

Jakość rozpoznawania klasyfikatorów po przejściu przez
procedurę *baggingu* z jakością klasyfikatorów pierwotnych

Karol Kozłowski
Karol Nikšcin

29 listopada 2007

Spis treści

0.1	Wprowadzenie	2
0.2	Bootstrap i rodzina bagging	2
0.3	Metoda walidacji krzyżowej	3
0.4	Założenia projektowe	3
0.5	Plan eksperymentu	3

0.1 Wprowadzenie

Bagging (*Bootstrap Aggregating*) jest jedną z pierwszych rodzin klasyfikatorów, zaproponowaną przez Breimana w 1994 r. Metoda ta często pozwala poprawić klasyfikację oraz modele regresyjne pod względem stabilności i dokładności. Dodatkowo zastosowanie metody bagging prowadzi do redukcji wariancji pojedynczych klasyfikatorów użytych do konstrukcji rodziny.

Bagging jest jedną z najbardziej efektywnych metod perturbowania i kombinowania. Dokonuje on wielokrotnych perturbacji zbioru treningowego i generuje odpowiadające im klasyfikatory, a następnie kombinuje je za pomocą prostego głosowania.

Założmy, że dany jest zbiór treningowy T posiadający N elementów. Dla każdego przypadku ze zbioru treningowego ustalone zostaje prawdopodobieństwo $p = \frac{1}{N}$, przy użyciu, którego dokonujemy N krotnego próbkowania ciągu treningowego z podstawieniem (bootstrap), tworząc zbiór treningowy T_b . W wyniku tego próbkowania niektóre przypadki ze zbioru T mogą się nie pojawić w zbiorze T_b , niektóre natomiast mogą pojawić się wielokrotnie. Powtarzanie tej procedury prowadzi do powstania sekwencji niezależnych zbiorów treningowych, które służą do tworzenia różnych klasyfikatorów przy zastosowaniu tego samego algorytmu tworzącego klasyfikator. Jak podaje L. Breiman w pracy [1], gdy klasyfikator jest dobry, pojedynczy klasyfikator może być daleki od optymalnego, natomiast kombinacje wielu dają klasyfikator bliski optymalnemu i stabilny. Niestety w przypadku słabych klasyfikatorów w wyniku kombinacji można otrzymać klasyfikator gorszy. [3]

0.2 Bootstrap i rodzina bagging

Założmy, że dany mamy zbiór uczący $\mathcal{L} \in \mathbb{L}$ posiadający n elementów oraz klasyfikator $d : \mathbb{X} \times \mathbb{L} \rightarrow \{1, 2, \dots, g\}$. Na podstawie zbioru \mathcal{L} tworzymy K psedoprób $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_K$. Każda psedopróba \mathcal{L}_k powstaje w wyniku wylosowania ze zwracaniem n -elementów z wyjściowego zbioru uczącego \mathcal{L} . Zakładamy przy tym, że wylosowanie każdego spośród n elementów jest równoprawdopodobne. Taki sposób generowania pseudoprób nazywamy metodą bootstrap. Rozważmy K reguł $d(\cdot, \mathcal{L}_k) : \mathbb{X} \rightarrow \{1, 2, \dots, g\}$. Mówimy, że reguła $d(\cdot, \mathcal{L}_k)$ jest k -tą wersją klasyfikatora d .

Definicja 1 Rodziną klasyfikatorów nazywamy dowolną rodzinę reguł (hipotez)

$$\mathcal{D} = \left\{ d_k : \mathbb{X} \rightarrow \{1, 2, \dots, g\} \right\}_{k=1,2,\dots,K}$$

gdzie $K \geq 2$

Zgodnie z definicją 1 rodzina:

$$\mathcal{B} = \left\{ d(\cdot, \mathcal{L}_k) : \mathbb{X} \rightarrow \{1, 2, \dots, g\} \right\}_{k=1,2,\dots,K}$$

jest rodziną klasyfikatorów. Wykorzystując regułę głosowania otrzymujemy klasyfikator $d_{\mathcal{B}}$ generowany rodziną \mathcal{B} . Klasyfikator $d_{\mathcal{B}}$ nazywamy klasyfikatorem otrzymanym metodą bagging (algorytm 1). [4]

Algorytm 1 Metoda Bagging

Dane Wejściowe: \mathcal{L} - próba ucząca, K - liczba iteracji

for $k = 1$ **to** K **do**

(1) Z próby uczącej \mathcal{L} wygeneruj pseudopróbkę \mathcal{L}_k metodą bootstrap

(2) Skonstruuj regułę decyzyjną $d(\cdot, \mathcal{L}_k)$

end for

Wyjście: Zlicz liczbę głosów $N_j(x)$, a następnie oblicz $d_{\mathcal{B}}(x) = \arg \max_j N_j(x)$

0.3 Metoda walidacji krzyżowej

pgflastimage

Metoda walidacji krzyżowej zwana również kroswalidacją to jedna z najwydajniejszych metod na ominięcie zjawiska overfittingu. Polega ona na tym, że ze zbior danych dzielimy na dwie części. Pierwsza zazwyczaj większa część wykorzystana będzie jako zbiór prób uczących. Druga natomiast posłuży do walidacji działania systemu.[2]

0.4 Założenia projektowe

Celem projektu jest ocena jakości klasyfikatorów zmodyfikowanych poprzez procedurę baggingu w porównaniu do jakości klasyfikatorów pierwotnych (tych, na których wykonywana była procedura baggingu). Jakość klasyfikacji oceniana będzie na podstawie metody *walidacji krzyżowej* oraz wariancji uzyskiwanych odpowiedzi. Badania wykonywane będą dla różnych długości ciągu uczącego. W badaniach wykorzystywane będą bazy danych z repozytorium UCI ¹

0.5 Plan eksperymentu

- Z repozytorium wybrane zostaną bazy danych.
- Baza zostanie podzielona na 2 części - uczącą i walidacyjną.
- Część ucząca zostanie poddana procedurze baggingu.
- Dane te zostaną dostarczone do systemu w celu nauki i późniejszej walidacji.

¹UCI Machine Learning Repository - <http://mllearn.ics.uci.edu/MLRepository.html> [5]

- Takiemu samemu procesowi zostaną poddane dane sprzed procedury baggingu (dla celów porównawczych).
- Powyższe czynności zostaną wykonane wielokrotnie dla takich samych danych (aby uśrednić wyniki pomiarów).
- Badanie zostanie wykonane dla różnych baz danych, różnych długości ciągów uczących jak i walidacyjnych.

Bibliografia

- [1] L. Breiman, *Bagging predictors*, Technical Report 420, Department of Statistics, University of California, CA, USA, September 1994.
- [2] Tom Mitchel, *Machine Learning*, McGraw Hill, 1997.
- [3] Rafał Adamczak *Zastosowanie sieci neuronowych do klasyfikacji danych doświadczalnych*. UMK, 2001
- [4] Iwona Głowacka *Metody łączenia klasyfikatorów w analizie dyskryminacyjnej* Warszawa, Wrzesień 2006
- [5] Asuncion, A & Newman, D.J. 2007. *UCI Machine Learning Repository* [<http://mllearn.ics.uci.edu/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.