Algorytm drzewa decyzyjnego C4.5(C5.0) w R

Indukcyjne metody analizy danych

Maksym Telepchuk

Kwiecień 2020

1 Drzewa decyzyjne

Drzewo decyzyjne jest grafem o strukturze drzewiastej, gdzie

1. każdy wierzchołek wewnętrzny reprezentuje podział według atrybutu
2. każdy łuk reprezentuje wynik podziału
3. każdy liść reprezentuje pojedynczą klasę

Algorytm drzewa decyzyjnego C4.5 rekurencyjnie dzieli zbiór treningowy na partycje do momentu, w którym każda partycja zawiera dane należące do jednej klasy, lub, gdy w ramach partycji dominują dane należące do jednej klasy.

Drzewo decyzyjne jest konstruowane w dwóch krokach. W pierwszym kroku drzewo decyzyjne jest tworzone z treningowej bazy danych. W drugim kroku następuje obcinanie drzewa (ang. pruning) w celu poprawy dokładności, interpretowalności i uniezależnienia się od efektu przetrenowania. Najczęściej używanym typem obcinania jest postpruning, w którym konstruowano pełne drzewo decyzyjne i usuwano z niego zawodne części.

W trakcie budowy drzewa decyzyjnego, wybierano taki atrybut i taki punkt podziału, określający wierzchołek wewnętrzny drzewa decyzyjnego, który „najlepiej" dzieli zbiór danych treningowych należących do tego wierzchołka. Do oceny jakości punktu podziału w C4.5 jest stosowany zysk informacyjny. Jako atrybut podziału wybierano atrybut o największym zysku informacyjnym.

$I(s_1, s_2, ...s_m)$ jest oczekiwaną ilością informacji niezbędna do zaklasyfikowania danego przykładu. Jest obliczana według wzoru:

$$I(s_1, s_2, ...s_m) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

gdzie s_i jest ilością przykładów o klasie c_i w danej partycji, p_i jest prawdopodobieństwem, że klasa i występuje w danej partycji.

Atrybut A posiadający v różnych wartości dzieli zbiór S na partycje $S_1, S_2, ..., S_v$, gdzie S_j zawiera przykłady ze zbioru S, których wartość atrybutu A wynosi a_j.

Entropię podziału zbioru S na partycje, według atrybutu A definiowano następująco:

$$E(S_1, S_2, ..., S_v) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + ... + s_{mj}}{|S|} I(s_{1j}, s_{2j}, ..., s_{mj})$$

gdzie s_{ij} oznacza liczbę przykładów z klasy C_i w partycji S_j.

Im mniejsza wartość entropii, tym większa "czystość" podziału zbioru S na partycje.

Zysk informacyjny, wynikający z podziału zbioru S na partycje według atrybutu A, definiowano następująco:

$$Gain = I(s_1, s_2, ..., s_m) - E(A)$$

Dla każdego atrybutu jest liczony zysk informacyjny. Drzewo jest dzielone według tego atrybutu, który ma największy zysk informacyjny. Jeśli wartości atrybutu są wartościami ciągłymi, podział odbywa się według przedziału, do którego należy ta wartość (mniejsze lub równe / większe).

