdego $i \in \mathbf{N}$. Wtedy prawdziwe są następujące nierówności:

$$h_1 \leq h_2 \leq \ldots \leq h_n \leq \ldots \tag{10}$$

Wykorzystując obserwacje ze zbioru uczącego, w każdym ze zbiorów H_i znajdujemy funkcję $f(\mathbf{x}, \boldsymbol{\alpha}^{*i})$ minimalizującą ryzyko empiryczne (3). Z tak otrzymanego ciągu funkcji $\left(f(\mathbf{x}, \boldsymbol{\alpha}^{*i})\right)_{i \in \mathbf{N}}$ wybieramy tę funkcję, która minimalizuje prawą stronę nierówności (7).

Powyższą zasadę nazywamy **zasadą minimalizacji ryzyka strukturalnego** (*Structural Risk Minimization* (SRM) *Inductive Principle*). Jest ona kompromisem między jakością dyskryminacji zbioru uczącego a złożonością funkcji klasyfikujących.

Istnieje też drugie, alternatywne podejście do zadania minimalizacji ryzyka, które również wykorzystuje strukturę łańcuchową zbiorów, zdefiniowaną w przestrzeni hipotez. Różnica polega na tym, że przedstawiona powyżej me-

toda ustala wymiar VC w poszczególnych zbiorach łańcucha i minimalizuje w nich ryzyko empiryczne, zaś alternatywę stanowi metoda ustalająca ryzyko empiryczne i minimalizująca wymiar VC w każdym ze zbiorów. Pierwsze podejście jest wykorzystywane w teorii sieci neuronowych, zaś drugie – w metodzie wektorów nośnych (*Support Vector Machines* (SVM)). Metoda wektorów nośnych w ogólnym zarysie oraz jej związek z zasadą minimalizacji ryzyka strukturalnego zostaną przedstawione w czwartej części opracowania.

Aby zasada minimalizacji ryzyka strukturalnego mogła być stosowana, struktura łańcucha zbiorów $H_1 \subset H_2 \subset \ldots \subset H_n \subset \ldots$ musi być taka, żeby można było obliczyć wymiar VC dla każdego zbioru H_i oraz aby możliwe było rozwiązanie zadania minimalizacji ryzyka empirycznego w każdym ze zbiorów H_i.

4. Zasada minimalizacji ryzyka strukturalnego w metodzie wektorów nośnych

Metoda wektorów nośnych polega na nieliniowym przekształceniu przestrzeni danych w przestrzeń o większym wymiarze, w której obserwacje są rozdzielane hiperpłaszczyznami o równaniu:

$$\boldsymbol{\alpha} \cdot k(\mathbf{x}) + \alpha_0 = 0 \tag{11}$$

gdzie $\boldsymbol{\alpha} \in \mathbf{R}^m, \alpha_0 \in \mathbf{R}, k : \mathbf{R}^d \to \mathbf{R}^m$ jest nieliniowym przekształceniem przestrzeni danych, zaś funkcja dyskryminująca ma postać:

$$\hat{G}(\mathbf{x}) = \operatorname{sign}\left[\hat{\boldsymbol{\alpha}} \cdot k(\mathbf{x}) + \hat{\alpha}_0\right] \tag{12}$$

Ze względu na nieliniowość transformacji przestrzeni danych, liniowemu rozdzieleniu danych w nowej przestrzeni cech odpowiada ich klasyfikacja opisana funkcjami nieliniowymi w przestrzeni pierwotnej (Hastie i in., 2001; Gunn, 1997).

W metodzie wektorów nośnych występuje parametr określający, jak duża część obserwacji ze zbioru uczącego może zostać błędnie sklasyfikowana. Zbadanie wartości tego parametru oznacza ustalenie poziomu ryzyka empirycznego (3).

Wyznaczana hiperpłaszczyzna rozdzielająca obserwacje należące do dwóch różnych klas jest optymalna, tzn. maksymalnie oddalona od elementów poszczególnych klas. Wymiar VC funkcji charakterystycznych, wyznaczonych przez hiperpłaszczyzny w przestrzeni $\mathbf{R}^n$, jest równy *n+1*. Jednak dla wektorów $\boldsymbol{\alpha}$ spełniających warunek:

$$\|\boldsymbol{\alpha}\| \leq c \tag{13}$$

zbiór funkcji $G(\mathbf{x}, \boldsymbol{\alpha}, \alpha_0) = \operatorname{sign}\left[\boldsymbol{\alpha} \cdot k(\mathbf{x}) + \alpha_0\right]$ ma wymiar VC ograniczony przez:

$$h \leq \min\left\{r^2 c, n\right\} + 1 \tag{14}$$

gdzie *r* to długość najmniejszego promienia hiperkuli zawierającej wszystkie obserwacje ze zbioru uczącego (Vapnik, 1995).

