Michał Trzęsiok

ŁĄCZENIE RÓWNOLEGŁE MODELI KLASYFIKACJI OTRZYMANYCH METODĄ WEKTORÓW NOŚNYCH

Wprowadzenie

Konstruowanie funkcji klasyfikujących przez łączenie wielu modeli składowych stanowi główny nurt badań naukowych nad metodami klasyfikacji w ciągu ostatnich pięciu lat. Powodem tak dynamicznego rozwoju metod agregujących są ich dobre własności, gdyż klasyfikacja danych na podstawie modeli zagregowanych daje na ogół mniejsze błędy klasyfikacji niż którakolwiek pojedyncza funkcja dyskryminująca, będąca składową modelu zagregowanego. Narzędziem, które umożliwia wyjaśnienie przewagi modeli łączonych nad pojedynczymi, a także pozwala na porównywanie oraz kreowanie nowych metod łączenia modeli, jest analiza błędu klasyfikacji podlegającego dekompozycji na obciążenie, wariancję i szum.

Zakładamy, że w zadaniu dyskryminacji dany jest zbiór uczący w postaci $D = \left\{\left(\mathbf{x}^1, y^1\right),\ldots,\left(\mathbf{x}^N, y^N\right)\right\}$, gdzie $\mathbf{x}^i \in \mathbf{R}^d$ oraz y^i – realizacje zmiennej opisującej klasę obiektu. Modele składowe $f_1,\ldots,f_B$ są otrzymywane w wyniku zastosowania metody klasyfikacji do B prób uczących $D_1,\ldots,D_B$, wyodrębnionych z podlegającego dyskryminacji zbioru uczącego D.

Jedną z metod agregacji jest metoda łączenia równoległego modeli (*bagging*) zaproponowana przez Breimana [1], w której obserwacje do każdej z prób uczących (*bootstrap samples*) są losowane ze zwracaniem ze zbioru D, przy czym liczebność każdej próby jest równa liczebności zbioru D, równa N, a prawdopodobieństwo wejścia do próby dla każdej obserwacji jest w każdym losowaniu stałe, równe $\frac{1}{N}$. W kolejnym kroku są wyznaczane funkcje dyskryminujące $f_1,\ldots,f_B$, odpowiadające próbom uczącym $D_1,\ldots,D_B$ w ten sposób, że wartości parametrów modelu składowego f_j wybierane są tak, aby błąd klasyfikacji był najmniejszy. Zagregowany model końcowy f^* łączy wskazania modeli składowych zgodnie z zasadą majoryzacji:

$$f^*(\mathbf{x}) = \arg\max_k \left\{ \sum_{j=1}^{B} I(f_j(\mathbf{x}) = k) \right\} \tag{1}$$

Pojedyncze modele otrzymane metodą wektorów nośnych mają dobre własności statystyczne: są odporne na występowanie w zbiorze uczącym obserwacji błędnie sklasyfikowanych oraz charakteryzują się dużym stopniem uogólnienia wyznaczanej funkcji dyskryminującej. W opracowaniu przedstawiono próbę empirycznego zweryfikowania hipotezy, iż model klasyfikacji zbudowany na podstawie metody łączenia równoległego wielu modeli uzyskanych metodą wektorów nośnych, generuje mniejsze błędy klasyfikacji niż każdy z pojedynczych modeli składowych.

1. Metoda wektorów nośnych

Metoda wektorów nośnych (*SVM – Support Vector Machines*) polega na nieliniowym przekształceniu oryginalnej przestrzeni danych w przestrzeń unitarną $\mathbf{Z}$ o dużo większym wymiarze, w której obserwacje są rozdzielane hiperpłaszczyznami. Ze względu na nieliniowość transformacji przestrzeni danych, liniowemu rozdzieleniu danych w nowej przestrzeni cech odpowiada nieliniowa ich dyskryminacja w przestrzeni pierwotnej.

