Wprowadzenie do sztucznej inteligencji - ćwiczenie nr 4

Jan Górski

6 maja 2024

Spis treści

1 Zadanie

1.1 Zadanie

Zaimplementować klasyfikator ID3 (drzewo decyzyjne). Atrybuty nominalne, testy tożsamościowe. Podać dokładność i macierz pomyłek na zbiorach: **Breast cancer** i **mushroom**. Dlaczego na jednym zbiorze jest znacznie lepszy wynik niż na drugim? Do potwierdzenia lub odrzucenia postawionych hipotez konieczne może być przeprowadzenie dodatkowych eksperymentów ze zmodyfikowanymi zbiorami danych. Sformułować i spisać wnioski.

1.2 Wskazówki dotyczące ćwiczenia

- Atrybuty nominalne - każdy atrybut może przyjmować jedną z kilku dozwolonych wartości, zakładamy, że wartość atrybutu to napis, np. “kot”, “a”, “20-34”, “>40”.
- Testy tożsamościowe - jeżeli atrybut testowany w danym węźle ma np. 3 dozwolone wartości, np. a, b, c, to z węzła tego wychodzą 3 krawędzie oznaczone: a, b, c.
- Na tym ćwiczeniu klasyfikator trenuje się na zbiorze trenującym, a ocenia jego jakość na zbiorze testującym. Należy losowo podzielić zbiór danych na trenujący i testujący w stosunku 3:2.
- Jeżeli zbiór danych zawiera numery lub identyfikatory wierszy to należy je wyrzucić - nie chcemy uczyć się identyfikatorów wierszy.
- Brakujące wartości atrybutów taktujemy jako wartość, np. jeżeli symbol “?” oznacza brakującą wartość, a symbole “a”, “b” wartości normalne, to z naszego punktu widzenia mamy 3 wartości normalne (fachowo: 3 wartości atrybutu): “a”, “b”, “?”.
- Tak naprawdę to nie musimy rozumieć dziedziny problemu - na wejściu mamy napisy, na wyjściu napisy, nie ważne czy klasyfikujemy sekwencje DNA, grzyby, czy samochody.
- Nazwa pliku ze zbiorem danych jest parametrem algorytmu klasyfikacji, kod klasyfikatora powinien być w stanie obsłużyć inny zbiór danych o tym samym rozkładzie kolumn (czyli nie należy wpisywać wartości atrybutów “na sztywno” w kodzie).
- W repozytorium ze zbiorami danych zwykle w plikach “.names” jest napisane, który atrybut to klasa (czyli wartości której kolumny mamy się nauczyć przewidywać).

2 Część teoretyczna

2.1 Drzewo decyzyjne

Przedmiotem klasyfikacji drzewa decyzyjnego są obiekty w danym zbiorze U charakteryzowane przez pewien zestaw D atrybutów nominalnych. Każdy obiekt z U ma $|D|$ atrybutów, z których każdy ma wartość ze skończonego zbioru. Każdy obiekt w U jest pewnej klasy, przy czym zbiór wszystkich klas to Y.

Zadanie polega na zbudowaniu klasyfikatora, który na podstawie atrybutów będzie odgadywał klasy obiektów.

2.2 Algorytm ID3: pseudokod

Algorithm 1: Iterative Dichotomiser 3

Data: Y : zbiór klas,
D : zbiór atrybutów wejściowych,
$U \neq 0$: zbiór par uczących
Result: Drzewo decyzyjne

begin
  if $y_i == y \quad \forall \langle x_i, y_i \rangle \in U$ **then**
    return Liść zawierający klasę y.
  end if
  if $|D| == 0$ **then**
    return Liść zawierający najczęstszą klasę w U.
  end if
  $d = \arg\max_{d \in D} InfGain(d, U)$
  $U_j = \{ \langle x_i, y_i \rangle \in U : x_i[d] = d_j \}$, ***gdzie***
   d_j - j-ta wartość atrybutu d
  return Drzewo z korzeniem d oraz krawiędziami:
   $d_1, d_2, \ldots$ prowadzącymi do drzew:
   $ID3(Y, D - \{d\}, U_1), ID3(Y, D - \{d\}, U_2) \ldots$
end

