Decision Tree Learning

(Part-II)

BY:

DR. VIPIN KUMAR

DEPARTMENT OF COMPUTER SCIENCE &IT

MAHATMA GANDHI CENTRAL UNIVERSITY

MOTIHARI, BIHAR

Outline…

- DESCRIBE THE INDUCTIVE BIAS OF ID3
- ADVANTAGES OF ID3
- DISADVANTAGES OF ID3
- C4.5 ALGORITHM
- ADVANTAGE OF C4.5 ALGORITHM
- DISADVANTAGES OF C4.5 ALGORITHM

ID3 : Inductive Bias

- Inductive bias is the policy by which ID3 generalizes from observed training instances to classify unseen instance.
- There are two inductive bias in ID3:
  - Approximate inductive bias of ID3
  - A closer approximation to the inductive bias of ID3

Advantages of ID3

1. Evident prediction rules are constructed form the training data.
2. It builds the short tree.
3. It searches entire dataset to create the tree.
4. It searches complete hypothesis space to predict unlabeled instances.
5. It is less sensitive toward errors of individual training examples because of statistical properties of instances are utilized.

Disadvantages of ID3

- Over-fitting of the data may happen while classification.
- It does not perform backtracking while searching
- It may converge in locally optimal solution.
- Computational complexity may be vey high for the continuous data.

C4.5 algorithm

(Quinlan 1986)

- It uses Gain Ratio measure for selecting the decision attribute.
- It sensitive toward, how uniformly and broadly the attribute splits the data.
- Split Information can be written as:

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

C4.5 algorithm

(Quinlan 1986)

- Then Gain Ratio can be define as:

$$GainRatio(S,A) = \frac{IG(S,A)}{SplitInfo(S,A)} \quad \ldots\ldots\ldots(2)$$

C4.5 Advantages over ID3 algorithm

- Its handle the continuous and discrete features.
- Handling missing attribute values. (missing values of the attribute are not considered while information gain and entropy calculation).
- Handling attributes with different cost.
- Its prune tree after creation.

Assignment-2: Construct the Decision tree using C4.5 algorithm

RID	Age	Income	Student	Credit-rating	Class: buys computer
1	youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle_aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle_aged	Medium	No	Excellent	Yes
13	Middle_aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

Bibliography

- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.
- Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
- Mitchell, Tom M. "Machine learning." (1997).
- Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2020.
- Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Burkov, Andriy. *The hundred-page machine learning book*. Quebec City, Can.: Andriy Burkov, 2019.

- Burkov, Andriy. *The hundred-page machine learning book*. Quebec City, Can.: Andriy Burkov, 2019.

Thank You