Dokumentacja

Mateusz Wezdeńko

Index: 304124

Zadanie 4.2

Opis algorytmu:

Do tworzenia drzew decyzyjnych został zastosowany algorytm id3 z metodą podziału „information gain”. Do sprawdzania poprawności wyników została użyta walidacja krzyżowa z różną ilością podziałów.

Wyniki walidacji krzyżowej:

- Zbiór danych:

buying price	vhigh	high	med.	low
maintenance	vhigh	high	med.	low
doors	2	3	4	5-more
people	2	4	more	
luggage boot	small	med	big	
safety	low	med	high	
class	unacc	acc	good	v-good

- Liczba instancji: 1728
- Dystrybucja danych:

class	**N**	**N[%]**
unacc	1210	70,12%
acc	387	22,22%
good	69	3,99%
v-good	65	3,76%

- Wyniki walidacji:

k-podziałów	poprawność
2	71,06%
3	70,75%
4	70,59%
5	70,50%
6	70,31%
7	70,41%
8	70,27%
średnia	70,56%

- Zbiór danych:

age	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99				
menopause	lt40	ge40	premeno										
tumor-size	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	
inv-nodes	0-2	3-5	6-8	9-11	12-14	15-17	18-20	21-23	24-26	27-29	30-32	33-35	36-39
noed-caps	yes	no											
deg-malig	1	2	3										
breast	left	right											
breast-quad	left-up		left-low		right-up		central						
irradiat	yes	no											
class	yes	no											

- Liczba instancji: 286
- Dystrybucja danych:

class	**N**	**N[%]**
yes	1210	70,12%
no	387	22,22%

- Wyniki walidacji:

k-podziałów	poprawność
2	74,76%
3	75,81%
4	76,36%
5	76,16%
6	76,00%
7	76,20%
8	75,74%
średnia	75,86%

Wnioski:

- Skuteczność z jaką algorytm daje poprawne wyniki zależy od zbioru danych, oraz ilości atrybutów i ilości klas
- Przy walidacji krzyżowej zwiększona ilość podziałów nie daje znaczących różnic w wynikach
- W obu testach uzyskana poprawność wyników to około 70% co oznacza że algorytm jest skuteczniejszy w przewidywaniu od zwykłego losowania wyników