

Budowa drzewa decyzyjnego za pomocą algorytmu ID3

Jarosław Arabas

Rozważmy przykładowy zbiór danych

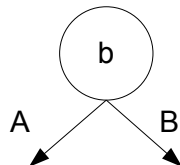
a	b	c	d	e	f
A	A	C	A	A	0
A	B	B	B	B	0
A	A	A	A	C	0
B	B	A	C	D	0
B	B	B	C	B	1
B	B	C	C	D	1
B	B	B	B	B	1
A	B	A	B	C	1

Atrybuty a-e są atrybutami warunkowymi, zaś f jest decyzyjny.

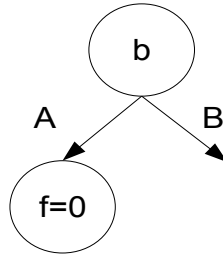
Uruchamiamy metodę ID3, która daje następujące wyniki pośrednie

		f=0	f=1	częstość wystąpień	entropia wartości atrybutu	entropia atrybutu
a	A	3	1	4	0.811278	0.811278
	B	1	3	4	0.811278	
b	A	2	0	2	0	0.688722
	B	2	4	6	0.918296	
c	A	2	1	3	0.918296	0.938722
	B	1	2	3	0.918296	
	C	1	1	2	1	
d	A	2	0	2	0	0.688722
	B	1	2	3	0.918296	
	C	1	2	3	0.918296	
e	A	1	0	1	0	0.844361
	B	1	2	3	0.918296	
	C	1	1	2	1	
	D	1	1	2	1	

Najmniejszą entropię mają atrybuty b i d, lecz b ma mniej wartości, dlatego on będzie korzeniem drzewa. Wygląda ono teraz tak



Uruchomienie metody ID3 dla lewego poddrzewa prowadzi do wstawienia węzła terminalnego z etykietą $f=0$



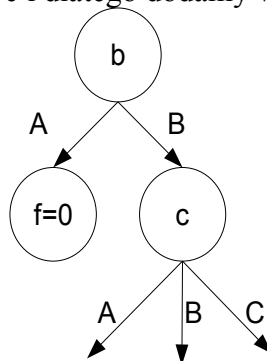
Uruchamiamy następnie ID3 dla prawego poddrzewa. Są do niego klasyfikowane następujące przykłady

a	b	c	d	e	f
A	B	B	B	B	0
B	B	A	C	D	0
B	B	B	C	B	1
B	B	C	C	D	1
B	B	B	B	B	1
A	B	A	B	C	1

Wyniki pośrednie są następujące

		f=0	f=1	częstość wystąpień	entropia wartości atrybutu	entropia atrybutu
a	A	1	1	2	1	0.874185
	B	1	3	4	0.811278	
c	A	1	1	2	1	0.792481
	B	1	2	3	0.918296	
	C	0	1	1	0	
d	A	0	0	0	0	0.918296
	B	1	2	3	0.918296	
	C	1	2	3	0.918296	
e	A	0	0	0	0	0.792481
	B	1	2	3	0.918296	
	C	0	1	1	0	
	D	1	1	2	1	

Atrybutem o najmniejszej entropii jest c i dlatego dodamy węzeł z testem na c



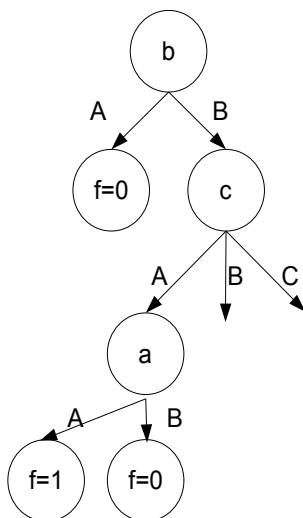
Uruchomienie ID3 dla lewego poddrzewa ($c=A$) daje następujące przykłady

a	b	c	d	e	f
B	B	A	C	D	0
A	B	A	B	C	1

i wyniki pośrednie

		f=0	f=1	częstość wystąpień	entropia wartości atrybutu	entropia atrybutu
a	A	0	1	1	0	0
	B	1	0	1	0	
d	A	0	0	0	0	0
	B	0	1	1	0	
	C	1	0	1	0	
e	A	0	0	0	0	0
	B	0	0	0	0	
	C	0	1	1	0	
	D	1	0	1	0	

Możemy wybrać dowolny atrybut, wybieramy a dlatego, że ma najmniej możliwych wartości.
Drzewo wygląda tak



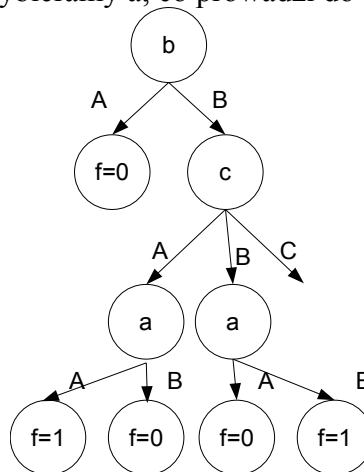
Teraz dla c=B: zbiór przykładów

a	b	c	d	e	f
A	B	B	B	B	0
B	B	B	C	B	1
B	B	B	B	B	1

i wyniki pośrednie

		f=0	f=1	częstość wystąpień	entropia wartości atrybutu	entropia atrybutu
a	A	1	0	1	0	0
	B	0	2	2	0	
d	A	0	0	0	0	1
	B	1	1	2	1	
	C	0	1	1	0	
e	A	0	0	0	0	0.918296
	B	1	2	3	0.918296	
	C	0	0	0	0	
	D	0	0	0	0	

Możemy wybrać a lub d, znowu wybieramy a, co prowadzi do drzewa



W końcu rozważamy przypadek $c=C$ i widzimy że wszystkie przykłady mają tę samą wartość atrybutu decyzyjnego, więc ostateczna postać drzewa to

