Het leren van een beslissingsboom

L. Schomaker, juni 2001

Bij het leren van een beslissingsboom moeten we de variabele of het kenmerk (feature) met de grootste informatiewinst bepalen, vervolgens kijken wat gegeven een gereduceerde tabel dan het meest informatieve kenmerk is, en zo verder. Tabel 1 is een voorbeeld van een dergelijke tabel voor een probleem waarin we willen voorspellen aan de hand van de antwoorden van twee beoordelaars (reviewers) of ze een artikel (paper) zullen goedkeuren. De laatste kolom geeft de te leren respons (klasse) aan: accepteren Y of niet N. We noemen een dergelijke tabel in machinaal leren de *leerset (training set)*. Een observatie (rij van kenmerkwaarden) in een dergelijke tabel noemen we een kenmerkvector. Hoe gaat het berekenen van de informatiewinst van een kenmerk in zijn werk? We laten hieronder de index voor het kenmerk weg, maar vergeet niet dat we deze waarde voor elk kenmerk (kolom in de datamatrix) moeten berekenen:

$$I_{gain} = I_{decision} - I_{remainder} \tag{1}$$

Bij een beslissing met twee mogelijke uitkomsten waarbij beide uitkomsten even vaak voorkomen geldt:

$$I_{decision} = -\sum_{i=1}^{N_{classes}} p_i \, log_2(p_i) = 0.5 + 0.5 = 1 \tag{2}$$

Waarin $N_{classes}$ het aantal mogelijke uitkomsten van de beslissing is (hier 2: Yes+No).

De restinformatie $I_{remainder}$ die geleverd wordt door het te onderzoeken kenmerk is:

$$I_{remainder} = \sum_{i=1}^{N_{values}} p_i \, (-p_{i,ok} \, log_2(p_{i,ok}) - p_{i,err} \, log_2(p_{i,err})) \tag{3}$$

Waarin N_{values} het aantal mogelijke (symbolische) waarden voor het onderzochte kenmerk is (in het voorbeeld ook gelijk aan twee: $|\{Yes, No\}|$), p_i de kans op elk van die waarden, $p_{i,ok}$ de kans dat de waarde de beslissing correct voorspelt, $p_{i,err}$ de kans dat de waarde i van het onderzochte kenmerk de uitkomst foutief voorspelt. Zie ook laatste paragraaf.

Table 1: Two reviewers, 6 papers. How are they deciding?

Paper	L_A	L_B	M_A	M_B	T_A	T_B	Accept?
X1	N	Y	Y	Y	N	Y	Y
X2	Y	N	Y	Y	Y	Y	Y
X3	N	N	N	N	Y	N	N
X4	N	Y	N	N	Y	Y	Y
X5	Y	Y	N	Y	N	Y	N
X6	N	N	Y	Y	N	N	N

Kenmerk
L_A = leesbaarheid volgens reviewer A
M_A = methodologisch OK volgens reviewer A?
T_A = theoretisch OK volgens reviewer A?
L_B = leesbaarheid volgens reviewer B
M_B = methodologisch OK volgens reviewer B?
T_B = theoretisch OK volgens reviewer B?

Vraag: maak een beslissingsboom die voorspelt of een paper door de reviewers zal worden geaccepteerd.

Tabel 2 geeft de informatiewaarden voor de complete ruwe datamatrix uit Tabel 1. Voor de entropie $I(p, q)$ zie Tabel 3.

Table 2: Informatiewaarde van de features

Feature	Yes: I(ok,err)	No: I(ok,err)	Rem	1-Rem=Gain
L_A	$\frac{2}{6}I(\frac{1}{2},\frac{1}{2})$	$\frac{4}{6}I(\frac{2}{4},\frac{2}{4})$	1	1-1=0
L_B	$\frac{3}{6}I(\frac{2}{3},\frac{1}{3})$	$\frac{3}{6}I(\frac{1}{3},\frac{2}{3})$	$I(\frac{1}{3},\frac{2}{3}) = 0.91$	1-0.91=0.09
M_A	$\frac{3}{6}I(\frac{2}{3},\frac{1}{3})$	$\frac{3}{6}I(\frac{1}{3},\frac{2}{3})$	$I(\frac{1}{3},\frac{2}{3}) = 0.91$	1-0.91=0.09
M_B	$\frac{4}{6}I(\frac{2}{4},\frac{2}{4})$	$\frac{2}{6}I(\frac{1}{2},\frac{1}{2})$	1	1-1=0
T_A	$\frac{3}{6}I(\frac{2}{3},\frac{1}{3})$	$\frac{3}{6}I(\frac{2}{3},\frac{1}{3})$	$I(\frac{1}{3},\frac{2}{3}) = 0.91$	1-0.91=0.09
T_B	$\frac{4}{6}I(\frac{3}{4},\frac{1}{4})$	$\frac{2}{6}I(\frac{2}{2},\frac{0}{0})$	0.66*0.81+0.33*0=0.53	1-0.53=0.47 (!)

Dus: T_B is het eerste feature om te onderzoeken: levert volledige zekerheid over de voorbeelden X3 en X6. Deze rijen worden verwijderd uit de matrix, de kolom van T_B wordt verwijderd, en met het restant gaan we de volgende knoop uitzoeken.