2.3 Miara zróżnicowania - entropia

Kluczowym elementem algorytmu jest wybór atrybutu przypisanego do korzenia drzewa. Najlepiej byłoby wtedy, gdyby na podstawie atrybutu dało się podzielić zbiór U na podzbiory, takie że w każdym z nich występują wyłącznie obiekty innej klasy. Nie jest to zwykle możliwe, dlatego stosuje się kryterium zmierzające do stworzenia sytuacji zbliżonej, tj. jak największego zróżnicowania występowania poszczególnych klas w podzbiorach. Miarą tego zróżnicowania jest entropia:

$$I(U) = -\sum_i f_i \ln f_i,$$

gdzie f_i - częstość i-tej klasy.

Entropia zbioru podzielonego na podzbiory jest to średnia ważona entropii podzbiorów, a mianowicie

$$Inf(d,U) = \sum_j \frac{|S_j|}{|S|} I(S_j),$$

gdzie $|S|$ to liczba elementów zbioru S, zaś S_j, $j = 1, 2, \ldots$ to zbiory powstałe przez podział zbioru S ze wzsględu na wartość atrybutu D.

Zdobycz informacyjna służąca do wyboru atrybutu d ma następującą definicję:

$$InfGain(d,U) = I(U) - Inf(d,U).$$

3 Część praktyczna

3.1 Badanie zbiorów "Breast cancer" i "Mushroom"

Pierwszą rzeczą po zaimplementowaniu algorytmu było przetestowanie go na zbiorach: "Breast cancer" oraz "Mushroom". Oba zbiory zostały podzielone na dwa podzbiory: zbiór trenujący oraz zbiór testujący w relacji 3:2 w sposób losowy.

W tabelach zostały przedstawione przykładowe wyniki uzyskane za pomocą funkcji *classification_report(actual, predicted)* z pakietu *sklearn.metrics*, gdzie *actual* to rzeczywiste klasy próbek, a *predicted* to zbiór klas wyznaczonych dla tych próbek za pomocą algorytmu.

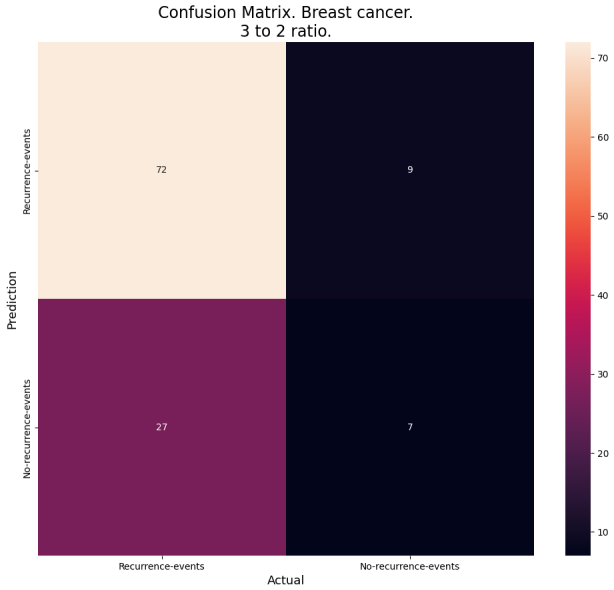
Na obrazkach zostały przydstawione macierze pomyłek dla zbiorów dla przykładowych wywołań algorytmu.

Wyniki dla zbioru "Breast cancer"				
	precision	recall	f1-score	support
no-recurrence-events	0,73	0,89	0,80	81
recurrence-events	0,44	0,21	0,28	34
accuracy	-	-	0,69	115
macro avg	0,58	0,55	0,54	115
weighted avg	0,64	0,69	0,65	115

