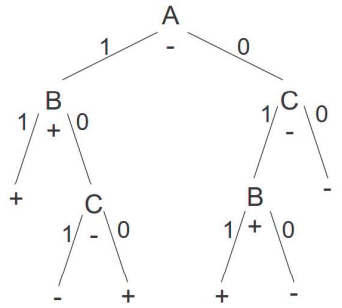
3.6 Rezanje: metoda zmanjševanja napake

Dano je klasifikacijsko drevo in rezalna tabela. Z rezanjem po metodi zmanjševanja napake poreži dano drevo.



A	B	C	R
0	0	1	−
0	0	1	+
0	1	1	−
0	0	0	−
0	1	0	−
0	1	1	−
0	1	1	−
0	1	1	−
0	0	1	+
1	1	1	+
1	1	0	+
1	0	1	+
1	0	1	+
1	0	0	+
1	0	1	−
1	0	0	−
1	0	0	−
1	0	0	−

Rešitev:

Rezanje klasifikacijskih dreves po metodi zmanjševanja napake (*angl. Reduced Error Prunning*) poteka od spodaj navzgor, t.j. od listov drevesa proti njegovemu korenu. Poleg drevesa postopek zahteva še rezalno tabelo podatkov. Pri gradnji drevesa se ti podatki ne upoštevajo, uporabljajo se samo pri rezanju. Z rezanjem začnemo v vozliščih tik nad listi. V vsakem vozlišču v izračunamo t.i. dobitek rezanja, $G(v)$, ki je definiran kot razlika med številom napačno klasificiranih primerov v poddrevesu T s korenom v v in številom napačnih klasifikacij pri v, če bi drevo tam porezali:

$$G(v) = \#\text{napak}_T - \#\text{napak}_v.$$

Režemo takrat, ko je $G(v) \geq 0$, torej če je napaka poddrevesa večja od napake v vozlišču v ali če sta napaki enaki, saj imamo ob enaki napaki raje manjše drevo.

Koristno je najprej vsakemu vozlišču drevesa pripisati porazdelitev razreda iz rezalne tabele (glej levo drevo na sliki 3.3). V oglatih oklepajih je najprej navedeno število primerov z razredom +, nato pa število primerov z razredom −. Sprotno preštevanje primerov hitro privede do napak. S pomočjo zgornje formule in izpisanih porazdelitev pa zlahka izračunamo dobitke rezanja za vsako vozlišče.

Začnimo pri vozlišču z atributom $(C, [3, 4], -)$. V tem vozlišču je torej 7 primerov, od tega 3 z razredom + in 4 z razredom −. Drevo v tem vozlišču klasificira v razred −. Poudarimo, da to ni vezano na porazdelitev primerov iz rezalne tabele, ampak je lastnost drevesa, ki je posledica porazdelitve v učnih podatkih, ki jih naloga ne navaja.

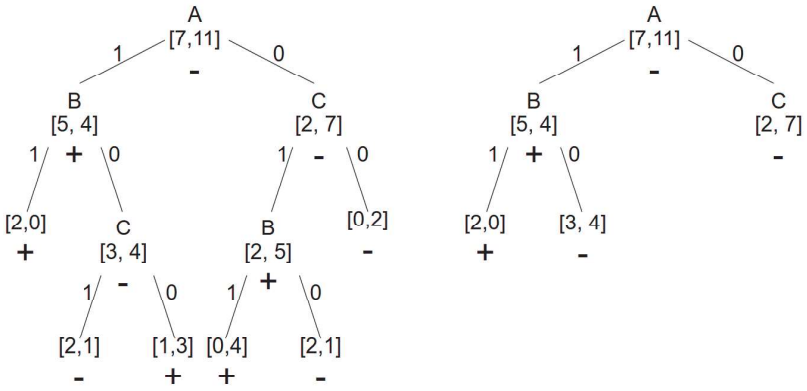
Število napak v tem vozlišču, $\#\text{napak}_v$, je 3, saj drevo napačno klasificira vse 3 primere z razredom +. Število napak v poddrevesu tega vozlišča je: $\#\text{napak}_T = 2 + 3 = 5$, 2 napaki v levem listu in 3 v desnem. Dobitek $G(v) = 5-3 = 2$, torej je dobro drevo porezati tako, da vozlišče $(C, [3, 4], -)$ postane list.

Naslednji kandidat za rezanje, po drevesu navzgor, je vozlišče $(B, [5, 4], +)$:
$\#\text{napak}_v = 4$
$\#\text{napak}_T = 0 + 3 = 3$
$G(v) = 3 - 4 = -1$
V tem vozlišču je bolje pustiti drevo tako kot je, kot ga porezati.

Nadaljujmo v vozlišču $(B, [2, 5], +)$:
$\#\text{napak}_v = 5$
$\#\text{napak}_T = 4 + 2 = 6$
$G(v) = 6 - 5 = 1$
Drevo tu porežemo in $(B, [2, 5], +)$ postane list.

Nadaljujemo v $(C, [2, 7], -)$:
$\#\text{napak}_v = 2$
$\#\text{napak}_T = 5 + 0 = 5$
$G(v) = 5 - 2 = 3$
Drevo porežemo in $(C, [2, 7], -)$ postane list.

Končno drevo je na sliki 3.3 desno.



Slika 3.3: Pomožno drevo (levo) in rešitev (desno).