A6M33DVZ - příprava pro cvičící – cvičení 2 - Rozhodovací stromy

1 Rozhodovací stromy - algoritmus ID3

Nechť S je trénovací množina (množina klasifikovaných příkladů)

1. Nalezni "nejlepší" atribut at_0 (t.j. atribut, jehož hodnoty nejlépe diskriminují mezi pozitivní a neg. příklady) a tím ohodnoť kořen vytvářeného stromu.

2. Rozděl množinu S na podmnožiny $S_1, S_2, \ldots, S_n$ podle hodnot atributu at_0 a pro každou množinu příkladů S_i vytvoř nový uzel jako následníka právě zpracovávaného uzlu (kořenu)

3. Pro každý nově vzniklý uzel s přiřazenou podmnožinou S_i proveď: Jestliže všechny příklady v S_i mají tutéž klasifikaci (všechny jsou pozitivní nebo všechny jsou negativní),

   (a) pak je uzel prohlášen za list vytvářeného rozhodovacího stromu (a tedy se už dále nevětví),

   (b) jinak jdi na bod 1 s tím, že $S := S_i$.

2 Výběr nejinformovanějšího atributu

2.1 Kritérium entropie

Nechť $\{at_1, \ldots, at_n\}$jsou všechny atributy v trénovací množině S. Pro každý atribut at_i, $i \in \{1, \ldots, n\}$, vypočteme entropii H^i. Říkáme, že atribut at je nejlepší pro rozdělení trénovací množiny S, jestliže platí $H(at) = \min\left\{H^1, \ldots, H^n\right\}$, kde entropii jednotlivých atributů H^i, $i \in \{1, \ldots, n\}$spočteme jako

$$H^i = \sum_{j=1}^{k} P_j H_j,$$

kde k je počet podmnožin daných atributem i,

$P_j$je poměr mohutnosti j-té podmnožiny k mohutnosti množiny všech příkladů S,

$H_j$je entropie j-té podmnožiny $H_j = -p_1 \log_2 p_1 - p_2 \log_2 p_2$, $p_1$je počet pozitivních příkladů / celkovému počtu příkladů v dané podmnožině, $p_2$je počet negativních příkladů / celkovému počtu příkladů v dané podmnožině.

2.2 Příklad

Nalezněte nejinformovanější atribut podle kritéria entropie v trénovací množině uvedené v tabulce 1.

Table 1: Příklad na rozhodovací stromy - weather.nominal.arff

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Řešení:
$H(at) = \min\left\{H^{outlook}, H^{temperature}, H^{humidity}, H^{windy}\right\}$
atribut outlook
$H^{outlook} = \frac{5}{14}H_{sunny} + \frac{4}{14}H_{overcast} + \frac{5}{14}H_{rainy} = 0.694$
$H_{sunny} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.442 + 0.529 = 0.971$
$H_{overcast} = -1\log_2 1 - 0\log_2 0 = 0$
$H_{rainy} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.442 + 0.529 = 0.971$
atribut temperature
$H^{temperature} = \frac{4}{14}H_{hot} + \frac{6}{14}H_{mild} + \frac{4}{14}H_{cool} = 0.910$
$H_{hot} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 0.5 + 0.5 = 1$
$H_{mild} = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.39 + 0.528 = 0.918$
$H_{cool} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.311 + 0.5 = 0.811$
atribut humidity
$H^{humidity} = \frac{7}{14}H_{high} + \frac{7}{14}H_{normal} = 0.7785$
$H_{high} = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} = 0.442 + 0.524 = 0.966$
$H_{normal} = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.190 + 0.401 = 0.591$
atribut windy
$H^{windy} = \frac{6}{14}H_{true} + \frac{8}{14}H_{false} = 0.921$
$H_{true} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 0.5 + 0.5 = 1$
$H_{false} = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.311 + 0.5 = 0.861$

$H(at) = \min\left\{H^{outlook}, H^{temperature}, H^{humidity}, H^{windy}\right\} = \min\{0.694, 0.91, 0.7785, 0.921\}$
nejinformovanější atribut je $at_{outlook}$.

2.3 Poznámky

Mezi další kritéria patří ziskové kritérium (gain), poměrné ziskové kritérium (information gain), jsou uvedeny v přednášce o rozhodovacích stromech. Existují ale i další složitější kritéria.

Další důležitou otázkou při konstrukci rozhodovacích stromů je volba diskretizace pro numerické atributy a tudíž i možnost použít pro takové atributy uvedená kritéria pro výběr toho nejvhodnějšího.