Oznaczmy przez φ nieliniową transformację przestrzeni danych:

$$\varphi : \mathbf{R}^d \to \mathbf{Z} \tag{2}$$

Dla dwóch klas zadanie dyskryminacji polega na wyznaczeniu optymalnej hiperpłaszczyzny:

$$\boldsymbol{\beta} \cdot \varphi(\mathbf{x}) + \beta_0 = 0 \tag{3}$$

rozdzielającej klasy zbioru uczącego $\left\{ \left(\varphi(\mathbf{x}^1), y^1\right), \ldots, \left(\varphi(\mathbf{x}^N), y^N\right) \right\}$, gdzie $\mathbf{x}^i \in \mathbf{R}^d$, $\varphi(\mathbf{x}^i) \in \mathbf{Z}$ oraz $y^i \in \{-1,1\}$ dla $i = 1,\ldots,N$. Rozważane zagadnienie można zapisać w postaci zadania optymalizacji wypukłej z kwadratową funkcją celu oraz liniowymi ograniczeniami nierównościowymi (zob. [6]):

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum\limits_{i=1}^{N} \xi_i \\ \xi_i \geq 0, \quad y^i\left(\boldsymbol{\beta} \cdot \varphi(\mathbf{x}^i) + \beta_0\right) \geq 1 - \xi_i, \quad i = 1,\ldots,N \end{cases} \tag{4}$$

Parametr metody $C \geq 0$, którego wartość ustala użytkownik, określa kompromis między dokładnością dopasowania funkcji dyskryminującej do danych ze zbioru uczącego a zdolnością uogólniania modelu, rozumianą jako poprawne klasyfikowanie nowych obiektów ze zbioru rozpoznawanego. Zmienne $\xi_i \geq 0$ zostały wprowadzone w celu umożliwienia błędnego klasyfikowania niektórych obserwacji ze zbioru uczącego, co zwiększa odporność metody na występowanie szumu w tym zbiorze.

Rozwiązań zadania (4) poszukujemy metodą mnożników Lagrange'a, przekształcając problem w postać dualną. Otrzymane rozwiązanie (optymalna hiperpłaszczyzna rozdzielająca klasy) definiuje funkcję dyskryminującą w postaci:

$$f(\mathbf{x}) = \operatorname{sign}\left[\sum_{i \in I_{SV}} \alpha_i y^i \varphi(\mathbf{x}^i) \cdot \varphi(\mathbf{x}) + \hat{\beta}_0\right] \tag{5}$$

która jest opisana tylko przez te wektory $\mathbf{x}^i$ ze zbioru uczącego, dla których odpowiadające im współczynniki Lagrange'a w rozwiązaniu zadania optymalizacyjnego (4) są większe od zera ($\alpha_i > 0$). Obserwacje te nazywamy wektorami nośnymi.

Postać funkcji transformującej φ nie musi być znana. Wystarczy bowiem postać iloczynu skalarnego $\mathrm{K}(\mathbf{u}, \mathbf{v}) = \varphi(\mathbf{u}) \cdot \varphi(\mathbf{v})$ w przestrzeni $\mathbf{Z}$. W metodzie wektorów nośnych najczęściej wykorzystuje się funkcje z rodziny funkcji jądrowych:

a) Gaussa: $\mathrm{K}(\mathbf{u}, \mathbf{v}) = \exp(-\gamma \|\mathbf{u} - \mathbf{v}\|^2)$,

b) wielomianową: $\mathrm{K}(\mathbf{u}, \mathbf{v}) = \gamma(\mathbf{u} \cdot \mathbf{v} + \delta)^d$, $d = 1,2,\ldots$,

c) sigmoidalną: $\mathrm{K}(\mathbf{u}, \mathbf{v}) = \tanh(\gamma \mathbf{u} \cdot \mathbf{v} + \delta)$,

d) liniową: $\mathrm{K}(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$.

Wobec powyższych uwag funkcję dyskryminującą można zapisać w postaci:

$$f(\mathbf{x}) = \operatorname{sign}\left[\sum_{i \in I_{SV}} \alpha_i y^i \mathrm{K}(\mathbf{x}^i, \mathbf{x}) + \hat{\beta}_0\right] \tag{6}$$

gdzie $I_{SV} = \{i \in \{1,\ldots,N\} : \alpha_i > 0\}$.

Zatem nieliniowe przekształcenie φ pierwotnej przestrzeni danych jest realizowane w algorytmie metody przez pewną funkcję jądrową, definiującą iloczyn skalarny w pewnej nowej przestrzeni cech. Jak wiadomo, zdefiniowanie iloczynu skalarnego w pewnej przestrzeni (zdefiniowanie przestrzeni unitarnej) implikuje możliwość liczenia normy wektorów w tej przestrzeni, odległości

między punktami oraz kątów między wektorami – podstawowe narzędzia do określania miar podobieństwa między obiektami. Wybór funkcji jądrowej definiującej iloczyn skalarny w pewnej przestrzeni, a także wybór wartości parametrów tej funkcji oznacza więc ustalenie miary podobieństwa obiektów ze zbioru uczącego i ma istotne znaczenie dla postaci oraz jakości otrzymanej klasyfikacji. Ponadto jakość dyskryminacji zależy również od wartości parametru C metody ($C \geq 0$). Jak zostało wykazane w [6], na ogół najmniejszy błąd klasyfikacji występuje przy zastosowaniu metody wektorów nośnych z wielomianową funkcją jądrową i stopniem wielomianu nieprzekraczającym 5.

