
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)).

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

Termasuk isu apa ?

Continuous-valued attribute

Handling missing attribute value

Handling attributes with differing costs

Alternative measures for selecting attributes

Overfitting training data

1. **What is GainRatio for Outlook?**
2. What are Examples (instances) for Outlook = sunny?
3. Based on the result of question (2), define the threshold in "Humidity" discretization, and the leaf node for each branch.

Gain(S,A) only consider the fraction of training examples with known value

Gain(S,A) = 10/11 * (Entropy(S) – [Σproportion*entropy_of_known_value])

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

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

set of

value v_i

CLSR-FIPM

CLSR-FIPM

Tanpa missing values

$$Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.940$$

5 $$Gain(S, Outlook) = Entropy(S) - \sum_{v \in value\ of\ Outlook} \frac{S_v}{S} Entropy(S_v)$$

$$= 0.940 - [\frac{5}{14}Entropy(Sunny) + \frac{4}{14}Entropy(Overcast) + \frac{5}{14}Entropy(Rain)]$$

$$= 0.940 - [(\frac{5}{14}(-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5})) + (\frac{4}{14}(-\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4})) + (\frac{5}{14}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}))] = 0.246$$

Handling missing attribute value

Gain(S,A) = 13/14 * (Entropy(S) – [Σproportion*entropy_of_known_value])

$$Entropy(S) = -\frac{8}{13}log_2\frac{8}{13} - \frac{5}{13}log_2\frac{5}{13} = \mathbf{0.961}$$

$$Gain(S, Outlook) = 13/14 * (0.961 - [\frac{5}{13}Entropy(Sunny) + \frac{3}{13}Entropy(Overcast) + \frac{5}{13}Entropy(Rain)])$$

$$= 13/14 * (0.961 - [\frac{5}{13}(-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}) + \frac{3}{13}(-\frac{3}{3}log_2\frac{3}{3} - \frac{0}{3}log_2\frac{0}{3}) + \frac{5}{13}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5})])$$

$$= 13/14 * (0.961 - [\frac{5}{13}0.971 + \frac{3}{13}0 + \frac{5}{13}0.971]) = 0.199$$

$$SplitInformation(S, Outlook) = -\sum_{i=1}^{4}\frac{|S_i|}{|S|}log_2\frac{|S_i|}{|S|} = -(\frac{5}{14}log_2\frac{5}{14} + \frac{3}{14}log_2\frac{3}{14} + \frac{5}{14}log_2\frac{5}{14} + \frac{1}{14}log_2\frac{1}{14}) = 1.809$$

$$GainRatio(S, Outlook) = \frac{Gain(S, Outlook)}{SplitInformation(S, Outlook)} = \frac{0.199}{1.809} = 0.110$$

EDUNEX ITB

CLSR-FIPM

Exercise

6

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?**
3. Based on the result of question (2), define the threshold in "Humidity" discretization, and the leaf node for each branch.

Outlook	Temp	Humidity	Windy	Class	Weight
sunny	75	70	TRUE	Play	1
sunny	80	90	TRUE	Don't Play	1
sunny	85	85	FALSE	Don't Play	1
sunny	72	95	FALSE	Don't Play	1
sunny	69	70	FALSE	Play	1
?	72	90	TRUE	Play	5/13

Termasuk isu apa ?

Alternative measures for selecting attributes

Overfitting training data

Continuous-valued attribute

Handling missing attribute values

Handling attributes with differing costs

EDUNEX ITB

Toolbar

L) missing value selalu masuk, kalaupun dia tanya outlook: overcast, dll missing value jg masuk tp bobotnya berbeda.

Exercise

7

Outlook	Temp	Humidity	Windy	Class	Weight
sunny	75	70	TRUE	Play	1
sunny	80	90	TRUE	Don't Play	1
sunny	85	85	FALSE	Don't Play	1
[illegible]	[illegible]	95	FALSE	Don't Play	1
		70	FALSE	Play	1
		90	TRUE	Play	5/13

Humidity	Class
70	Play
70	Play
85	Don't Play
90	Don't Play
90	Play
95	Don't Play

1. What is GainRatio for Outlook?
2. What are Examples (instances) for Outlook = sunny?
3. **Based on the result of question (2), define the threshold in "Humidity" discretization, and the leaf node for each branch.**

⇒ Threshold = (70+85)/2=77.5

...

Outlook=sunny:

-- humidity < 77.5 = True: Play

-- humidiy < 77.5 = False: <5/13 Play, 3 Don't Play>

THANK YOU


INSTITUT TEKNOLOGI BANDUNG
1920

LATIHAN MINGGU 4

Jika terdapat training examples sebagai berikut, pembatas (threshold) yang mungkin diperiksa untuk attr2 adalah:

sorting nilai attr2 dulu

Instance	attr1	attr2	attr3	Label
1	v1_1	8	v3_1	0
2	v1_2	5	v3_1	1
3	v1_1	5	v3_2	1
4	v1_2	7	v3_1	1
5	v1_3	6	v3_1	0
6	v1_3	4	v3_2	0
7	v1_3	5	v3_1	1
8	v1_1	8	v3_1	0
9	v1_2	5	v3_2	1
10	v1_3	6	v3_1	0

a. 1 pembatas yaitu (attr2 ≤ 6 dan attr2 > 6)

b. 2 pembatas yaitu (attr2 ≤ 5 dan attr2 > 5), (attr2 ≤ 7 dan attr2 > 7)

c. 3 pembatas yaitu (attr2 ≤ 5 dan attr2 > 5), (attr2 ≤ 6 dan attr2 > 6), (attr2 ≤ 7 dan attr2 > 7)

d. 4 pembatas yaitu (attr2 ≤ 4 dan attr2 > 4), (attr2 ≤ 5 dan attr2 > 5), (attr2 ≤ 6 dan attr2 > 6), (attr2 ≤ 7 dan attr2 > 7)

Semua pilihan a, b, c, d ini salah. cuman yg paling mendekati = d. cuman batasnya harusnya 4.5, 5.5, 6.5, dll (lihat slide bawah)

LATIHAN MINGGU 4

2. Jika terdapat training examples sebagai berikut, pembatas (threshold) yang mungkin diperiksa untuk attr2 adalah:

Instance	attr1	attr2	attr3	Label
1	v1_1	8	v3_1	0
2	v1_2	5	v3_1	1
3	v1_1	5	v3_2	1
4	v1_2	7	v3_1	1
5	v1_3	6	v3_1	0
6	v1_3	4	v3_2	0
7	v1_3	5	v3_1	1
8	v1_1	8	v3_1	0
9	v1_2	5	v3_2	1
10	v1_3	6	v3_1	0

Attr2	Label
4	0
5	1
5	1
5	1
5	1
6	0
6	0
7	1
8	0
8	0

$\frac{4+5}{2} = 4.5$

Kandidat 1:
4.5

Kandidat 2:
5.5

Kandidat 3:
6,5

Kandidat 4:
7.5