Szymon SMAGA[1]

NADMIERNE DOPASOWANIE W DRZEWACH DECYZYJNYCH

Streszczenie

W pracy staramy się sprawdzić wpływ jaki ma dobór kryterium wyboru testu na nadmierne dopasowanie w drzewach decyzyjnych. Uważamy, że losowe kryterium doboru może okazać się nie gorsze od kryterium entropijnego. Nasze przypuszczenia potwierdzają wstępne badania wykonane dla trzech (niewielkich rozmiarów) zbiorach trenujących, co w naszej opinii zasługuje na dalsze eksperymenty.

Abstract

In this paper we try to check the influence of selection criteria for the test selection for excessive fit in decision trees. We believe that a random criteria selection may not be worse than the criteria of entropy. Our supposition is confirmed by preliminary tests performed for three training sets, which in our opinion deserves further experiments.

1 WSTĘP

Istnieje prawdopodobieństwo, że w zbiorze treningowym mogą występować przypadkowe regularności w danych. Nadmierne dopasowany klasyfikator w swoich regułach decyzyjnych może wykorzystywać te przypadkowe regularności, których nie ma w całej przestrzeni danych.

W pracy tej chcemy sprawdzić jaki wpływ ma dobór kryterium wyboru testu na nadmierne dopasowanie danych podczas indukcji drzew decyzyjnych. Jako kryterium doboru chcemy zastosować kryterium entropijne i losowe.

1.1 Entropijne kryterium doboru testu

Dobór kryterium testu ma ogromne znaczenie dla złożoności drzewa decyzyjnego. Jednym z proponowanych kryterium doboru testu w naszej pracy jest liczbowe określenie jakości testu poprzez wyliczanie entropii zbioru przykładów:

[1] Mgr Szymon Smaga jest wykładowcą Warszawskiej Wyższej Szkoły Informatyki.

$$E_{tr}(P) = \sum_{d \in C} -\frac{|P_{tr}^d|}{|P_{tr}|} \log \frac{|P_{tr}^d|}{|P_{tr}|},$$

a następnie wyliczenie dla nich średniej ważonej entropii dla poszczególnych wyników testu

$$E_t(P) = \sum_{r \in R_t} \frac{|P_{tr}|}{|P|} E_{tr}(P).$$

Takie podejście daje nam w rezultacie optymalne drzewo, którego konstrukcja nie jest zbyt kosztowna.

2 NADMIERNE DOPASOWANIE

W budowie drzewach decyzyjnych ważną rzeczą jest dostarczenie zestawu reguł na podstawie których system nauczy się wnioskowania, a dzięki temu będzie w stanie klasyfikować nowe dostarczone przypadki. Często zdarza się jednak, że podczas konstruowania drzewa następuje jego zbytni rozrost lub nadmierne dopasowanie do zbioru uczącego. Skutkiem takiej przypadłości jest niemożność poprawnego rozpatrywania dostarczonych nowych przykładów lub zbytnie obciążenie systemu uczącego.

2.1 Sprawdzenie błędu próbki

Istnieje wiele metod sprawdzenia nadmiernego dopasowania drzewa decyzyjnego do zbioru trenującego. Na potrzeby naszego doświadczenia został wybrany jeden ze sposobów polegający na wyliczeniu błędu próbki, a więc sprawdzeniu stopnia zgodności klasyfikatora hipotezy z klasyfikatorem zbioru uczącego.

$$e_P^c = \frac{|\{x \in P \mid h(x) \neq c(x)\}|}{P}$$

gdzie:

h (x) – klasyfikator hipotezy

c (x) – klasyfikator zbioru trenującego

P – zbiór walidacyjny

Podzbiór dla wyliczenia błędu przyjęło się określać jako różnicę ilości elementów zbioru trenującego i liczby elementów zbioru próby uczącej, wybranej poprzez losowe wygenerowanie ze zbioru trenującego z piątej jego części

$$P = T - T_1,$$

gdzie:

T – zbiór trenujący

T_1 – zbiór próby uczącej

Wiadomym jest, że czym mniejszy błąd próbki lub w granicznych przypadkach jeżeli zbiór próby uczącej jest spójny ze zbiorem trenującym (błąd próbki jest równy 0) tym większe prawdopodobieństwo nadmiernego dopasowanie drzewa, które może odzwierciedlać przypadkowe regularności w danych.