2. Dekompozycja błędu klasyfikacji

W dalszych rozważaniach wykorzystamy dekompozycję błędu klasyfikacji dla modeli zagregowanych na obciążenie, wariancję i szum, pochodzącą od Domingosa [4].

Niech t będzie rzeczywistą klasą obiektu $\mathbf{x}$. Jeżeli oznaczymy przez $f(\mathbf{x})$ wskazane klasy wyznaczone przez metodę dyskryminacji dla obiektu $\mathbf{x}$, to błąd klasyfikacji takiego modelu będzie równy $E_t[L(t,f(\mathbf{x}))]$, gdzie:

$$L(t,f(\mathbf{x})) = \begin{cases} 0, & gdy \quad t = f(\mathbf{x}) \\ 1, & gdy \quad t \neq f(\mathbf{x}) \end{cases} \tag{7}$$

jest funkcją straty. Jednakże wskazanie klasy $f(\mathbf{x})$ nie jest funkcją deterministyczną, gdyż zależy od zbioru uczącego D_j, który został wykorzystany w metodzie do wyznaczenia funkcji dyskryminującej f, więc $f(\mathbf{x}) = f_{D_j}(\mathbf{x})$. Biorąc więc jednocześnie pod uwagę losowość wynikającą zarówno z wyboru zbioru uczącego, jak i z wyboru obiektu ($\mathbf{x}$, t) ze zbioru testowego, średni błąd klasyfikacji można zapisać jako:

$$E_{D_j}[E_t[L(t,f_{D_j}(\mathbf{x}))]] \tag{8}$$

Celem tej części opracowania jest przedstawienie analizy średniego błędu klasyfikacji () oraz jego dekompozycja na obciążenie, wariancję i szum.

Do zdefiniowania składowych błędu klasyfikacji potrzebne są pojęcia: „klasyfikator optymalny" oraz „klasyfikator główny".

Klasyfikatorem optymalnym będziemy nazywać taką funkcję dyskryminującą $y^*(\mathbf{x}) = f(\mathbf{x})$, która minimalizuje błąd klasyfikacji:

$$E_t[L(t,f(\mathbf{x}))] \tag{9}$$

Zatem dla każdego punktu x, klasyfikator optymalny można zapisać jako:

$$y^*(\mathbf{x}) = \arg\min_k E_t[L(t,k)] \tag{10}$$

Klasyfikator główny definiujemy jako funkcję dyskryminującą, minimalizującą dla każdego punktu x, wartość oczekiwaną funkcji straty po wszystkich próbach uczących $D_1,\ldots,D_B$:

$$y^m(\mathbf{x}) = \arg\min_k E_{D_j}[L(f_j(\mathbf{x}),k)] \tag{11}$$

Zatem klasyfikator główny wskazuje klasę dla obiektu x zgodnie z najczęściej występującym wskazaniem klasyfikatorów składowych $f_1,\ldots,f_B$.

Obciążenie w punkcie x definiujemy jako wartość funkcji straty dla klasyfikatora optymalnego i głównego:

$$B(\mathbf{x}) = L(y^*, y^m) \tag{12}$$

Wariancję w punkcie x definiujemy jako wartość oczekiwaną funkcji straty między klasyfikatorem głównym a klasyfikatorami składowymi:

$$V(\mathbf{x}) = E_{D_j}[L(f_j(\mathbf{x}), y^m)] \tag{13}$$

Wariancja pokazuje, jak bardzo wskazania klasyfikatorów składowych różnią się w zależności od zbioru uczącego D_j. Szum w punkcie x definiujemy jako:

$$N(\mathbf{x}) = E_t[L(t, y^*)] \tag{14}$$

Przy tak zdefiniowanych składowych (12), (13) i (14) błędu klasyfikacji, można przedstawić jego dekompozycję w postaci:

$$E_{D_j}[E_t[L(t, f_{D_j}(\mathbf{x}))]] = c_1 N(\mathbf{x}) + B(\mathbf{x}) + c_2 V(\mathbf{x}) \tag{15}$$

Powyższa dekompozycja nie jest addytywna, gdyż stała c_2 przyjmuje wartość ujemną dla obciążonych obserwacji x [4].

