
INSTITUT TEKNOLOGI BANDUNG
1920

Modul : Issues in Decision Tree Learning (DTL)

Continuous-valued 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

Discretization

Continuous valued attributes → new discrete valued (boolean) attribute A_c

True: A < c

False: A < c (or A ≥ c)

Temperature:	40	48	60	72	80	90
PlayTennis:	No	No	Yes	Yes	Yes	No

Potential optimal breakpoints

What is Best Value for threshold c?

C = (48+60)/2 = 54
Or
C = (80+90)/2 = 85

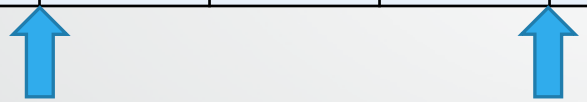
Use Information Gain for each potential breakproint

Illustration

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1		72			Yes
D2		40			No
D3		90			No
D4		60			Yes
D5		48			No
D6		80			Yes

1. Sort The Continuous-valued attribute

Temperature	40	48	60	72	80	90
Play Tennis	No	No	Yes	Yes	Yes	No



2. Identify Adjacent examples that differ in their target class

3. Candidates: midway between corresponding values → C : 54 or C : 85

4. Find the greatest Gain from the candidates, and other discrete-valued attributes

For C: 54
Temperature < 54: 2 examples → yes/0, no/2
Temperature ≥ 54: 4 examples → yes/3, no/1
$Gain(S,T_{54}) = Entropi(S) – [(2/6*Entropi(0,2))+(4/6*Entropi(3,1)]$

For C: 85
Temperature < 85: 5 examples → yes/3, no/2
Temperature ≥ 85: 1 examples → yes/0, no/1
$Gain(S,T_{85}) = Entropi(S) – [(5/6*Entropi(3,2))+(1/6*Entropi(0,1)]$

THANK YOU


INSTITUT TEKNOLOGI BANDUNG
1920