Data Mining

Classification IV - Random Forests (Part A)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

/

Overview

- In the training phase, a number of Classification and Regression Trees (CART, a binary decision tree) will be generated. User can specify how many trees are going to grow.
- In the testing phase, a test sample will be classified by majority votes from the CART decision trees.

Tree Growing Algorithm (1)

- Suppose the number of training records is N. Randomly pick records N times with replacement (repeatedly picking the same record is allowed).
- According to $(1 - 1/N)^N = 1/e = 0.368$ when $N \rightarrow \infty$, about 63.2% of training records will be picked to grow each tree.
- The remaining set with about 36.8% of training data will be used for error rate estimation.

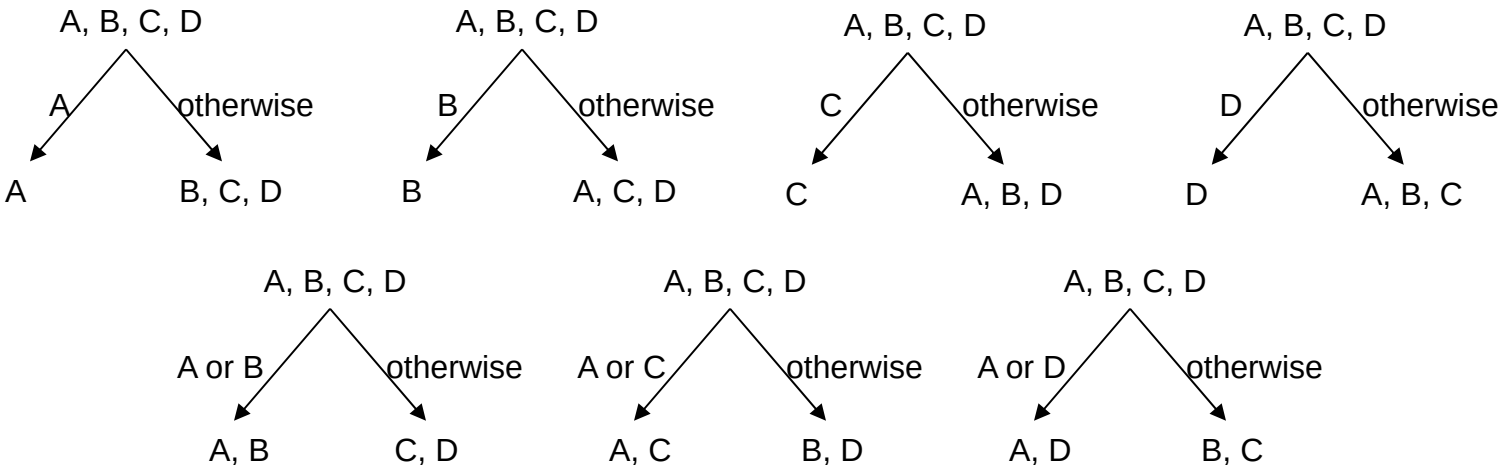
Tree Growing Algorithm (2)

- Suppose the number of attributes in each training record is M. When splitting each node, we randomly pick $\sqrt{M}$ attributes and examine each picked attribute. The best split among the picked attributes is used to split the node.
- The best split is determined by gini impurity measure.

Tree Growing Algorithm (3)

- Suppose an attribute is a categorical variable with n different categories. There are $2^{n-1}-1$ possible splits.

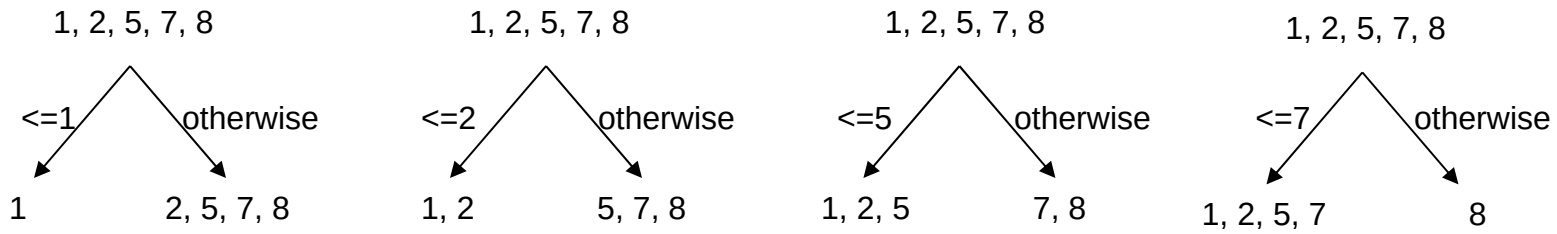
E.g. There are 4 different categories, A, B, C and D.
Then there will be $2^{4-1}-1=7$ possible splits as follows:



Tree Growing Algorithm (4)

- Suppose an attribute is a numerical variable with n different values associated with a comparison operator. There are n – 1 possible splits.

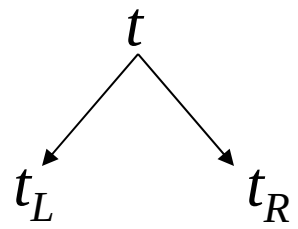
E.g. There are 5 different values, 1, 2, 5, 7 and 8.
Then there will be 5 – 1 = 4 possible splits as follows:



Tree Growing Algorithm (5)

- **What is the gini impurity measure?**

Suppose there are m classes in node t.
The gini impurity measure for t is:



$$g(t) = 1 - \sum_{i=1}^{m} f_i^2$$

where f_i is the fraction of class i among all training records in t. If there is only one class in node t, then $g(t)$ is zero; otherwise, $g(t)$ is greater than zero.

- **How to determine the best split?**

$$\Delta g(s,t) = g(t) - P_L g(t_L) - P_R g(t_R)$$

$$s^* \leftarrow \arg\max_{s}(\Delta g(s,t))$$

where s is a split, P_L and P_R are the proportion of training records assigned to t_L and t_R respectively according to s.
s^* is the best split among all possible splits.

End of
Random Forests Module (Part A)