Metoda łączenia równoległego jest techniką redukującą przede wszystkim wariancję modelu. Ponieważ metoda wektorów nośnych jest skonstruowana tak, iż jej algorytm zawiera mechanizm redukcji wariancji, dlatego uzasadniona wydaje się potrzeba empirycznego sprawdzenia, czy model zagregowany, uzyskany metodą łączenia równoległego funkcji dyskryminujących wyznaczonych metodą wektorów nośnych, będzie się charakteryzować mniejszym błędem klasyfikacji niż każdy z modeli składowych.

W dalszej analizie porównamy średnie błędy klasyfikacji wyliczone dla zbioru testowego w przypadku:

a) najlepszego pojedynczego modelu (*SVM*),

b) modelu zagregowanego (AGR_{SVM}), otrzymanego w wyniku zastosowania metody łączenia równoległego z minimalizacją błędu klasyfikacji na próbach uczących D_j.

3. Procedura badawcza

Procedura badawcza obejmuje następujące kroki:

1. Zbiór danych jest dzielony na dwie części – zbiór uczący D i zbiór testowy S w stosunku 2:1. Zbiór testowy posłuży do wyznaczenia średnich błędów klasyfikacji dla pojedynczych modeli oraz modelu zagregowanego, w celu ich porównania.
2. Ze zbioru uczącego jest losowanych ze zwracaniem $B = 100$ prób uczących $D_1, \ldots, D_B$, przy czym liczebność każdej próby jest równa liczebności zbioru D, równa N, a prawdopodobieństwo inkluzji dla każdej obserwacji jest w każdym losowaniu stałe, równe $\frac{1}{N}$.
3. Dla każdej próby uczącej D_j oraz każdego układu parametrów (d, γ, C), gdzie d to stopień wielomianowej funkcji jądrowej, przebiega zakres od 2 do 4, $\gamma \in \{0{,}5; 5\}$ oraz $C \in \{0{,}1; 10; 100\}$, wyznaczana jest, metodą wektorów nośnych, funkcja dyskryminująca f_j (pojedynczy model *SVM*).
4. Spośród wszystkich wyznaczonych funkcji dyskryminujących wyodrębnione zostają:
   a) model, któremu odpowiada najmniejszy średni błąd klasyfikacji,
   b) grupa modeli $\left\{f_j^{\min Err}\right\}_{j=1}^{B}$ takich, że funkcja $f_j^{\min Err}$ charakteryzuje się najmniejszym błędem klasyfikacji na próbie uczącej D_j, wśród wszystkich funkcji dyskryminujących wyznaczonych dla tej próby.
5. Zbiór testowy S zostaje poddany dyskryminacji klasyfikatorami z punktu 4a) oraz 4b).
6. Wskazania klas zbioru funkcji z punktu 4b) zostają połączone w jeden model zagregowany, zgodnie z zasadą majoryzacji.
7. Na podstawie wskazań klas z punktu 6. zostają wyznaczone (dla modelu zagregowanego): klasyfikator główny, średni błąd klasyfikacji oraz obciążenie i wariancja.
8. Dla najlepszego modelu pojedynczego, zidentyfikowanego w punkcie 4a), zostaje wyznaczony średni błąd klasyfikacji, obciążenie i wariancja na zbiorze testowym S.
9. Obliczone błędy klasyfikacji oraz jego składowe, dla najlepszego modelu pojedynczego oraz modelu zagregowanego, przedstawia tab. 1.

Ponieważ dla zbiorów danych rzeczywistych wyznaczenie wartości szumu jest trudne (gdyż dysponując jedynie obserwacjami zaburzonymi szumem, nie znamy rzeczywistej przynależności danego obiektu do klas), będziemy rozważać dekompozycję średniego błędu klasyfikacji na dwie składowe – obciążenie i wariancję. Tym samym wartości obciążenia będą nieco zawyżone.

4. Zbiór danych i wyniki analizy

Metodę łączenia równoległego składowych modeli otrzymanych metodą wektorów nośnych zilustrowano na zbiorze danych *Vehicle*, standardowo wykorzystywanym do badania i porównywania własności metod wielowymiarowej analizy statystycznej. Zbiór ten zawiera 846 obserwacji charakteryzowanych przez 19 zmiennych, z których jedna opisuje klasę obiektu. Liczba klas jest równa 4. Zbiór *Vehicle* zawiera dane rzeczywiste i pochodzi z ogólnodostępnej bazy „UCI Repository Of Machine Learning Databases" zlokalizowanej na Uniwersytecie Kalifornijskim[1]. Do obliczeń wykorzystano program komputerowy napisany przez autora w języku pakietu statystycznego R z wykorzystaniem dodatkowej biblioteki `e1071`, zawierającej implementację metody wektorów nośnych.

