Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Uczenie Maszynowe

Dokumentacja wstępna, Drzewo decyzyjne w zadaniu klasyfikacji danych z brakującymi wartościami atrybutów

Daniel Adamkowski, Mateusz Zembroń

Warszawa, 2021

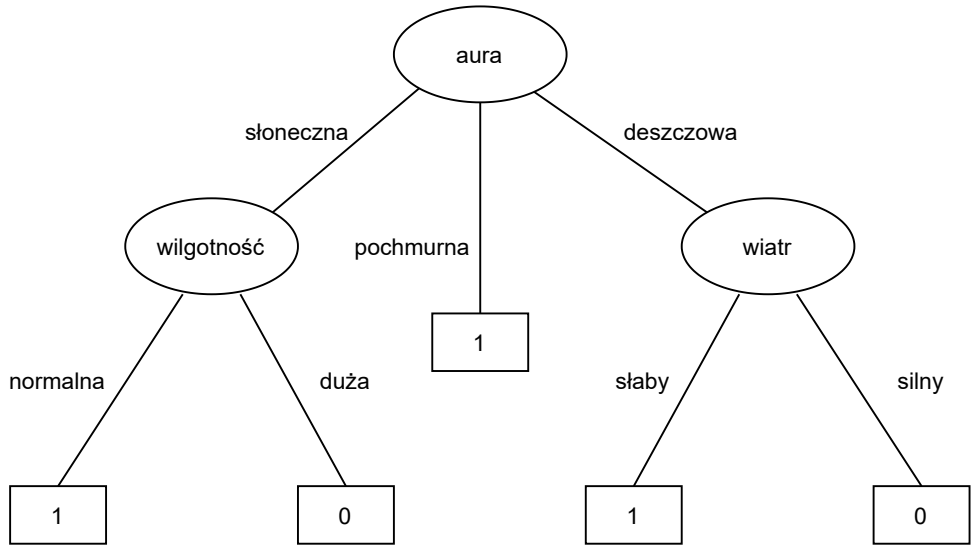
Spis treści

1. Wstęp

Problem brakujących wartości w zbiorze danych jest powszechny i naturalny. W rzeczywistości niemożliwe jest zapewnienie: całkowitego wypełnienia bardzo dużego zbioru danych (np. zmuszenie ankietowanych do odpowiedzi na każde pytanie), niezawodności działania przyrządów pomiarowych w dłuższym okresie lub warunków które zniwelują wpływ potencjalnych zakłóceń. Są to jedynie wybrane przykłady źródeł powstania tego zjawiska.
Istnieje również wiele podejść obsługi danych z brakującymi wartościami, w tym dokumencie zostaną przytoczone wybrane sposoby uodpornienia drzewa decyzyjnego od takich przypadków.

2. Drzewo decyzyjne

Drzewo decyzyjne jest jednym z najbardziej intuicyjnych sposobów reprezentacji hipotez. Składa się ono z węzłów przechowujących testy, które sprawdzają wartości atrybutów przykładów oraz liści które przypisują im konkretne kategorie. Dla każdego z możliwych wyników testu, z węzła prowadzi odpowiadająca mu gałąź do pewnego poddrzewa. Działanie tego algorytmu bardzo dobrze oddaje przykładowe drzewo decyzyjne widoczne na rys. 2.1, klasyfikujące stany pogody - sprawdzające czy jest ona odpowiednia do gry w piłkę nożną.



Rys. 2.1. Drzewo decyzyjne dla stanów pogody

2.1. Budowa drzewa decyzyjnego

Konstruowanie drzewa decyzyjnego nie jest głównym zadaniem tego projektu, jednak jest ono niezbędne aby móc w jakikolwiek sposób podjąć próbę implementacji algorytmów które zostaną przedstawione. Przed rozpoczęciem tego procesu należy podjąć kilka ważnych decyzji:

— W jaki sposób dobrane zostanie kryterium stopu?
— Jaka etykieta zostanie ustalona liściowi, jeśli kryterium stopu będzie spełnione?
— W jaki sposób w każdym węźle będzie wybierany test?

Wstępną odpowiedź na te pytania można przedstawić już na etapie planowania, jednak autorzy tego dokumentu zastrzegają sobie prawo do ich zmiany w trakcie implementacji oraz podania ostatecznych parametrów w dokumentacji końcowej.