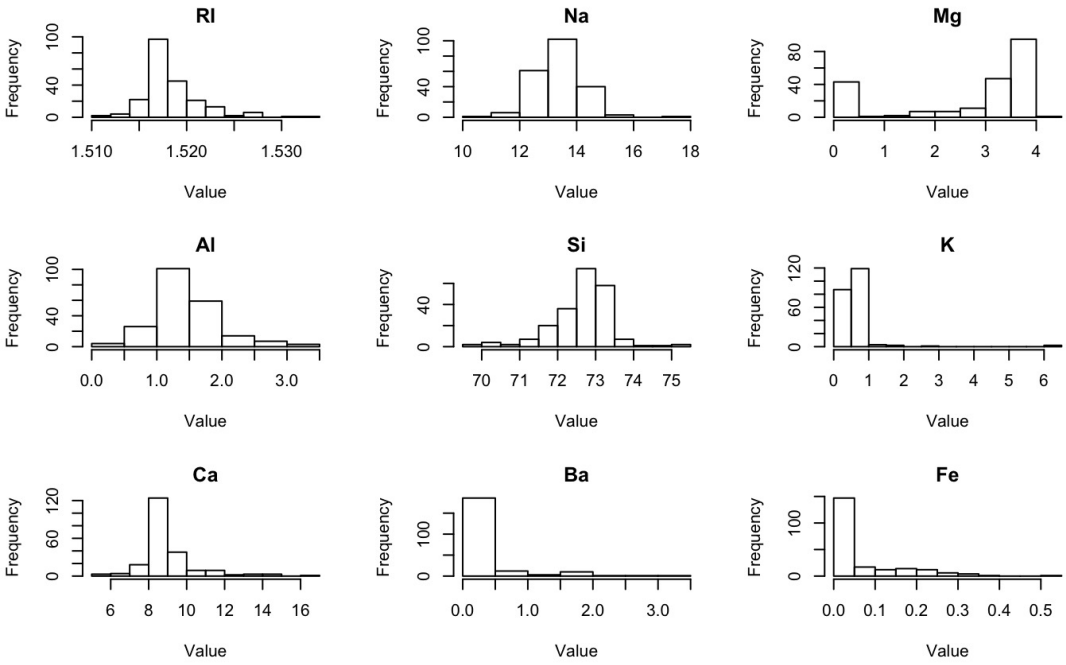
W tym badaniu stosowano rozszerzoną wersję algorytmu C4.5, nazywaną C5.0. Szczegóły rozszerzeń są w dużej mierze nieudokumentowane. W C5.0 jest opcja korzystania z tzw. macierzy kosztu, gdzie można niejednolicie karać model za niepoprawną klasyfikację. Ma również opcję boostingu oraz korzystania z selekcji cech (winnow). Ma nieco inny sposób pruningu oraz drzewa C5.0 są mniejsze od odpowiedników C4.5.

2 Dane

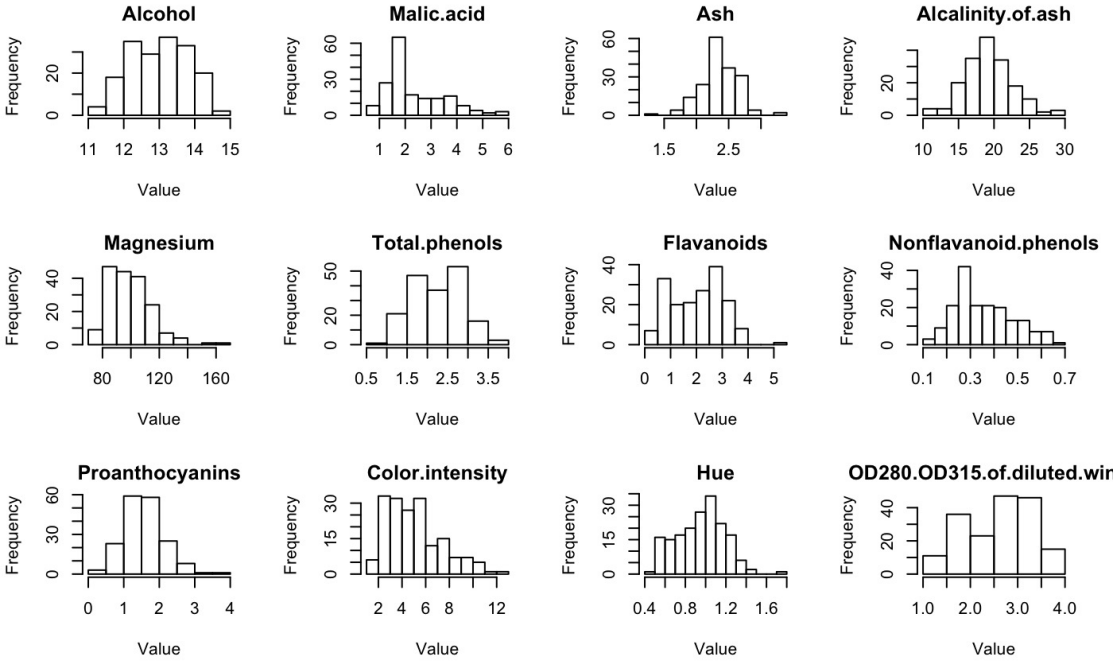
Dla analizy zostały wybrane następujące zbiory danych :

- glass - 6 typów szkła, skład z 9 chemicznych elementów (rys. 1)
- wine - 3 rodzaje wina, skład z 13 chemicznych elementów (rys. 2)
- seeds - 3 typy nasienia, 7 pomiarów nasienia (rys. 3)