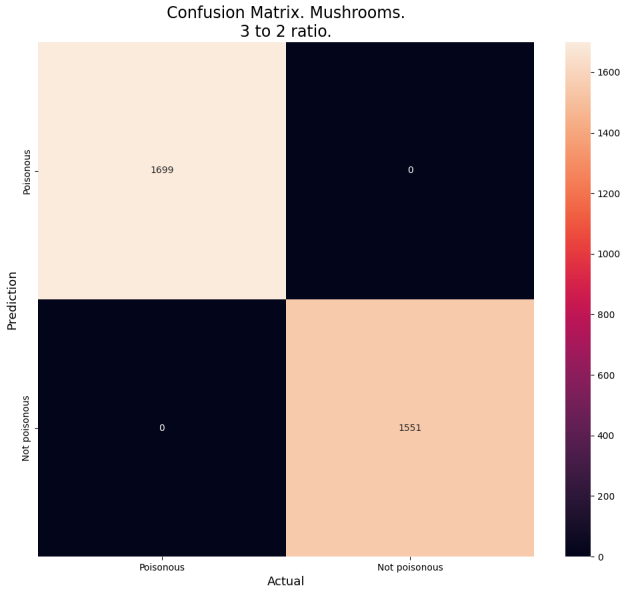
Tablica 1: Wyniki pomiarów dla zbioru "Breast cancer" przy podziale o stosunku 3:2.

Wyniki dla zbioru "Mushroom"				
	precision	recall	f1-score	support
e	1,00	1,00	1,00	1699
p	1,00	1,00	1,00	1551
accuracy	-	-	1,00	3250
macro avg	1,00	1,00	1,00	3250
weighted avg	1,00	1,00	1,00	3250

Tablica 2: Wyniki pomiarów dla zbioru "Mushroom" przy podziale o stosunku 3:2.



(a) Macierz pomyłek dla zbioru “Breast cancer”.



(b) Macierz pomyłek dla zbioru “Mushroom”.

Rysunek 1: Macierze pomyłek dla badanych zbiorów przy podziale o stosunku 3:2.

Dokładność algorytmu na zbiorze “Breast cancer” wyniosła 0.69, podczas gdy dla zbioru “Mushrooms” wyniosła 1.00. Są to wyniki porównywalne z tymi przedstawiony na stronie uzyskanymi dla innych algorytmów, z której zbiory zostały pozyskane.

3.2 Badanie zbioru “Mushroom” dla różych parametrów

W dalczej części ćwiczenia zostały wykonane dodatkowe testy mające na celu sprawdzenie, co jest przyczyną odmiennych wyników dla obu zbiorów testowanych.

Pierwszymi parametrami, które zostały sprawdzone były rozmiar zbiorów oraz liczba atrybutów uwzględnionych w klasyfikacji.

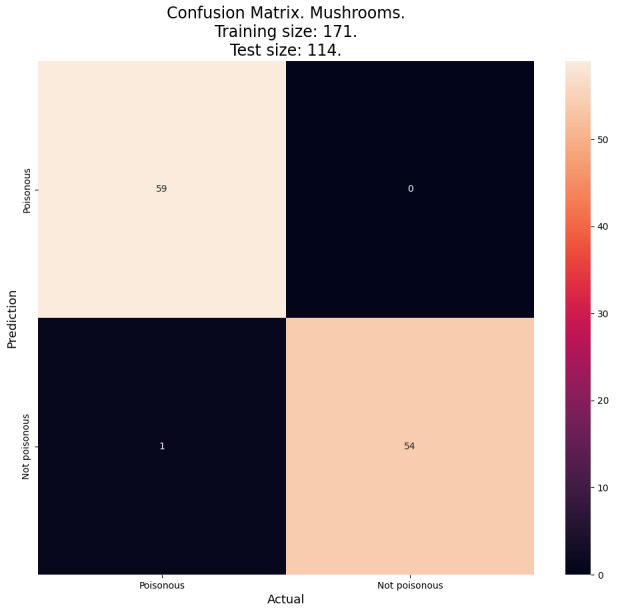
3.2.1 Sprawdzenie wpływu rozmiaru zbioru

Zbiór “Breast cancer” zawiera 286 obiektów, podczas gdy “Mushroom” zawiera ich 8124. W celu sprawdzenia, czy ma to wpływ na wyniki badań, zamiast wykorzystywać cały zbiór “Mushrooms” wykorzystałem tyle obiektów, ile zawiera zbiór “Breast cancer”. Obiekty były wybierane w sposób losowy.