Najbardziej intuicyjnymi kryteriami stopu są: wyczerpanie przykładów o różnych etykietowaniach (jeśli do danego węzła wchodzą przykłady o takim samym etykietowaniu, nie ma powodu aby testować ich atrybuty - węzeł można zamienić na liść), całkowite wyczerpanie przykładów (do węzła nie wchodzą jakiekolwiek przykłady) lub wyczerpanie możliwych do przeprowadzenia testów. W każdym z tych przypadków powstałemu liściowi powinna zostać przydzielona odpo-

wiednia etykieta. W przypadku pierwszej przytoczonej sytuacji - brak jakichkolwiek przykładów, liściowi przypisana zostanie kategoria występująca najczęściej na bezpośrednio wcześniejszym poziomie rekursji. W drugim przypadku sytuacja jest trywialna - liściowi należy nadać taką etykietę jaką mają wszystkie przykłady które do niego dotarły. W ostatnim przypadku przyjęta zostanie wartość jaką reprezentuje większość wchodzących do niego przykładów.

Nieustalona pozostaje jeszcze kwestia wyboru testu. W wypadku tego projektu zastosowane zostaną jedynie testy nierównościowe (związane jest to z wybranymi zbiorami danych, które zostaną omówione później). Jako kryterium wyboru testu, najprawdopodobniej zostanie wybrany przyrost informacji, obliczany zgodnie ze wzorem:

$$g_t(P) = I(P) - E_t(P) \tag{2.1}$$

Gdzie:

— $I(P)$ to informacja zawarta w zbiorze etykietowanych przykładów P:

$$I(P) = \sum_{d \in C} -\frac{|P^d|}{|P|} \log \frac{|P^d|}{|P|} \tag{2.2}$$

— $E_t(P)$ to średnia ważona entropia:

$$E_t(P) = \sum_{r \in R_t} -\frac{|P_{tr}|}{|P|} E_{tr}(P) \tag{2.3}$$

wyliczana przy pomocy entropii zbioru przykładów P:

$$E_{tr}(P) = \sum_{d \in C} -\frac{|P_{tr}^d|}{|P_{tr}|} \log \frac{|P_{tr}^d|}{|P_{tr}|} \tag{2.4}$$

Dzięki ustaleniu prawdopodobnych odpowiedzi na przytoczone wcześniej pytania można przedstawić algorytm budowy drzewa decyzyjnego:
Jako argumenty wejściowe funkcji $buduj - drzewo(P, d, S)$, zwracającej drzewo dyzyjne reprezentujące hipotezę c na zbiorze P, wykorzystane zostaną:

— P - zbiór przykładów etykietowanych pojęcia c,
— d - domyślna etykieta kategorii,
— S - zbiór możliwych testów

```
if kryterium-stopu(P, S) then
    utwórz liść 1;
    d1 := kategoria(P,d);
    return 1;
endif
utwórz węzeł n;
tn := wybierz-test(P,S);
d := kategoria(P,d);
for r in Rtn
    n[r] := buduj-drzewo(Ptnr, d, S - {tn});
endfor
return n;
```

2.2. Algorytmy uodparniające drzewo decyzyjne od brakujących wartości atrybutów

Procesy uczenia maszynowego są wrażliwe na błędy w danych, w większym lub mniejszym stopniu. Błędy te mogą być różnej natury i mieć różne przyczyny, ale w rzeczywistym środowisku są one nieuniknione. Istnieje wiele podejść stosowanych do radzenia sobie z brakującymi wartościami, m.in.

— Usuwanie próbek,
— Podział próbki (wykorzystanie obiektów ułamkowych),
— Uzupełnianie atrybutów,
— Dla atrybutów nominalnych - potraktowanie braku jako kolejnej wartości.

W zależności od rozwiązywanego problemu oraz typu danych należy posłużyć się stosownym algorytmem.

2.2.1. Brakujące wartości niektórych atrybutów w niektórych próbkach

Aby rozwiązać ten problem możemy wyeliminować odpowiednie próbki lub atrybuty. W każdym przypadku tracimy dane, co może być nieakceptowalne.
Alternatywnym podejściem może być uzupełnienie brakujących wartości sztucznie generowanymi danymi. Muszą one być wygenerowane losowo, przez staranne dopasowanie rozkładu wartości danego atrybutu, z odpowiednim skalowaniem liczby takich próbek.

2.2.2. Algorytm wykorzystujący podziały zastępcze

W przypadku występowania brakujących wartości w danych w drzewie wykorzystuje się podziały zastępcze (surrogate splitters), które są swego rodzaju zabezpieczeniem pozwalającym uzyskać możliwie podobny wynik podziału przykładów co wybrany warunek. Węzły w drzewie zawierają listę takich alternatywnych podziałów opartych na różnych atrybutach. Podziały z listy wykorzystuje się w przypadku braku wartości, na podstawie której można by było wnioskować przy użyciu podstawowej reguły węzła.
Przykładem stosowania takiego rozwiązania jest CART (Classification And Regression Tree), służący do budowania drzew binarnych, co oznacza, że dla atrybutów nominalnych rozważa się wszystkie możliwości podziału na dwie rozłączne grupy wartości.

