
INSTITUT TEKNOLOGI BANDUNG
1920

Modul : Issues in Decision Tree Learning (DTL)

Alternative Measures for Selecting Attribute

Nur ULFA Maulidevi

KK IF - Teknik Informatika- STEI ITB

Pembelajaran Mesin
(Machine Learning)

Issues in DTL

Overfitting training data

Continuous -valued attribute

Handling attributes with differing costs

Handling missing attribute value

Alternative measures for selecting attributes

Attribute with many values (C4.5)

Gain will always select it → example *Date=2021_Jan_31*

Date will perfectly classify training examples,
but very poor predictor for unseen data

GAIN RATIO

$$GainRatio(S,A) \equiv \frac{Gain(S,A)}{SplitInformation(S,A)}$$

$$SplitInformation(S,A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

where S_i is subset of S for which A has value v_i

Illustration

Date	Atr2	Atr3	Class
2021_Jan_01	v1		No
2021_Jan_02	v1		No
2021_Jan_03	v2		Yes
2021_Jan_04	v2		Yes
2021_Jan_05	v1		Yes
2021_Jan_06	v1		No

$$\text{SplitInformation(S,Date)} = -\sum_{i=1}^{6} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$$

$$= -\left(\frac{1}{6} log_2 \frac{1}{6} + \frac{1}{6} log_2 \frac{1}{6} + \frac{1}{6} log_2 \frac{1}{6} + \frac{1}{6} log_2 \frac{1}{6} + \frac{1}{6} log_2 \frac{1}{6} + \frac{1}{6} log_2 \frac{1}{6}\right)$$

$$\text{SplitInformation(S,Atr2)} = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$$

$$= -\left(\frac{4}{6} log_2 \frac{4}{6} + \frac{2}{6} log_2 \frac{2}{6}\right)$$

What if SplitInformation is very small or zero (|Si|≈ |S|) → GainRatio undefined or very large

Heuristic: Apply GainRatio test only for Attribute with above average Gain

THANK YOU


INSTITUT TEKNOLOGI BANDUNG
1920