Rozważmy strukturę łańcuchową zbiorów funkcji klasyfikujących typu (12) wyznaczonych przez hiperpłaszczyzny rozdzielające o rosnących wartościach parametru *c*. W celu wyznaczenia hiperpłaszczyzny optymalnej, w metodzie wektorów nośnych, jest rozwiązywane zadanie programowania wypukłego minimalizacji wyrażenia $\|\boldsymbol{\alpha}\|$. Wobec nierówności (13) i (14) oznacza to minimalizację wymiaru VC rozważanego zbioru funkcji klasyfikujących. Przypomnijmy, że wcześniej została ustalona wartość ryzyka empirycznego. Zatem metoda wektorów nośnych jest zgodna z zasadą minimalizacji ryzyka strukturalnego i jest skonstruowana na podstawie drugiego podejścia do tego zagadnienia przedstawionego w części trzeciej niniejszego opracowania. Zauważyć można ponadto, że nieliniowa transformacja przestrzeni daje możliwość przeszukiwania dużych przestrzeni hipotez, co w znaczący sposób wpływa na jakość dyskryminacji.

5. Przykład ilustrujący metodę wektorów nośnych

Metoda wektorów nośnych zostanie zilustrowana na przykładzie zbioru „Spirals", sztucznie wygenerowanego za pomocą funkcji zawartych w pakiecie `mlbench`, który jest dodatkową biblioteką programu statystycznego R. Zbiór „Spirals", jak i cały pakiet `mlbench` zostały stworzone na potrzeby porównywania własności metod wielowymiarowej analizy statystycznej. Podstawę do wygenerowania zbioru stanowią dwie spirale (klasy) leżące w płaszczyznie $\mathbf{R}^2$,

przy czym jedną spiralę można otrzymać z drugiej przez obrót o 180^o względem wspólnego środka. Z analizowanego zbioru wylosowano 1000 punktów, których współrzędne zaburzono białym szumem (zmienną o rozkładzie $N(0; 0,1)$). Tak otrzymany zbiór podzielono losowo na dwie równoliczne części – zbiór uczący (używany w metodzie wektorów nośnych do wyznaczania funkcji dyskryminującej) i zbiór testowy (nieuczestniczący w procesie wyznaczania funkcji dyskryminującej). Posługując się metodą symulacyjną przeszukano pewien zakres parametrów metody wektorów nośnych. Do końcowego modelu wykorzystano układ wartości parametrów minimalizujący błąd klasyfikacji w zbiorze uczącym (ryzyko empiryczne). Funkcję dyskryminującą otrzymaną metodą wektorów nośnych wykorzystano następnie do dyskryminacji klas w zbiorze uczącym i testowym. Wyniki klasyfikacji oraz jej jakość w badanych dwóch zbiorach przedstawia tab. 1.

Tabela 1

Tabela kontyngencji rzeczywistych przynależności do klas i wskazań uzyskanych metodą wektorów nośnych dla zbioru uczącego (po lewej) i testowego (po prawej) oraz błędy klasyfikacji w obydwu przypadkach

	A_1	B_1
SVM(A_1)	238	13
SVM(B_1)	17	232
Błąd klasyfikacji		0,06

	A_2	B_2
SVM(A_2)	229	25
SVM(B_2)	16	230
Błąd klasyfikacji		0,082

Podsumowanie

W przedstawionym przykładzie błąd klasyfikacji w zbiorze testowym różni się nieznacznie od błędu klasyfikacji w zbiorze uczącym, czyli można oczekiwać, że nieznana wartość ryzyka rzeczywistego nie odbiega znacząco od wartości ryzyka empirycznego. Oznacza to, że otrzymana funkcja dyskryminująca nadaje się do klasyfikowania nowych obiektów ze zbioru rozpoznawanego (co jest celem dyskryminacji). Ponadto, zważywszy że punkty spirali nie są liniowo separowalne w przestrzeni pierwotnej, a klasy częściowo się nakładają, uzyskany ośmioprocentowy błąd klasyfikacji jest zadowalający. Zasada minimalizacji ryzyka strukturalnego pozwala otrzymywać modele o dużym stopniu uogólnienia, tj. poprawnie klasyfikujące obserwacje nie tylko ze zbioru uczącego, ale także ze zbioru rozpoznawanego.

Literatura

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

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

Vapnik V.: *The Nature of Statistical Learning Theory*. Springer-Verlag, New York 1995.

Vapnik V., Chervonenkis A.: *On the Uniform Convergence of Relative Frequencies of Events to their Probabilities*. „Doklady Akademii Nauk USSR" 1968, 181 (4).

Vapnik V., Chervonenkis A.: *Theory of Pattern Recognition* (in Russian). Nauka, Moscow 1974 (German translation: Wapnik W., Tschervonenkis A.: *Theorie der Zeichenerkennung*. Akademia, Berlin 1979).

Vapnik V., Chervonenkis A.: *The Necessary and Sufficient Conditions for Consistency of the Method of Empirical Risk Minimization* (in Russian). Yearbook of the Academy of Science of the USSR on Recognition, Classification and Forecasting. Nauka, Moscow 1989 (English transl.: Pattern Recog. And Image Analysis, 3, 1991).

AN OUTLINE OF THE THEORETICAL BASES OF THE DISCRIMINATION METHOD USING SUPPORT VECTOR MACHINES (SVM)

Summary

Usually, when the problem of finding the best classifier is considered, it is based on minimizing the error on the training data (Empirical Risk Minimization). But in order to have a model with good generalization ability, the concept of Structural Risk Minimization (SRM) principle has been introduced. It defines a trade off between the quality of the approximation of the given data and the complexity of the approximating function. The formulation of the Support Vector Machines (SVM) embodies SRM principle. The very short overview of the theory of SVM has been presented and as an illustration a numerical example has been given.