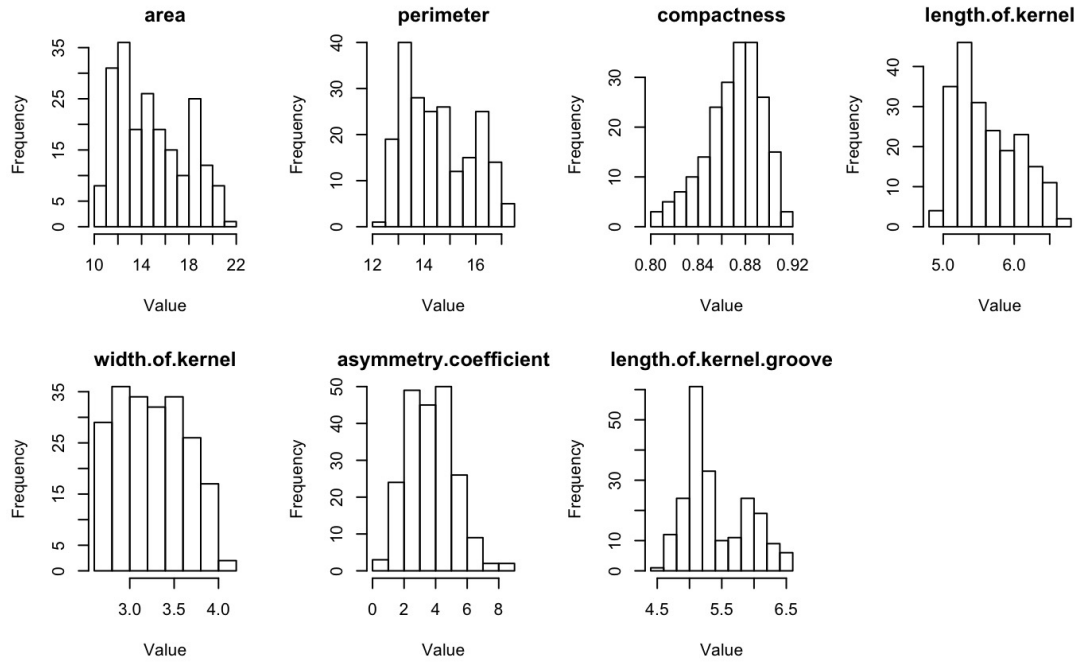
Rozkłady cech są pokazane na załączonych wykresach.



Rysunek 1: Cechy zbioru danych **glass**



Rysunek 2: Cechy zbioru danych **wine**



Rysunek 3: Cechy zbioru danych **seeds**

3 Walidacja modelu

Dla porównania efektywności algorytmu na różnych zbiorach danych, użyto k-fold Cross Validation w wersji zwykłej oraz stratyfikowanej. Walidację dokonywano n=10 razy oraz wynikiem jest średnia wartości. W walidacji zwykłej zostały wyliczone metryki *accuracy* i *f1 score* oraz w walidacji stratyfikowanej została filoczona metryka *accuracy*. Wyniki są podane w tablicy 1.

zbiór	k	accuracy	accuracy (stratified)	f1 score
glass	5	0.70	0.703	0.72
	10	0.693	0.71	0.714
wine	5	0.928	0.93	0.942
	10	0.929	0.93	0.955
seeds	5	0.910	0.92	0.909
	10	0.906	0.918	0.848

Tablica 1: Wyniki krosswalidacji

4 Analiza parametrów C5.0

Do przeanalizowania działania algorytmu zostały wybrane 3 argumenty do analizy:

1. winnow : wartość logiczna (FALSE)
2. noGlobalPruning : wartość logiczna (FALSE)
3. CF : wartość numeryczna (0.25)

Winnowing to etap selekcji cech przeprowadzany przed modelowaniem. Zestaw danych jest losowo dzielony na pół, oraz model jest uczony na pierwszej połowie. Każdy predyktor jest kolejno usuwany i określany jest wpływ na wydajność modelu (z wykorzystaniem drugiej połowy losowego podziału). Predyktory są oflagowane, jeśli ich usunięcie nie zwiększa poziomu błędu. Ostateczny model jest dopasowywany do wszystkich próbek zestawu treningowego, używając tylko nieflagowanych predyktorów.