Reguła	Przypisanie do gałęzi 1	Przypisanie do gałęzi 2
Główna	X $\geqslant$ 1	X $<$ 1
Zastępcza nr 1	Brakujący X, Y $<$ 0	Brakujący X, Y $\geqslant$ 0
Zastępcza nr 2	Brakujące X i Y, Z $\geqslant$ 100	Brakujące X i Y, Z $<$ 100
Domyślna	Brak	Brakujące X, Y, Z

Tab. 2.1. Przykład klasyfikacji z wykorzystaniem alternatywnych atrybutów

2.3. Brakujące wartości podczas klasyfikacji - klasyfikacja probabilistyczna

Jeśli wynik konkretnego testu t (w danym węźle) nie może zostać ustalony ze względu na brakującą wartość atrybutu, można posłużyć się podejściem probabilistycznym. Dla każdego węzła za pomocą zbioru trenującego i zawartych w nim elementów o znanym wyniku testu t, można wyznaczyć prawdopodobieństwa osiągnięcia konkretnych liści przez element o nieznanym wyniku testu t (wyznaczyć to można na etapie budowy drzewa). Prawdopodobieństwo to można oszacować zgodnie ze wzorem:

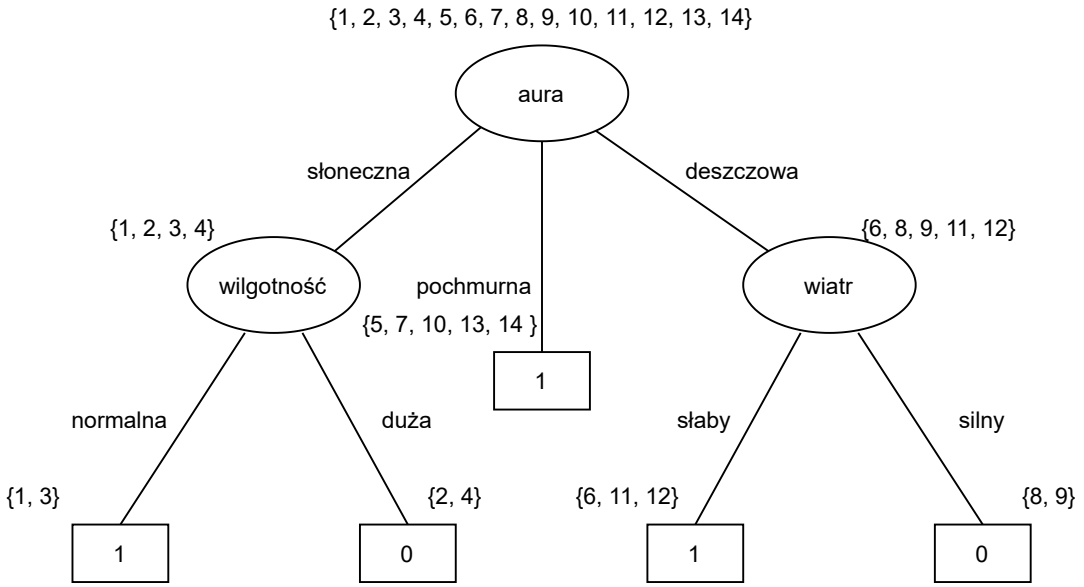
$$Pr(c(x_*) = d) = \sum_{l \in \boldsymbol{L_T}} Pr(l|x_*) \cdot Pr_{x \in \Omega}(c(x) = d|l) \tag{2.5}$$

Gdzie: $Pr_{x\in\Omega}(c(x) = d|l)$ to prawdopodobieństwo tego, że kategoria przykładu x wybranego z dziedziny zgodnie z rozkładem Ω i zaliczonego do liścia $\boldsymbol{l}$ jest d. Możemy je oszacować w następujący sposób:

$$Pr_{x\in\Omega}(c(x) = d|l) = \frac{|P^d_{T,l}|}{|P_{T,l}|} \tag{2.6}$$

Dla lepszego zrozumienia algorytmu można przytoczyć przykład:

Dane jest następujące drzewo decyzyjne (przytoczone wcześniej - w tym wypadku uzupełnione o numery przykładów - w nawiasach klamrowych):



