成绩

北京航空航天大學
BEIHANG UNIVERSITY

Experiments for “Pattern Recognition and Machine Learning”

Experiment 3
Decision Tree Learning for Classification

院（系）名称 自动化科学与电气工程学院

专业名称 自动化

学生学号 15071135

学生姓名 刘雨鑫

2018年5月21日

3.1 Introduction

Decision tree induction is one of the simplest and yet most successful learning algorithms. A decision tree (DT) consists of internal and external nodes and the interconnections between nodes are called branches of the tree. An internal node is a decision-making unit to decide which child nodes to visit next depending on different possible values of associated variables. In contrast, an external node also known as a leaf node, is the terminated node of a branch. It has no child nodes and is associated with a class label that describes the given data. Adecision tree is a set of rules in a tree structure, each branch of which can be interpreted as a decision rule associated with nodes visited along this branch.

3.2 Principle and Theory

Decision trees classify instances by sorting them down the tree from root to leaf nodes. This tree-structured classifier partitions the input space of the data set recursively into mutually exclusive spaces. Following this structure, each training data is identified as belonging to a certain subspace, which is assigned a label, a value, or an action to characterize its data points. The decision tree mechanism has good transparency in that we can follow a tree structure easily in order to explain how a decision is made. Thus interpretability is enhanced when we clarify the conditional rules characterizing the tree.

Entropy of a random variable is the average amount of information generated by observing its value. Consider the random experiment of tossing a coin with probability of heads equal to 0.9, so that P(Head) = 0.9 and P(Tail) = 0.1. This provides more information than the case where P(Head) = 0.5 and P(Tail) = 0.5.

Entropy is used to evaluate randomness in physics, where a large entropy value indicates that the process is very random. The decision tree is guided heuristically according to the information content of each attribute. Entropy is used to evaluate the information of each attribute; as a means of classification. Suppose we have m classes, for a particular attribute, we denoted it by pi by the proportion of data which belongs to class Ci where i = 1, 2, … m.

The entropy of this attribute is then:

$$Entropy = \sum_{i=1}^{m} - p_i \cdot \log_2 p_i$$

We can also say that entropy is a measurement of the impurity in a collection of training examples: larger the entropy, the more impure the data is. Based on entropy, Information Gain (IG) is used to measure the effectiveness of an attribute as a means of discriminating between classes.

$$IG(S, A) = Entropy(S) - \sum_{v=Valuse(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where all examples S is divided into several groups (i.e. Sv for v ∈ Values(A)) according to the value of A. It is simply the expected reduction of entropy caused by partitioning the examples according to this attribute.

3.3 Objective

(1) To understand why we use entropy-based measure for constructing a decision tree.

(2) To understand how Information Gain is used to select attributes in the process of building a decision tree.

(3) To understand the equivalence of a decsion tree to a set of rules.

(4) To understand why we need to prune the tree sometimes and how can we prune? Based on what mesure we prune a decsion tree.

(5) To understand the concept of Soft Decsion Treees and why they are imporant extensions to classical decision trees.

3.4 Contents and Procedure

(1) consider the case of continuous attributes

Because the number of values available for continuous attributes is no longer limited, we can not build the tree with the values of continuous attributes. I think using the bi-partition is the simplest way to do discretization, which is also used in the C4.5 algorithm.

First we suppose that there is a class D and the attribute a, the attribute a has n different values in class D. We sort these values from big to small, get the aggregate

$$\{a^1, a^2, \ldots, a^n\}$$

Then we get n-1 partition elements aggregate

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \mid 1 \le i \le n-1 \right\}$$

So we use the middle site $\frac{a^i + a^{i+1}}{2}$ in $\left[a^i, a^{i+1}\right)$ as candidate partitioning points. Then we can deal it like the discrete attributes with these partitioning points. The Information Gain (IG) can be used as follows.

$$Gain(D,a) = \max_{t \in T_a} Gain(D,a,t) = \max_{t \in T_a} Entropy(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} Entropy(D_t^\lambda)$$

Gain(D,a) is the Information Gain based on bi-partition of the division point *t*. So we can choose the division boundary to get the maximal Information Gain.

(2) finish the code and build the decision tree

First I write the code according the computational formula of the entropy and the information gain of the attribute, while considering the case of continuous attributes.

$$Entropy = \sum_{i=1}^{m} - p_i \cdot \log_2 p_i$$

$$Gain(D,a) = \max_{t \in T_a} Gain(D,a,t) = \max_{t \in T_a} Entropy(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} Entropy(D_t^\lambda)$$

Then I use the continuous attributes data 'Iris Plants Database' from the UCI Machine Learning Repository, and the attributes information are shown in table 1.