W tabeli oraz na obrazku przedstawiono rezultat badań.

Wyniki dla zbioru “Mushroom”				
	precision	recall	f1-score	support
e	0,98	1,00	0,99	59
p	1,00	0,98	0,99	55
accuracy	-	-	0,99	114
macro avg	0,99	0,99	0,99	114
weighted avg	0,99	0,99	0,99	114

Tablica 3: Wyniki pomiarów dla zbioru “Mushroom” przy podziale o stosunku 3:2.



Rysunek 2: Macierz pomyłek dla zbioru “Mushroom” dla rozmiaru zbioru odpowiadającemu rozmiarowi zbioru “Breast cancer” przy podziale o stosunku 3:2.

Z testu nie wynika, że zmniejszenie liczby próbek jest w tym przypadku źródłem różnicy w wynikach. Algorytm ma niemal taką samą dokładność. Znacznie częściej pojawiają się sytuacje, w których w zbiorze trenującym nie było wartości atrybutu, który wystąpił w zbiorze testowym. Takich sytuacji nie rozwiązywałem: wyszukiwałem inne ziarno dla generatora liczb pseudolosowych dla którego testy przebiegały bez wystąpienia wyjątku.

Mniejszy rozmiar zbioru obiektów nie jest tutaj powodem różnic.

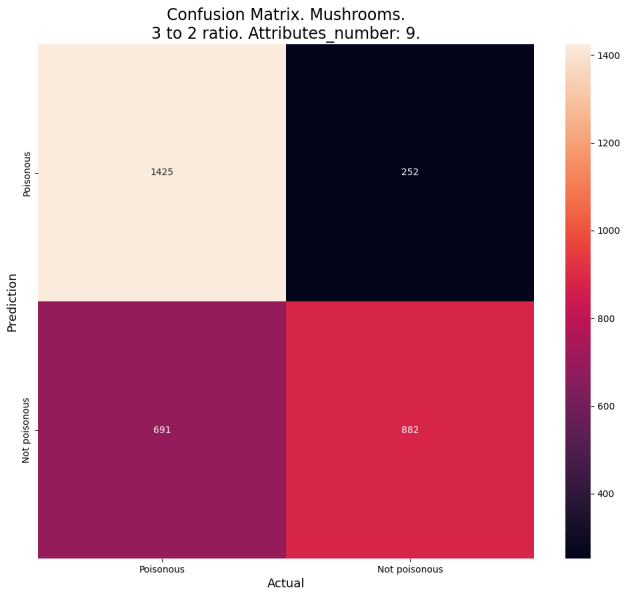
3.2.2 Sprawdzenie wpływu liczby atrybutów

Obiekty ze zbioru “Mushroom” są charakteryzowane przez zestaw 22 atrybutów. W przypadku zbioru “Breast cancer” liczba atrybutów jest równa 9. W tym eksperymencie, zamiast korzystać ze wszystkich atrybutów obiektów ze zbioru “Mushroom”, wybierałem losowo 9 z nich i przeprowadzałem testy uwzględniając tylko te wybrane.

W tabeli oraz na obrazku zostały przedstawione wyniki dla najgorszego przypadku uzyskanego w trakcie testów.

Wyniki dla zbioru “Mushroom”				
	precision	recall	f1-score	support
e	0,67	0,85	0,75	1677
p	0,78	0,56	0,65	1573
accuracy	-	-	0,71	3250
macro avg	0,73	0,71	0,70	3250
weighted avg	0,72	0,71	0,70	3250

Tablica 4: Wyniki pomiarów dla zbioru “Mushroom” dla liczby atrybutów odpowiadającej liczbie atrybutów zbioru “Breast cancer” przy podziale o stosunku 3:2.



Rysunek 3: Macierz pomyłek dla zbioru “Mushroom” dla liczby atrybutów odpowiadającej liczbie atrybutów zbioru “Breast cancer” przy podziale o stosunku 3:2.