Do zbioru danych *Vehicle* zastosowano procedurę opisaną powyżej.

Tabela 1

Średni błąd klasyfikacji i jego składowe dla najlepszego modelu pojedynczego oraz modelu zagregowanego, wyznaczone na zbiorze testowym zbioru danych *Vehicle* (zł)

Nazwa	Pojedynczy model *SVM*	Model zagregowany AGR_{SVM}
Średni błąd klasyfikacji	0,19	0,17
Obciążenie	0,12	0,11
Wariancja	0,13	0,12

Tabela 1 przedstawia średnie błędy klasyfikacji wyliczone na zbiorze testowym dla najlepszego pojedynczego modelu (*SVM*) oraz modelu zagregowanego (AGR_{SVM}), otrzymanego w wyniku zastosowania metody łączenia równoległego z minimalizacją błędu klasyfikacji na próbach uczących D_j.

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

Dla analizowanego zbioru danych łączenie wielu modeli składowych w jeden końcowy model zagregowany polepszyło jakość klasyfikacji obserwacji ze zbioru testowego – błąd klasyfikacji modelu zagregowanego jest mniejszy od tegoż błędu dla najlepszego z pojedynczych modeli składowych. Łączenie modeli spowodowało zmniejszenie zarówno wariancji, jak i obciążenia.

Podsumowanie

Koncepcja Breimana polepszania jakości dyskryminacji przez metody łączenia równoległego wielu modeli składowych została pozytywnie zweryfikowana metodami empirycznymi, w przypadku zastosowania do konstrukcji modeli składowych metody wektorów nośnych. Powyższy wniosek jest zgodny z wynikami badań nad metodą łączenia równoległego modeli składowych w postaci drzew klasyfikacyjnych. Jednak polepszenie jakości klasyfikacji nie było znaczne. Wydaje się, iż lepsze rezultaty będzie można osiągnąć, jeśli wykorzysta się zmodyfikowaną metodę agregacji, a mianowicie adaptacyjną metodę łączenia równoległego (*adaptive bagging*). Modyfikacja ta polega na tym, aby przy wyborze funkcji składowych nie kierować się minimalizacją błędu klasyfikacji na danej próbie uczącej, lecz na pewnej próbie testowej, zawierającej obiekty, które nie uczestniczyły w procesie wyznaczania zbioru klasyfikatorów. Do stworzenia prób testowych można wykorzystać konsekwencje losowania obiektów do próby ze zwracaniem, polegające na tym, że w każdej próbie uczącej D_j znajduje się około 63,2% obserwacji ze zbioru uczącego D. Pozostałe 36,8% obserwacji można łatwo zidentyfikować i utworzyć z nich próbę testową T_j (*out-of-bag*), $T_j = D \setminus D_j$, odpowiadającą próbie uczącej D_j.

Literatura

1. Breiman L.: *Bagging Predictors*. „Machine Learning" 1996, nr 24.
2. Cristianini N., Shawe-Taylor J.: *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. Cambridge University Press, Cambridge 2000.
3. Dietterich T.G., Valentini G.: *Bias-Variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods*. „Journal of Machine Learning Research" 2000.
4. Domingos P.: *A Unified Bias-Variance Decomposition and Its Applications*. Proceedings of the Seventeenth International Conference on Machine Learning, Stanford 2000.

5. Smola A., Schölkopf B.: *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge 2002.
6. Trzęsiok M.: *Analiza wybranych własności metody dyskryminacji wykorzystującej wektory nośne.* W: *Postępy ekonometrii.* Red. A.S. Barczak. AE, Katowice 2004.
7. Vapnik V.: *Statistical Learning Theory.* John Wiley & Sons, New York 1998.

PARALLEL UNIFICATION OF THE CLASSIFICATION MODELS OBTAINED THROUGH SUPPORT VECTORS' METHOD

Summary

The paper presents a unified bias-variance decomposition of zero-one loss and its application to ensemble method using Support Vector Machines. We have used Breiman's bagging technique to aggregate base learners trained on the repeated bootstrap samples. Then, we present a numerical experiment to compare bagged ensemble of SVMs versus single SVMs.