Table 3: Informatiewaarden (entropie) van een aantal kansen p in $[bits]$

p	1-p	I
0.05	0.95	0.29
0.10	0.90	0.47
0.15	0.85	0.61
0.20	0.80	0.72
0.25	0.75	0.81
0.30	0.70	0.88
0.33	0.67	0.91
0.35	0.65	0.93
0.40	0.60	0.97
0.45	0.55	0.99
0.50	0.50	1.00

Table 4: Two reviewers problem, reduced: minus two rows (X3,X6), minus one column (T_B)

Paper	L_A	L_B	M_A	M_B	T_A	-	Accept?
X1	N	Y	Y	Y	N	-	Y
X2	Y	N	Y	Y	Y	-	Y
X3	-	-	-	-	-	-	-
X4	N	Y	N	N	Y	-	Y
X5	Y	Y	N	Y	N	-	N
X6	-	-	-	-	-	-	-

Table 5: Informatiewaarde van de features, knoop 2

Feature	Yes: I(ok,err)	No: I(ok,err)	Rem	1-Rem=Gain
L_A	$\frac{2}{4}I(\frac{1}{2},\frac{1}{2})$	$\frac{2}{4}I(\frac{0}{0},\frac{2}{2})$	0.5	1-0.5=0.5
L_B	$\frac{3}{4}I(\frac{2}{3},\frac{1}{3})$	$\frac{1}{4}I(\frac{0}{0},\frac{1}{1})$	0.75*0.91	1-0.68=0.32
M_A	$\frac{2}{4}I(\frac{2}{2},\frac{0}{0})$	$\frac{2}{4}I(\frac{1}{2},\frac{1}{2})$	0.5	1-0.5=0.5
M_B	$\frac{3}{4}I(\frac{2}{3},\frac{1}{3})$	$\frac{1}{4}I(\frac{0}{0},\frac{1}{1})$	0.75*0.91	1-0.68=0.32
T_A	$\frac{2}{4}I(\frac{2}{2},\frac{0}{0})$	$\frac{2}{4}I(\frac{1}{2},\frac{1}{2})$	0.5	1-0.5=0.5

We kunnen nu kiezen uit L_A, M_A en T_A (elk 0.5 bit). Als we T_A kiezen kunnen we een correcte Yes! opleveren voor observaties X2 en X4. Blijven over X1 en X5, die we met de laatste knoop kunnen oplossen, dat kan met: M_A (Yes $\rightarrow$ Yes!, No $\rightarrow$ No!).

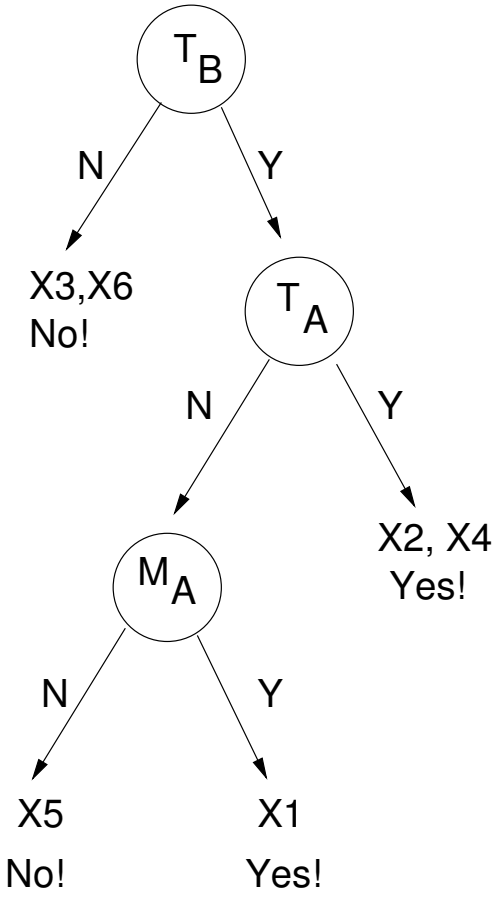

Figure 1: De resulterende beslissingsboom

Figuur 1 geeft de resulterende beslissingsboom, met per beslissing de data ('papers') uit de leerset die gebruikt zijn.

De volgende fase is om een onafhankelijke test set te nemen en te kijken of nog ongeziene artikelen ook goed geklassificeerd worden. Het systeem maakt een aantal fouten waarvoor je correcties kunt aanbrengen (pruning). Je hebt vervolgens weer een onbesmette test set nodig om de prestaties van de beslissingsboom objectief te kunnen beoordelen. Hoewel beslissingsbomen erg nuttig kunnen zijn (wat er gebeurt is volledig expliciet), is in de praktijk gebleken dat de prestaties achterblijven bij andere methoden (neurale netwerken, statistische klassificatie, of soms zelfs de simpele Euclidische afstandsmaat: '(k) nearest neighbour', kNN). De reden is gelegen in het fenomeen 'premature commitment'. Gedurende het sequentiele beslissingsproces kun je niet herstellen van een eerder genomen foute beslissing. Het gebruik van de informatie in de klasseruimte is bij de beslissingsboom lokaal, terwijl een neural net of een kNN alle informatie in de kenmerkvector simultaan gebruiken om de meest waarschijnlijke respons (klasse) op te leveren.

Algemeen

Als het beslissingsprobleem zowel meerdere waarden voor de kenmerken (attributen,"features") omvat,
als meerdere klassen (uitspraken) in de beslissing,
dan is de algemene vergelijking voor de restinformatie:

$$I_{remainder} = \sum_{i=1}^{N_{values}} \left(p_i \sum_{j=1}^{N_{classes}} -p_{(i\&j)}\, log_2(p_{(i\&j)}) \right) \tag{4}$$