W tym eksperymencie zaobserwowałem największą do tej pory różnicę w wynikach testów. W większości przypadków, dokładność algorytmu wynosiła ok. 100% lub ok. 90%. Dla jednego przypadku jednak, uzyskałem rezultat wynoszący 71%, co znacząco odbiegało od dotychczasowych wyników.

Rozbieżność w wynikach może wynikać z tego, że w niektórych wywołaniach testu do zbioru testowanego nie zostały uwzględnione atrybuty "ważniejsze" od pozostałych. Mniejsza liczba atrybutów zmniejsza możliwość wyboru najlepiej dzielącego zbiór atrybutu w danej gałęzi drzewa.

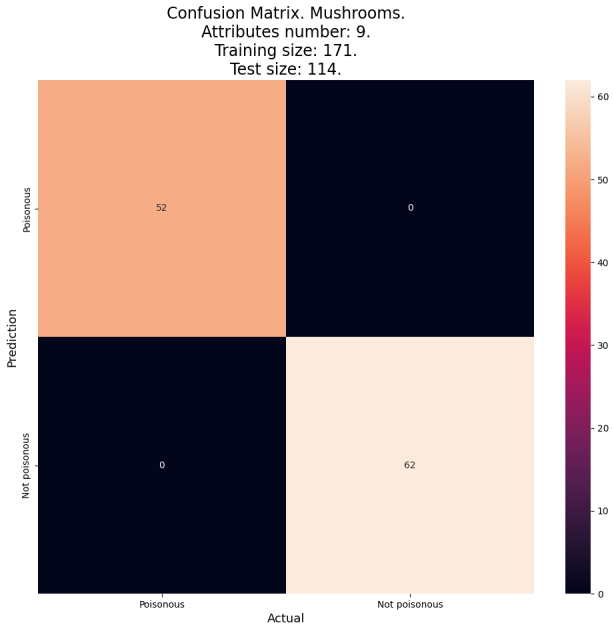
3.2.3 Zastosowanie rozmiaru zbiorów oraz liczby atrybutów takiej, jaka jest w zbiorze "Breast cancer"

W tym eksperymencie niejako zostały połączone dwa podejścia wykonane w powyższych punktach.

W tabeli i na obrazku przedstawiono rezultaty jednego z testów.

Wyniki dla zbioru "Mushroom"				
	precision	recall	f1-score	support
e	1,00	1,00	1,00	52
p	1,00	1,00	1,00	62
accuracy	-	-	1,00	114
macro avg	1,00	1,00	1,00	114
weighted avg	1,00	1,00	1,00	114

Tablica 5: Wyniki pomiarów dla zbioru "Mushroom" dla liczby atrybutów oraz rozmiaru zbioru odpowiadającym liczbie atrybutów i romiarowi zbioru "Breast cancer" przy podziale o stosunku 3:2.



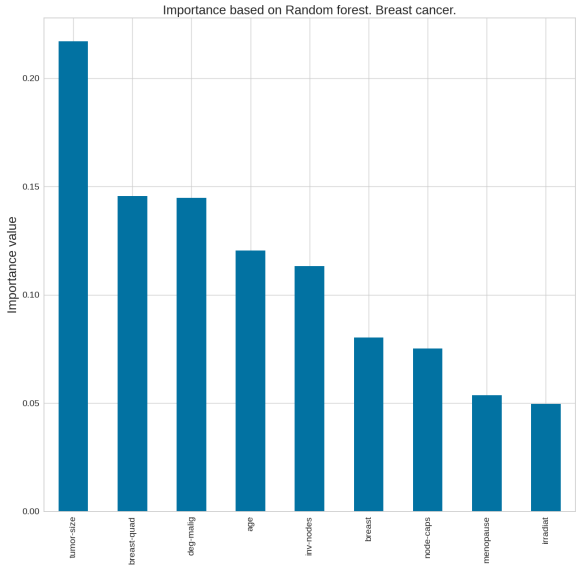
Rysunek 4: Macierz pomyłek dla zbioru "Mushroom" dla liczby atrybutów oraz rozmiaru zbioru odpowiadającym liczbie atrybutów i romiarowi zbioru "Breast cancer" przy podziale o stosunku 3:2.

Nie zaobserwowano pogorszenia się wyników względem pierwotnego wyniku.

