Rješenje zadatka 4.1 predmeta Strojno učenje

Siniša Biđin

20. siječnja 2013.

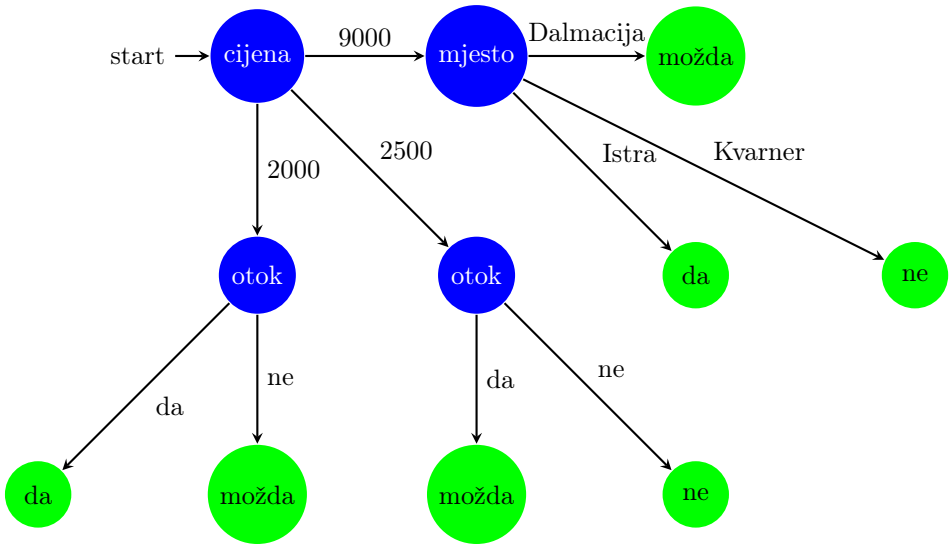
(a) Model je stablo odluke, funkcija gubitka je funkcija informacijske dobiti, a optimizacijski postupak je metoda "uspona na vrh" (engl. *hill-climbing*).

(b) Prvo računamo informacijsku dobit svih sedam atributa. Atribut s najvećom informacijskom dobiti postaje korijen stabla odluke.

$$E(D) = -p_{\text{da}} \log_2 p_{\text{da}} - p_{\text{ne}} \log_2 p_{\text{ne}} - p_{\text{možda}} \log_2 p_{\text{možda}} =$$
$$= -\frac{7}{15} \log_2 \frac{7}{15} - \frac{1}{3} \log_2 \frac{1}{3} - \frac{1}{5} \log_2 \frac{1}{5} = 1.506$$

$$\text{ID}(D, \text{Mjesto}) = E(D) - \frac{1}{3} E(D_{\text{Istra}}) - \frac{1}{3} E(D_{\text{Kvarner}}) - \frac{1}{3} E(D_{\text{Dalmacija}}) =$$
$$= 1.506 - \frac{0.721}{3} - \frac{1.371}{3} - \frac{1.522}{3} = 0.301$$

Na sličan način računamo i preostale informacijske dobiti. Najveća je ona za atribut Cijena (1.106), pa stoga taj atribut postaje korijen stabla odluke. Za cijene 2000, 2500 i 9000 koristimo atribute Otok, Otok i Mjesto, za koje je informacijska dobit najveća i iznosi 1.



Za vrijednost Kvarner čvora Mjesto odabiremo klasu **ne**, jer je ta klasa najčešća za primjere za koje vrijedi Mjesto==Kvarner.

(c) Vrijednostima atributa pridijelio bih vjerojatnosti temeljene na relativnim frekvencijama poznatih primjera. Te bih vjerojatnosti zatim koristio za računanje informacijske dobiti.

(d) Algoritam je sklon prenaučenosti kada je skup za učenje malen ili sadrži šum, jer stablo odluke raste sve dok svi primjeri ne budu ispravno klasificirani.

Problem prenaučenosti možemo riješiti na više načina. Možemo prisilno zaustaviti rast stabla na određenoj razini prije nego ispravno klasificiramo sve primjere skupa za učenje ili naknadno podrezati čitavo izgrađeno stablo.

Drugim pristupom, koji je u praksi uspješniji, iz stabla uklanjamo svaki čvor koji, nakon što je uklonjen, ne utječe negativno na performanse stabla na skupu za vrednovanje. Pritom nam je potreban velik broj primjera za učenje.