Uniwersytet Ekonomiczny w Katowicach
Katedra Uczenia Maszynowego

Jan Kozak

Systemy uczące się
(drzewa decyzyjne ID3 / C4.5)

Katowice, 2022 r.

Spis treści

Informacje wstępne – drzewo decyzyjne, jako reprezentacja wiedzy

W kolejnych częściach dokumentu przedstawiony jest przykład budowy drzewa decyzyjnego ID3 / C4.5 wykonany na podstawie źródła:

`www.mimuw.edu.pl/~awojna/SID/referaty/strzelczak/c4_5Main.html`

1.1 Przykładowe tabele decyzyjne

W poniższych obliczeniach zaprezentowane są wyniki dla danych testowych giełda zapisanych w tabeli 1.1 oraz spreparowanych danych testowych zapisanych w tabeli 1.2.

Tabela 1.1: Tabela decyzyjna danych giełdowych (gielda.txt)

	Atrybuty warunkowe			Decyzja
	a_1	a_2	a_3	d
x_1	old	yes	swr	down
x_2	old	no	swr	down
x_3	old	no	hwr	down
x_4	mid	yes	swr	down
x_5	mid	yes	hwr	down
x_6	mid	no	hwr	up
x_7	mid	no	swr	up
x_8	new	yes	swr	up
x_9	new	no	hwr	up
x_{10}	new	no	swr	up

Tabela 1.2: Tabela decyzyjna danych testowych (testowaTabDec.txt)

	Atrybuty warunkowe				Decyzja
	a_1	a_2	a_3	a_4	d
x_1	0	1	0	1	0
x_2	1	1	0	1	0
x_3	2	2	0	1	1
x_4	2	2	1	0	2
x_5	2	2	0	0	1
x_6	1	2	0	1	0
x_7	2	1	1	1	2
x_8	0	0	1	0	1
x_9	0	0	0	0	1

1.2 Przygotowanie tabeli decyzyjnej

W początkowej fazie przygotowania algorytmu należy przygotować implementację, która pozwala na:

1. Wczytanie danych zapisanych w pliku tekstowym i odpowiadającym tabeli decyzyjnej (z jednym atrybutem decyzyjnym znajdującym się na końcu).

   ```
   Wersja podstawowa: wartości atrybutów, to 0 i 1. Atrybuty
   oddzielone spacją.
   Wersje rozszerzone: wartości, to dowolne liczby
   całkowite, liczby rzeczywiste, literały. Atrybuty
   oddzielone dowolnym symbolem (np. z ograniczonego
   zestawu).
   ```

2. Obliczenie możliwej liczby wartości każdego atrybutu.

3. Obliczenie wystąpień każdej wartości każdego atrybutu.

Entropia

2.1 Wyznaczenie entropii

$$I(P) = -(p_1 \cdot log_2(p_1) + p_2 \cdot log_2(p_2) + \cdots + p_n \cdot log_2(p_n)), \tag{2.1}$$

gdzie $p_1 \ldots p_n$ to wartości prawdopodobieństwa wystąpień każdego z elementów (w tym przypadku każdej z wartości atrybutu decyzyjnego).

Entropia jest miarą nieuporządkowania, im wyższa wartość entropii, tym większe nieuporządkowanie i odwrotnie. Przykładowe wartości entropii, dla różnych prawdopodobieństw są następujące:

- $P = (1,0; 0,0)$, czyli $p_1 = 1,0; p_2 = 0,0;$, to $I(P) = 0,0$;
- $P = (0,9; 0,1)$, to $I(P) = 0,4689955935892812$;
- $P = (0,8; 0,2)$, to $I(P) = 0,7219280948873623$;
- $P = (0,7; 0,3)$, to $I(P) = 0,8812908992306927$;
- $P = (0,6; 0,4)$, to $I(P) = 0,9709505944546686$;
- $P = (0,5; 0,5)$, to $I(P) = 1,0$;

Należy pamiętać, że maksymalna wartość entropii (najgorsze przypadki) jest zależna dla liczby parametrów (przypadków, prawdopodobieństw), co nie pozostaje bez znaczenia przy dalszych etapach budowy drzewa decyzyjnego. Jest to w zasadzie $log_2(n)$ dla n ze wzoru (2.1). Nie ma konieczności wyznaczania tych wartości, jednak są one następujące:

- $n = 2$, to 1,0;
- $n = 3$, to 1,584962500721156;

- $n = 4$, to 2,0;
- $n = 5$, to 2,321928094887362;
- $n = 6$, to 2,584962500721156;
- $n = 7$, to 2,807354922057604;
- $n = 8$, to 3,0;

2.2 Entropia według klas decyzyjnych

Entropię według klas decyzyjnych $Info(T)$ (T, to tabela decyzyjna) wyznacza się na podstawie wystąpień każdej z klasy decyzyjnych (wystąpień każdej wartości atrybutu decyzyjnego).

Wystąpienia te determinują prawdopodobieństwo wystąpienia danej klasy, dlatego wprost wyznaczana jest entropia na podstawie wzoru (2.1), czyli $Info(T) = I(P)$. W tym przypadku P przyjmuje wartości:

$$P = \left(\frac{C_1}{|T|}, \frac{C_2}{|T|}, \cdots, \frac{C_k}{|T|} \right), \tag{2.2}$$

gdzie $|T|$, to liczba obiektów w tabeli decyzyjnej; k, to liczba klas decyzyjnych (możliwych wartości atrybutu decyzyjnego); $C_1 \dots C_k$, to liczba wystąpień kolejnych wartości klas decyzyjnych.

Posiadając te wartości możliwe jest wyznaczenia entropii dla klas decyzyjnych, czyli informacji o zbiorze danych (wzór 2.3).

$$Info(T) = I \left(\frac{C_1}{|T|}, \frac{C_2}{|T|}, \cdots, \frac{C_k}{|T|} \right) \tag{2.3}$$

Dla przykładowej tabeli decyzyjnej dotyczącej giełdy (tab. 1.1) występują dwie klasy decyzyjne: *down* (5 razy) i *up* (5 razy), w związku z czym $|T| = 10$, a $Info(T)$ następujące:

$$Info(T) = I \left(\frac{5}{10}, \frac{5}{10} \right) = 1,0$$

Natomiast dla danych testowych (tab. 1.2) występują trzy klasy decyzyjne: 0 (3 razy), 1 (4 razy) i 2 (2 razy), w związku z czym $|T| = 9$, a $Info(T)$ następujące:

$$Info(T) = I\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) = 1,5304930567574824$$