Chosen data mining problems e.g. decision trees

dr Piotr Wąsiewicz

1. From the training set shown below in the table with the help of top-down decision tree induction algorithm create a decision tree (use the entropy cost). The attribute age should be discretized using two threshholds 30 and 65 years. The attribute risk will be the label class.

x	age	car	risk
1	18	mini	big
2	35	mini	small
3	50	racer	big
4	66	van	big
5	18	racer	big
6	35	van	small
7	60	mini	small
8	70	racer	big
9	25	van	small

Solution:
The attribute age has three values after discretization:
w_1: age < 30, w_2: age $\geqslant 30 \wedge$ age < 65, w_3: age $\geqslant 65$.
First the information of the whole set I(P) and attributes is calculated using the entropy impurity measures.

$$I(P) = -\frac{|P^{small}|}{|P|}\log_2(\frac{|P^{small}|}{|P|}) - \frac{|P^{big}|}{|P|}\log_2(\frac{|P^{big}|}{|P|}) = -\frac{4}{9}\log_2(\frac{4}{9}) - \frac{5}{9}\log_2(\frac{5}{9}) = 0.991,$$

$$E_{\mathsf{age,w_1}}(P) = -\frac{|P^{small}_{\mathsf{age,w_1}}|}{|P_{\mathsf{age,w_1}}|}\log_2(\frac{|P^{small}_{\mathsf{age,w_1}}|}{|P_{\mathsf{age,w_1}}|}) - \frac{|P^{big}_{\mathsf{age,w_1}}|}{|P_{\mathsf{age,w_1}}|}\log_2(\frac{|P^{big}_{\mathsf{age,w_1}}|}{|P_{\mathsf{age,w_1}}|}) = -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) = 0.918,$$

$$E_{\mathsf{age,w_2}}(P) = -\frac{|P^{small}_{\mathsf{age,w_2}}|}{|P_{\mathsf{age,w_2}}|}\log_2(\frac{|P^{small}_{\mathsf{age,w_2}}|}{|P_{\mathsf{age,w_2}}|}) - \frac{|P^{big}_{\mathsf{age,w_2}}|}{|P_{\mathsf{age,w_2}}|}\log_2(\frac{|P^{big}_{\mathsf{age,w_2}}|}{|P_{\mathsf{age,w_2}}|}) = -\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4}) = 0.811,$$

$$E_{\mathsf{age,w_3}}(P) = -\frac{|P^{small}_{\mathsf{age,w_3}}|}{|P_{\mathsf{age,w_3}}|}\log_2(\frac{|P^{small}_{\mathsf{age,w_3}}|}{|P_{\mathsf{age,w_3}}|}) - \frac{|P^{big}_{\mathsf{age,w_3}}|}{|P_{\mathsf{age,w_3}}|}\log_2(\frac{|P^{big}_{\mathsf{age,w_3}}|}{|P_{\mathsf{age,w_3}}|}) = -\frac{0}{2}\log_2(\frac{0}{2}) - \frac{2}{2}\log_2(\frac{2}{2}) = 0,$$

$$E_{\mathsf{car,mini}}(P) = -\frac{|P^{small}_{\mathsf{car,mini}}|}{|P_{\mathsf{car,mini}}|}\log_2(\frac{|P^{small}_{\mathsf{car,mini}}|}{|P_{\mathsf{car,mini}}|}) - \frac{|P^{big}_{\mathsf{car,mini}}|}{|P_{\mathsf{car,mini}}|}\log_2(\frac{|P^{big}_{\mathsf{car,mini}}|}{|P_{\mathsf{car,mini}}|}) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.918,$$

$$E_{\mathsf{car,van}}(P) = -\frac{|P^{small}_{\mathsf{car,van}}|}{|P_{\mathsf{car,van}}|}\log_2(\frac{|P^{small}_{\mathsf{car,van}}|}{|P_{\mathsf{car,van}}|}) - \frac{|P^{big}_{\mathsf{car,van}}|}{|P_{\mathsf{car,van}}|}\log_2(\frac{|P^{big}_{\mathsf{car,van}}|}{|P_{\mathsf{car,van}}|}) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.918,$$

$$E_{\mathsf{car,racer}}(P) = -\frac{|P^{small}_{\mathsf{car,racer}}|}{|P_{\mathsf{car,racer}}|}\log_2(\frac{|P^{small}_{\mathsf{car,racer}}|}{|P_{\mathsf{car,racer}}|}) - \frac{|P^{big}_{\mathsf{car,racer}}|}{|P_{\mathsf{car,racer}}|}\log_2(\frac{|P^{big}_{\mathsf{car,racer}}|}{|P_{\mathsf{car,racer}}|}) = -\frac{0}{3}\log_2(\frac{0}{3}) - \frac{3}{3}\log_2(\frac{3}{3}) = 0,$$

Next the weighted entropies are obtained:

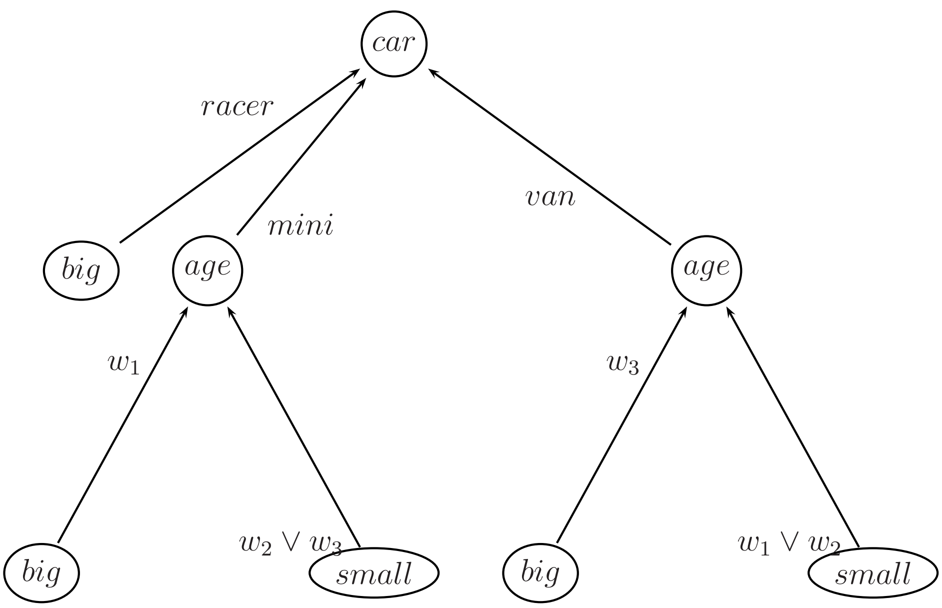
$E_{\mathsf{age}}(P) = \frac{|P_{\mathsf{age,w_1}}|}{|P|} E_{\mathsf{age,w_1}}(P) + \frac{|P_{\mathsf{age,w_2}}|}{|P|} E_{\mathsf{age,w_2}}(P) + \frac{|P_{\mathsf{age,w_3}}|}{|P|} E_{\mathsf{age,w_3}}(P) = \frac{3}{9}(0.918) + \frac{4}{9}(0.811) + \frac{2}{9}0 = 0,666,$
$E_{\mathsf{car}}(P) = \frac{|P_{\mathsf{car,mini}}|}{|P|} E_{\mathsf{car,mini}}(P) + \frac{|P_{\mathsf{car,van}}|}{|P|} E_{\mathsf{car,van}}(P) + \frac{|P_{\mathsf{car,racer}}|}{|P|} E_{\mathsf{car,racer}}(P) = \frac{3}{9}(0.918) + \frac{3}{9}(0.918) + \frac{3}{9}0 = 0,612,$

And all attributes information measures:

$IV_{\mathsf{age}}(P) = -\frac{|P_{\mathsf{age,w_1}}|}{|P|} \log_2(\frac{|P_{\mathsf{age,w_1}}|}{|P|}) - \frac{|P_{\mathsf{age,w_2}}|}{|P|} \log_2(\frac{|P_{\mathsf{age,w_2}}|}{|P|}) - \frac{|P_{\mathsf{age,w_3}}|}{|P|} \log_2(\frac{|P_{\mathsf{age,w_3}}|}{|P|}) = -\frac{3}{9}\log_2(\frac{3}{9}) - \frac{4}{9}\log_2(\frac{4}{9}) - \frac{2}{9}\log_2(\frac{2}{9}) = 0,528 + 0,519 + 0,482 = 1,53,$
$IV_{\mathsf{car}}(P) = -\frac{|P_{\mathsf{car,mini}}|}{|P|} \log_2(\frac{|P_{\mathsf{car,mini}}|}{|P|}) - \frac{|P_{\mathsf{car,van}}|}{|P|} \log_2(\frac{|P_{\mathsf{car,van}}|}{|P|}) - \frac{|P_{\mathsf{car,racer}}|}{|P|} \log_2(\frac{|P_{\mathsf{car,racer}}|}{|P|}) = -\frac{3}{9}\log_2(\frac{3}{9}) - \frac{3}{9}\log_2(\frac{3}{9}) - \frac{3}{9}\log_2(\frac{3}{9}) = 0,528 + 0,528 + 0,528 = 1,584,$

At last information increase coefficients:

$$\vartheta_{\mathsf{age}}(P) = \frac{I(P) - E_{\mathsf{age}}(P)}{IV_{\mathsf{age}}(P)} = \frac{0,991 - 0,666}{1,53} = 0,212$$
$$\vartheta_{\mathsf{car}}(P) = \frac{I(P) - E_{\mathsf{car}}(P)}{IV_{\mathsf{car}}(P)} = \frac{0,991 - 0,612}{1,584} = 0,239$$



The attribute car has the biggest coefficient and wins to be the first node.
For its value racer every example with this value has the label big of the target risk.