2.2 Wyniki eksperymentu

Zadanie sprawdzenia wpływu sposobu doboru testu na stopień nadmiernego dopasowania drzew decyzyjnych, zrealizowane zostało na trzech zbiorach trenujących tj. Golf, Car oraz Iris. Dwa pierwsze zbiory wzięte zostały z książki [1], gdzie posłużyły autorowi do zobrazowania zagadnień dotyczących systemów uczących się, trzeci zaś zbiór pobrany został z systemu RSES.

W ramach eksperymentu dla każdego ze zbiorów trenujących, wygenerowano losowo dziesięć podzbiorów z zbioru trenującego, a następnie obliczony został dla nich błąd próbki przy kryterium doboru testu: entropijnym, losowym. Wyniki przedstawione są poniżej w trzech tabelach

Tabela 1. Wyniki testu dla zbioru trenującego „Golf"

Nr losowania	Błąd Dlos	Błąd De
1	0,36	0,27
2	0,45	0,45
3	0,54	0,45
4	0,27	0,27
5	0,36	0,36
6	0,45	0,45
7	0,63	0,63
8	0,27	0,27
9	0,36	0,27
10	0,54	0,18
Średni błąd na zbiorze P	0,42	0,36

Tabela 2. Wyniki testu dla zbioru trenującego „Car"

Nr losowania	Błąd Dlos	Błąd De
1	0,22	0,22
2	0,22	0,33
3	0,11	0
4	0,55	0,22
5	0,33	0,33
6	0,22	0,22
7	0,33	0,22
8	0,22	0,22
9	0,22	0,33
10	0,44	0,22
Średni błąd na zbiorze P	0,29	0,23

Tabela 3. Wyniki testu dla zbioru trenującego „Iris"

Nr losowania	Błąd Dlos	Błąd De
1	0,27	0,18
2	0,27	0,27
3	0,36	0,27
4	0,44	0,27
5	0,36	0,36
6	0,36	0,36
7	0,58	0,27
8	0,44	0,27
9	0,27	0,36
10	0,36	0,27
Średni błąd na zbiorze P	0,34	0,26

Tabela 4. Średnie błędy próbki dla drzew decyzyjnych

Zbiór przykładów T	Średni błąd drzew Dlos na T	Średni błąd drzew De na T
Golf	0,42	0,36
Car	0,29	0,23
Iris	0,34	0,26

Wnioski z wyników eksperymentu można wyciągnąć następujące:

- błąd próbki dla drzew decyzyjnych z entropijnym doborem testu jest mniejszy od błędu próbki dla drzew decyzyjnych z losowym doborem testu,
- indukcja drzew decyzyjnych z entropijną klasyfikacją może więc odzwierciedlać przypadkowe zależności występujące w zbiorze próby uczącej,
- nie jesteśmy w stanie dokładnie ocenić błędu próbki dla kryterium losowego wyboru testu,
- nie jesteśmy w stanie określić, które z kryterium doboru testu daje większe prawdopodobieństwo dopasowywania się do przypadkowych regularności występujących w zbiorach trenujących.

3 ZAKOŃCZENIE

Na zakończenie należy powiedzieć, że wiarygodna ocena nadmiarowości podczas indukcji drzew decyzyjnych wymaga dodatkowych eksperymentów na dużych zbiorach trenujących. Na obecną chwilę nie możemy jednoznacznie określić, które z kryteriów doboru testu dało by w rezultacie lepszy wynik.

Dodatkowym elementem nad jakim należało by się zastanowić to ujednolicenie wyników błędu próbki z losowym kryterium doboru testu. Podczas tej pracy dało się zauważyć, że dla tego sposobu doboru, istnieje wiele rozwiązań, nie zawsze zbliżonych do siebie, a wręcz odwrotnie, mających duże rozbieżności. Jedną z proponowanych metod mogła by być wielokrotna budowa drzewa decyzyjnego dla zbioru próby uczącej i wyliczenie dla niej błędu, a następnie wynik należało by uśrednić.

Literatura

1. Paweł Cichosz, *Systemy uczące się*, Wydawnictwo Naukowo Techniczne Warszawa 2000, 2007
2. Jacek Koronacki, Jan Ćwik, *Statystyczne systemy uczące się*, Akademicka Oficyna Wydawnicza EXIT, 2008