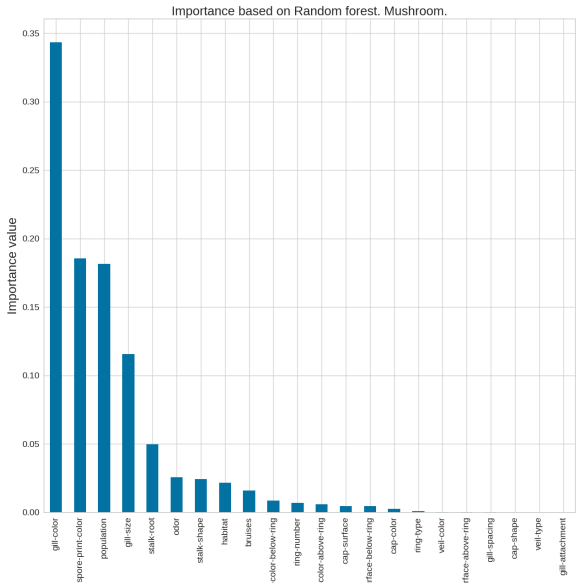
3.3 Wpływ poszczególnych atrybutów w zbiorze na wyniki predykcji

Wiedząc, że atrybuty mogą mieć wpływ na wyniki testów chciałem sprawdzić, jaki wpływ na klasę obiektu ma każdy atrybut. Zostało to przetestowane z użyciem klas *RandomForestRegressor* *RandomForestClassifier* z biblioteki *sklearn.ensemble*. Wartości atrybutów zostały zamapowane na liczby.

Na obazka zostały przedstawione "ważności" poszczególnych atrybutów dla określonych zbiorów.



(a) “Breast cancer”



(b) “Mushroom”

Rysunek 5: Wykresy przedstawiające “ważności” poszczególnych atrybutów przy klasyfikacji.

W zbiorze “Mushroom” najbardziej znaczącym atrybutem jest *gill-color*. Jego ważność została oceniona na 0.35. W zbiorze “Mushrooms” duża część atrybutów ma znikome znaczenie przy tworzeniu drzewa decyzyjnego. Może to sugerować, że decyzja o tym, do jakiej klasy należy dany obiekt jest podejmowana znacznie wcześniej, czyli i głebokość drzewa decyzyjnego jest niższa. Świadczy to o tym, że istnieją atrybuty które bardzo dobrze rozdzielają zbiór obiektów na maksymalnie różne podzbiory za pomocą wyliczania entropii zbiorów.

Dla zbioru “Mushroom” 5 atrybutów ma wpływ większy niż 0.05, podczas gdy dla zbioru “Breast cancer” wszystkie atrybuty miały znaczenie o wartości równej co najmniej 0.05.

4 Wnioski

Udało się zaimplementować algorytm ID3.

Wyniki działania algorytmu mogą się w znaczący sposób od zbioru, na którego podstawie jest tworzone drzewo decyzyjne i są przeprowadzane testy.

Rozmiar zbioru wejściowego nie musi koniecznie oznaczać, że jeden ze zbiorów jest łatwiejszy do zbadania niż inny. Możliwe jest uzyskiwanie dobrych wyników, trenując na małym zbiorze, jeśli dobrze są w nim odzwierciedlone właściwości całej populacji.

Znacząct wpływ na wyniki algorytmu ma to, jak atrybuty obiektów ze zbioru badanego mogę podzielić zbiór na podzbiory ze względu na klasy tych obiektów. IM uzyskiwane podzbiory bardziej różnią się od siebie nawzajem, tym lepsze wyniki możemy uzyskać. Wiąże się to z krokiem działania algorytmu, w którym to jest wyznaczany najlepszy atrybut pod względem uzyskiwanego zysku zbioru przy podziale opartym na wartościach tego atrybutu.

W celu sprawdzenia tej hipotezy możnaby wykonać więcej testów na zbiorach, o których wiemy, jak dobrze różna wartość atrybutów skutkuje różną wartością klasy obiektu i porównać to ze zbioramy, w których nie ma widocznie sensownych sposobów podziału zbioru.