
INSTITUT TEKNOLOGI BANDUNG
1920

Modul : Issues in Decision Tree Learning (DTL)

Attributes with Differing Costs

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 Different Cost

Attributes: Temperature, BiopsyResult, Pulse, BloodTestResults

Have Different Cost (monetary and patient comfort)

Use low cost attribute where possible, high cost only when required to produce reliable classification

Cost is considered in calculating Gain of each attribute

Approaches

Tan and Schlimmer (1990) and Tan (1993):

$$\frac{Gain^2(S,A)}{Cost(A)}$$

Nunez (1988):

$$\frac{2^{Gain(S,A)}-1}{(Cost(A)+1)^w}$$

Where $w \in [0,1]$ determine importance of cost

Exercise

Outlook	Temp	Humidity	Windy	Class
sunny	75	70	TRUE	Play
sunny	80	90	TRUE	Don't Play
sunny	85	85	FALSE	Don't Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
?	72	90	TRUE	Play
overcast	83	78	FALSE	Play
overcast	64	65	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play
rain	65	70	TRUE	Don't Play
rain	75	80	FALSE	Play
rain	68	80	FALSE	Play
rain	70	96	FALSE	Play

1. What is GainRatio for Outlook?
2. What are Examples (instances) for Outlook = sunny? (if "Outlook" is the root)
3. Based on the result of question (2), if the next attribute is "Humidity" with threshold <= 75 and >75, illustrate the Decision Tree for that branch (assumption: stop learning after "Humidity" is selected)
4. If we have unseen data: <Outlook = sunny, Temp = 70, Humidity=?, and Windy =FALSE>, what is the class prediction? (Based on the result of question (3)).

THANK YOU


INSTITUT TEKNOLOGI BANDUNG
1920