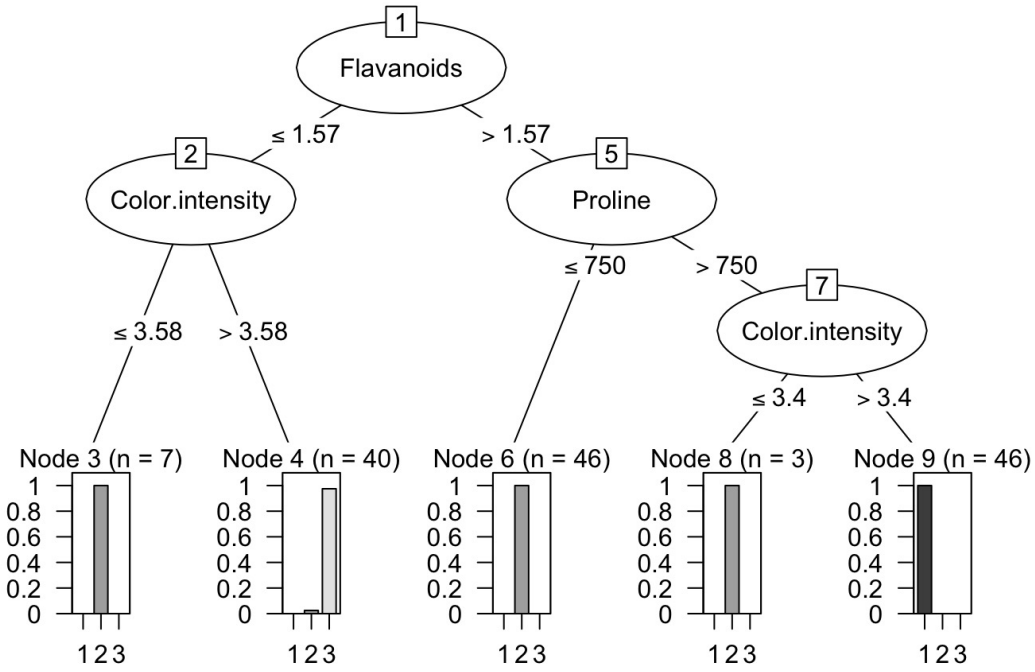
Można również kontrolować czy będzie wykonywane krok przycinania w celu uproszczenia drzewa.

CF (ang. confidence factor) - kontroluje stopień przycinania: im wyższa wartość, tym mniej gałęzi będzie przycinanych; im niższa wartość, tym więcej gałęzi zostanie przyciętych.

Trenowanie modelu zostało wykonywane n=30 razy dla różnych zestawów wartości atrybutów. Wynikami są średnie wartości metryk. Wyniki są przedstawione w tablicy 2, gdzie "-" oznacza wartość domyślną.

zbiór	train ratio	winnow	noGlobalPrunning	CF	accuracy	f1 score
glass	0.8	-	-	-	0.702	0.705
glass	0.8	TRUE	-	-	0.691	0.735
glass	0.8	-	TRUE	-	0.668	0.679
glass	0.7	-	-	0.75	0.666	0.691
wine	0.8	-	-	-	0.922	0.938
wine	0.8	-	TRUE	-	0.922	0.946
wine	0.5	-	-	-	0.855	0.886
wine	0.5	TRUE	-	-	0.894	0.919
seeds	0.8	-	-	-	0.895	0.841
seeds	0.8	-	TRUE	-	0.907	0.856
seeds	0.8	-	-	0.75	0.917	0.863
seeds	0.8	-	-	0.1	0.924	0.876