Rys. 2.2. Drzewo decyzyjne dla stanów pogody

Na wejściu otrzymujemy przykład x_* o nieznanej wartości atrybutu *aura*. Załóżmy, że pozostałe atrybuty są znane i wynoszą:

— wilgotność(x_*) = duża,
— wiatr(x_*) = silny

Na podstawie zbioru trenującego (przedstawionego na drzewie decyzyjnym rys. 2.2). Można oszacować prawdopodobieństwa poszczególnych wyników testu *aura*:

— $Pr_{x\in\Omega}$(aura(x) = słoneczna) = $\frac{4}{14}$,
— $Pr_{x\in\Omega}$(aura(x) = pochmurna) = $\frac{5}{14}$,
— $Pr_{x\in\Omega}$(aura(x) = deszczowa) = $\frac{5}{14}$,

Wartości innych atrybutów, przykładu x_* są znane, zatem możemy obliczyć prawdopodobieństwa osiągnięcia poszczególnych liści. Kolejno dla liści od lewej do prawej:

— $Pr(\mathbf{l_1}|x_*) = \frac{4}{14} \cdot 0 = 0$
— $Pr(\mathbf{l_2}|x_*) = \frac{4}{14} \cdot 1 = \frac{4}{14}$
— $Pr(\mathbf{l_3}|x_*) = \frac{5}{14}$
— $Pr(\mathbf{l_4}|x_*) = \frac{5}{14} \cdot 0 = 0$
— $Pr(\mathbf{l_5}|x_*) = \frac{5}{14} \cdot 1 = \frac{5}{14}$

Zatem możemy przyjąć prawdopodobieństwa poszczególnych kategorii dla przykładu x_*:

— $Pr(c(x_*) = 1) = \frac{5}{14}$

— $Pr(c(x_*) = 0) = \frac{9}{14}$

Z czego wynika, że przykładowi x_* zostanie przypisana etykieta 0.

3. Plan eksperymentów oraz zbiory danych jakie zostaną podczas nich wykorzystane

3.1. Zbiory danych

Do eksperymentów wykorzystane zostaną zbiory danych ze strony: http://archive.ics.uci.edu/ml/ (UC Irvine).
Wstępnie wyselekcjonowano trzy zbiory:

— Irysy - kwiaty (150 elementów, 4 atrybuty),
— Rozpoznawanie liter (20000 elementów, 16 atrybutów),
— Rak piersi w Wisconsin (569 elementów, 32 atrybuty).

Każdy z nich posiada atrybuty liczbowe, początkowo nie występują w nich brakujące wartości (związane jest to z planowanymi eksperymentami które omówiono w następnej części).
Zostaną one podzielone na zbiór testowy oraz weryfikujący - zbiór testowy obejmie 30% całości, weryfikujący pozostałą część. Przykłady do każdej z grup zostaną dobrane w taki sposób aby każdy z parametrów był odpowiednio reprezentowany - żaden nie znajdował się w zbyt dużej lub zbyt małej ilości w porównaniu do zawartości w całym zbiorze.
Wyselekcjonowane zbiory różnią się od siebie; pierwszy z nich jest mały (zawiera niewiele elementów o jedynie czterech atrybutach) - będzie idealny do sprawdzenia poprawności działania zaimplementowanych algorytmów, pozostałe zawierają wyraźnie więcej przykładów i atrybutów - będą stanowiły większe wyzwanie.

3.2. Planowane eksperymenty

W planie do wykonania są następujące eksperymenty:

— Sprawdzanie skuteczności algorytmu przy zwiększającej się ilości brakujących wartości atrybutu. Przykłady dla których atrybut zostanie usunięty (pojawi się wartość `nieznany`) będą wybierane w sposób losowy. Badane algorytmy będą porównywane z metodą usunięcia tego atrybutu dla wszystkich przykładów i zbudowaniu drzewa bez niego (metoda referencyjna). Celem będzie również znalezienie ilości brakujących wartości od której warto usunąć atrybut.
— Sprawdzenie działania dla większej ilości atrybutów o brakujących wartościach.
— Sprawdzenie jaki wpływ ma ilość wszystkich atrybutów (czy jeśli wiadomo, że mogą wystąpić brakujące wartości to warto poszukać innych możliwych parametrów do zmierzenia?)
— Sprawdzenie działania algorytmu na zbiorze danych w którym brakujące wartości występowały pierwotnie - nie zostały sztucznie usunięte.

Bibliografia

[1] P. Cichosz: Systemy uczące się, Wydawnictwo Naukowo - Techniczne Warszawa, 2000.