Table 1 attributes using to build the decision tree

#	Attribute	domain
1	sepal length	4.3 - 7.9
2	sepal width	2.0 - 4.4
3	petal length	1.0 - 6.9
4	petal width	0.1 - 2.5
5	class	Iris Setosa, Iris Versicolour, Iris Virginica

Because it has more than two attributes, I decide to use the binary tree structure to represent a decision tree, divide the samples in two parts, and finish the code. I use 150 samples to run the program and the result is shown in fig.1.

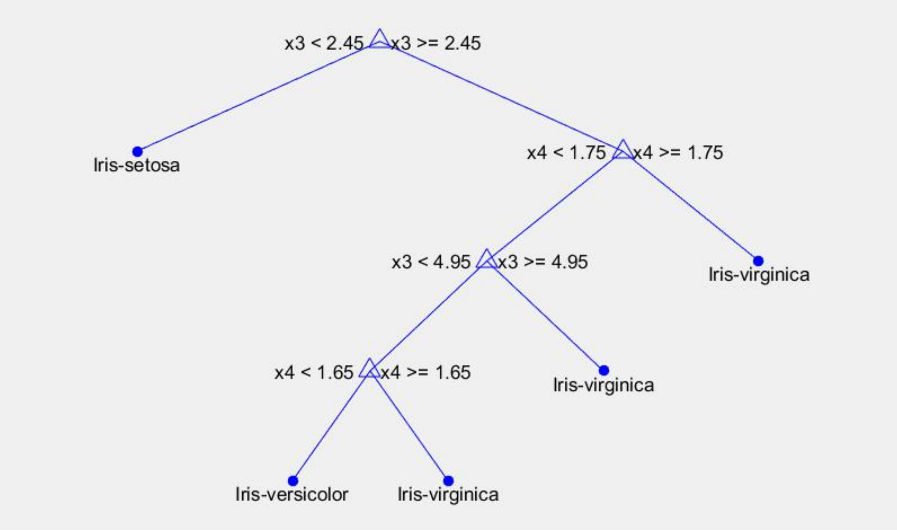

Fig.1. the decision tree (1)

The x1, x2, x3 and x4 in the fig.1 means the attributes shown in the table 1, sepal length, sepal width, petal length and petal width. The decision tree is built successfully.

(3) Is there a tradeoff between the size of the tree and the model accuracy?

Look at the result in fig.3, the tree is very accurate for the classification of training data, but the classification of unknown test data is not so accurate, that is, there is a phenomenon of over-fitting. The reason of over-fitting is to improve the correct classification of training data in order to construct an overly complex decision tree.

The solution to this problem is to simplify the generated decision tree by considering the complexity of the decision tree. There are two ways: Pre-pruning and Post-pruning. It will build a better decision tree in both compactness and performance.

(4) For one data element, the classical decsion tree gives a hard bounday to decide which branch to follow, can you propose a “soft approach” to increase the robustness of the decision tree?

To induce a decision tree classifier for data having continuous valued attributes, the most common approach is, split the continuous attribute range into a hard (crisp) partition having two or more blocks, using one or several crisp (sharp) cut points. But, this can make the resulting decision tree, very sensitive to noise.

An existing solution to this problem is to split the continuous attribute into a fuzzy partition (soft partition) using soft or fuzzy cut points which is based on fuzzy set theory and to use fuzzy decisions at nodes of the tree. These are called soft decision trees in the literature which are shown to perform better than conventional decision trees, especially in the presence of noise. [1]

(5) Compare to the Naïve Bayes, what are the advantages and disvantages of the decision tree learning?

Compare to the Naïve Bayes, the advantage and disadvantage of decision tree learning are as follows.

Advantage:

1) It is easy to understand and explain.

2) It can process both the data type and the general type property at the same time.

3) It will deal with data with many attributes.

4) It will give feasible and effective results for large data sources in a relatively short period of time.

Disadvantage:

1) It ignores the correlation between attributes.

2) It will cause the emergence of the problem of over-fitting

3) It is difficult to deal with missing data.

4) The results of Information Gain tend to lean the attributes with more values in a data in which it has different sample sizes for each attribute.

3.5 References

[1] Kumar G K, Viswanath P, Rao A A. Ensemble of randomized soft decision trees for robust classification[J]. Sādhanā, 2016, 41(3):273-282.

3.6 Code

The code can be download at https://github.com/Jackboomer/Experiments-for-Pattern-Recognition-and-Machine-Learning.git .