Tablica 2: Wyniki trenowania C5.0



Rysunek 4: Drzewo decyzyjne dla zbioru **wine**

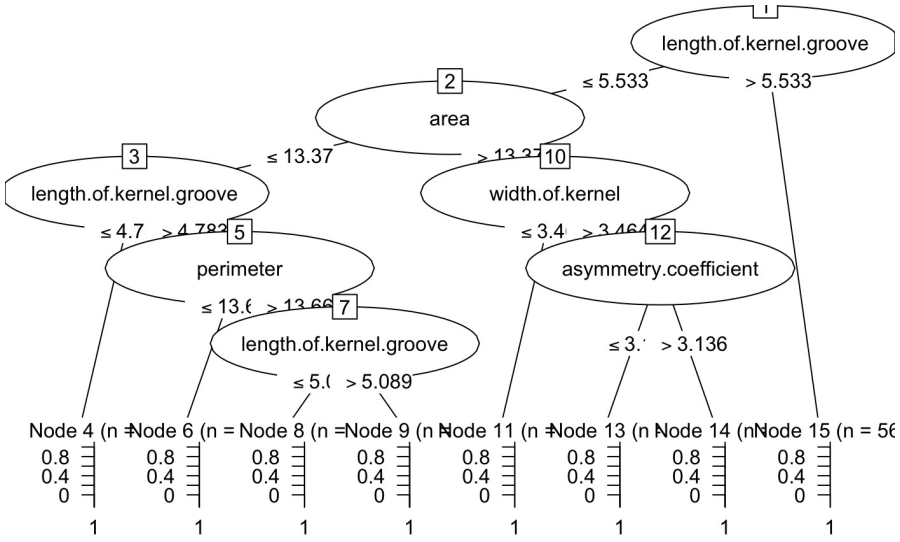
5 Wizualizacja

Zaletą drzew decyzyjnych jest ich wysoka interpretowalność. Na rysunkach 4, 5, 6 przedstawiona wizualizacja drzew decyzyjnych dla każdego ze zbiorów. W zbiorze **glass** najbardziej użytecznymi cechami byli Mg (100% użycia) i RI (72.51 %). W zbiorze **wine** byli to flavanoids (100%), color intensity (67.61%) i proline (66.90 %). Dla zbioru **seeds** byli to length of kernel groove (100 %) i area (66.67%).

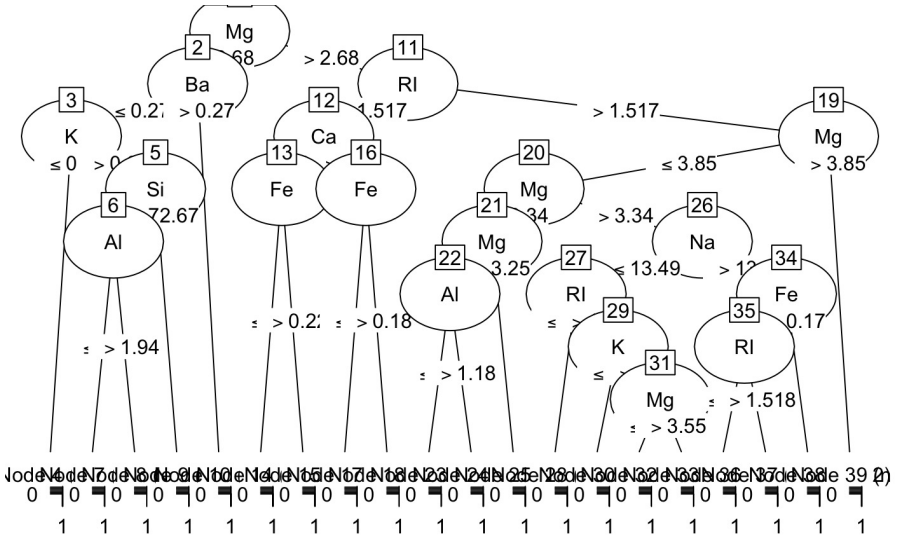
6 Wnioski

Z wyników badań można wywnioskować, że prunnig pozwala zapobiec przetrenowaniu. Badania wykazały, że wyłączenie przecinania pogarsza wyniki końcowe. Selekcja cech może polepszyć wyniki w przypadku małego zbioru treningowego. Współczynnik pewności CF kontroluje obcinanie oraz jest ważnym hyperparametrem dla modelu drzewa decyzyjnego.

W porównaniu do wyników Naiwnego Bayesa, zauważalne istotne różnice w zachowywaniu. Tak dla zbioru **wine** klasyfikator Naiwnego Bayesa daje o wiele lepsze wyniki (accuracy = 0.99, f1 score = 0.99). Dla zbioru **glass** Naiwny Bayesa okazał się bardzo nie skutecznym w porównaniu do modelu drzewa decyzyjnego



Rysunek 5: Drzewo decyzyjne dla zbioru **seeds**



Rysunek 6: Drzewo decyzyjne dla zbioru **glass**

(accuracy = 0.4, f1 score = 0.35). Jednak dla zbioru **seeds** obaj modeli wykazały w jednakowym stopniu dobre wyniki (accuracy = 0.92, f1 score = 0.93).

Z powyższego porównania można wywnioskować, że są modele działające lepiej lub gorzej dla różnych zbiorów danych. Nie ma jednego modelu, który by działał najlepiej w każdym zbiorze danych.

Literatura

[1] https://cran.r-project.org/web/packages/C50/vignettes/C5.0.html.

[2] https://static1.squarespace.com/static/51156277e4b0b8b2ffe11c00/t/51e7e42ce4b0fd2e32684bca/1374151724529/user_C5.0.pdf.

[3] https://cran.r-project.org/web/packages/C50/C50.pdf.

[4] http://wazniak.mimuw.edu.pl/index.php?title=ED-4.2-m08-1.0-toc.

[5] https://rpubs.com/cyobero/C50.