3.

(CCCCalcularea unor entropii; aplicarea algoritmului ID3)

Centrul [de Medicină] pentru Controlul şi Prevenţia Maladiilor a fost sesizat în legătură cu o creştere surprinzătoare a apariţiilor de vampiri. Acest centru a strâns anumite date preliminare privind atât caracteristicile vampirilor [deja] cunoscuţi cât şi ale non-vampirilor, şi doreşte să construiască un arbore de decizie ca să-i ajute pe cetăţeni să identifice noi vampiri. Datele culese sunt prezentate în tabelul următor:

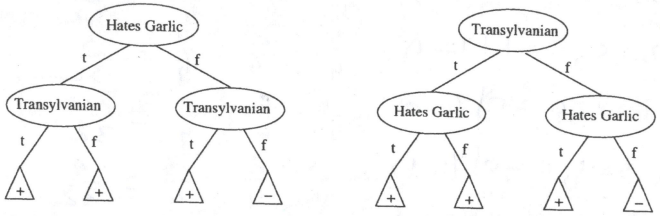
V (Vampire)	G (Hates Garlig)	T (Transylvanian)	Nr. apariţii
+	t	t	9
+	t	f	4
+	f	t	3
+	f	f	0
−	t	t	1
−	t	f	3
−	f	t	1
−	f	f	6

Fiecare linie indică ce caracteristici au fost „observate", şi de câte ori a fost observată fiecare combinaţie de caracteristici. De exemplu, combinaţia (+, t, t) a fost observată de 9 ori, pe când combinaţia (+, f, f) n-a fost observată niciodată. (Simbolii 't' şi 'f' au fost folosiţi în locul lui True şi False, pentru a evita confuzia cu atributul T.)

a. Calculaţi entropia condiţională medie $H(V|G)$. (Toate calculele intermediare trebuie făcute cu o precizie de cel puţin 4 zecimale, pentru a ne asigura că răspunsul final are primele 3 zecimale corecte.)

b. Calculaţi entropia condiţtionalăl medie $H(V|T)$. (Din nou, toate calculele intermediare trebuie făcute cu o precizie de cel puţin 4 zecimale.)

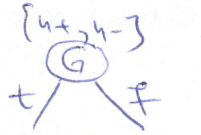
c. Care dintre arborii de mai jos reprezintă rezultatul învăţării realizate de algoritmul ID3 pe aceste date?



d. Adevărat sau Fals: Arborele produs de către ID3 va clasifica o persoană căreia nu-i place usturoiul (engl., hates garlic) dar nu este transilvănean ca fiind vampir.

Indicaţie: Este posibil să aveţi nevoie de următoarele valori pentru entropia ($H(p)$) unei variabile aleatoare Bernoulli de parametru p: $H(11/27) = 0.9751$, $H(4/17) = 0.7871$, $H(3/10) = 0.8812$, $H(1/7) = 0.5916$, $H(4/13) = 0.8904$.

a. [4+, 4-]



[2+, 2-] [2+, 2-]

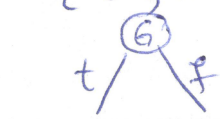
greşit

~~$H(Y|G) = \frac{4}{8} H[2+,2-] + \frac{4}{8} H[2+,2-] = 1$~~

1 1

~~b.~~

a. [16+, 11-]



[13+, 4-] [3+, 7-]

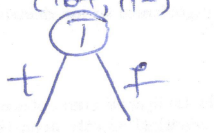
$H(Y|G) = \frac{17}{27} \cdot H[13+,4-] + \frac{10}{27} H[3+,7-] =$

$= \frac{17}{27} \cdot \left(\frac{13}{17} \cdot \log_2 \frac{17}{13} + \frac{4}{17} \log_2 \frac{17}{4} \right) +$

$+ \frac{10}{27} \left(\frac{3}{10} \log_2 \frac{10}{3} + \frac{7}{10} \log_2 \frac{10}{7} \right) =$

$= \frac{13 \lg \frac{17}{13} + 4 \lg \frac{17}{4} + 3 \lg \frac{10}{3} + 7 \lg \frac{10}{7}}{27 \lg 2} = 0,822$ ✓

b. [16+, 11-]



[12+, 2-] [4+, 9-]

$H(Y|T) = \frac{14}{27} H[12+,2-] + \frac{13}{27} H[4+,9-] =$

$= \frac{14}{27} \cdot \left(\frac{12}{14} \cdot \log_2 \frac{14}{12} + \frac{2}{14} \log_2 \frac{14}{2} \right) +$

$+ \frac{13}{27} \left(\frac{4}{13} \log_2 \frac{13}{4} + \frac{9}{13} \log_2 \frac{13}{9} \right) =$

$= \frac{12 \lg \frac{14}{12} + 2 \lg 7 + 4 \lg \frac{13}{4} + 9 \lg \frac{13}{9}}{27 \lg 2} = 0,7356$ ✓

c. Conform subpunctelor a şi b, în rădăcina va fi ales atributul T (Transylvanian) deoarece entropia sa este mai mică decât a atributului G, ceea ce înseamnă că va avea câştig de informaţie mai mare. Asadar arborele ID3 este cel de-al doilea.

d. Adevărat. Se urmăreşte pe arbore ramura f pentru atributul Transylvanian şi ramura t pentru atributul Hates Garlic