An Improvement on the Algorithm of Decision Tree

Liu Xumin[1,2] Huang Houkuan[2] Xu Weixiang[3]

[1] School of Information Engineering, Capital Normal University, Beijing 100037
[2] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044
[3] School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044
E-MIAL: liuxmxxxy@263.net

Abstract

Decision tree learning is one of the most widely used and practical methods for inductive inference. In this paper, the idea of algorithm for building a decision tree is introduced by comparing the algorithm of information gain or entropy. The produced process of decision tree is given as an example. According to the theory of rough sets, the method of constructing decision tree is discussed. Using the algorithm, the complexity of decision tree is decreased, the construction of decision tree is optimized and the rule of data mining could be built.

Key words: data mining；rough sets; decision tree；information entropy

1. Introduction

Decision tree learning is one of the most widely used and practical methods for inductive inference. It is a method for approximating discrete-valued functions that is robust to noisy date. Decision tree has many advantages, such as its fast speed, high accuracy as well as the easy mode of production, which attracts many researchers in data mining [1]. There are many algorithms of decision tree. In 1986 J. Ross Quinlan discovered the famous decision tree induction algorithm, ID3 version [2], which caused much effect. Making supplement and improvement on it in 1993, he raised the popular C4.5 algorithm [3]. Later C5.0 algorithm appeared as an improved commercial version of C4.5. Besides some scalability algorithms like SLIQ [4], SPRINT [5] and RainForest algorithms [6] also have wide application.

The theory of Rough Set (RS) [7] put forward by Professor Z.Pawlak in 1982, which identifies the knowledge from a new angle and associates knowledge with classification, provides a mathematical tool which are more sharable for human’s recognition to deal with the inaccurate and incomplete data classifying problems. The theory is mainly applied to reduce knowledge and analyze knowledge dependence, and have been widely used in fields like artificial intelligence, information process, pattern recognition, machine learning and knowledge discover.

In this paper, a data-mining algorithm of building a decision trees is introduced by comparing the gain or entropy. According to the theory of rough sets, the method of simplifying the structure of decision tree is discussed. A method of constructing decision tree with rough sets theory is given, and the production process of decision tree is illustrated by example.

2. Decision Tree

A decision tree [8] is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision tree in a top-down recursive and divide-and-conquer manner. In the process of constructing a decision tree, one of the most important steps is to choose a proper attribute as a node so as to produce the simplest decision tree.

Basic idea of ID3 algorithm is: The decision tree starts with a single node presenting the training samples. If the samples are all of the same class, then the node become a leaf and is labeled with that class. Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separates the samples into individual classes. This attribute becomes the “test” or “decision” attribute at the node. A new node is created and labeled with that attribute. A branch is created for each known value of the test attribute, and the samples are partitioned accordingly. The algorithm uses the same process to recursively form a decision tree for the samples at each partition. The recursive partitioning stops when all samples for a given node belongs to the same class or there are no remaining attributes on which the samples may be further partitioned. Each route from root to leaves of the decision tree can obtain partition regulations. And the accuracy of the regulations can be estimated by using training samples.

3. Foundation of the RS Theory

RS theory defines the knowledge from a new angle. It regards knowledge as the partition of domain, supposes it has granularity and discusses knowledge by introducing the equivalence relation in algebra. [9]

The problem to be dealt with is expressed as a knowledge expression system *S*=<*U*, *C*, *D*>, where *U* is a limited set called domain, composed of interested subjects. *C* is a set of condition attributes, while *D* is a set of decision attributes. $Q=C\cup D$ is called attribute set. $A\subseteq Q$, *A* is a subset of *Q*. $X\subseteq U$, *X* is a subset of domain *U*.

The lower approximation of set *X* about *A* is defined as:

$$A_o(X)=\bigcup\{Y\in U/A: \quad Y\subseteq X\} \qquad (1)$$

Which means the biggest set composed of subjects, which, according to the owned knowledge, belong to X. It is also called the positive area of X.

The upper approximation of set *X* about *A* is defined as:

$$A^o(X)=\bigcup\{Y\in U/:A \quad Y\cap X\neq\phi\} \qquad (2)$$

Which means the smallest set composed of subjects, which probably belong to X.

The negative area of set *X* about A is defined as:

$$NEG_A(X)=U-A^o(X) \qquad (3)$$

Which means the set composed of subjects, which, according to the owned knowledge, don't belong to X.

The boundary area of set X about A is defined as:

$$R_sB=A^o(X)-A_o(X) \qquad (4)$$

R_SB is the approximate margin between the upper approximation and lower approximation of the set.

According to the definitions above, in rough set theory, knowledge is defined as a capacity of classification, a capacity to classify domain U that is based on an equivalence relation.

4. Construction of Decision Trees

Consider an information system example. Training data tuples (Tab. 1) is a data set. There are 12 tuples in the table, among which *A*, *B*, *C* and *D* are condition attributes, while d is a decision attribute.

Tab.1: Training data tuples from an information system

U	Condition attributes(*C*)				Decision attribute(*D*)
	A	*B*	*C*	*D*	Class(*d*)
1	1	2	2	1	1
2	1	2	3	2	1
3	1	2	2	3	1
4	2	2	2	1	1
5	2	3	2	2	2
6	1	3	2	1	1
7	1	2	3	1	2
8	2	3	1	2	1
9	1	2	2	2	1
10	1	1	3	2	1
11	2	1	2	2	2
12	1	1	2	3	1

Using ID3 algorithm, the decision tree is formed eventually, as shown in Fig. 1.

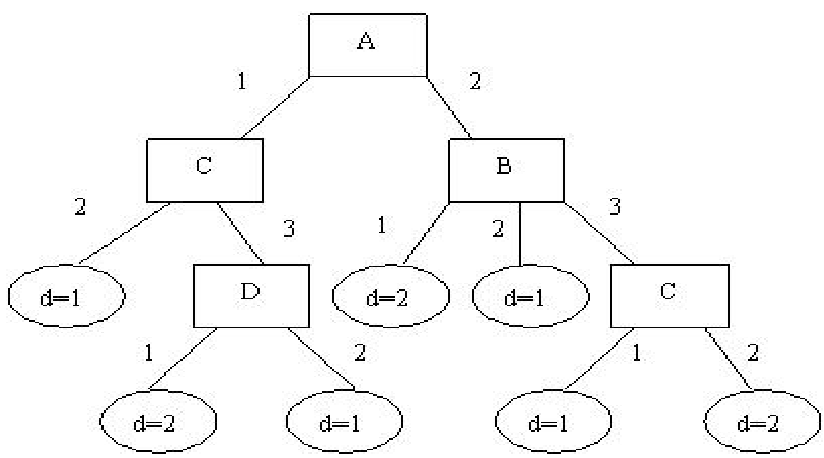

Fig. 1: A decision tree based on information entropy theory

As shown in Fig. 1, the complexity of the tree (the number of nodes of the tree) is 12. There are altogether 7 leaf nodes; so 7 classification rules can be gained.

From the process mentioned above of the construction of decision tree, we can see that:

When all the nodes of a leaf are d=1 or d=2, there's no need to partition, thus the downward recursive process stops. In the process of subtree production, the earlier such "pure" nodes appear, the better.

The general procedure of constructing decision trees is:

(1) Set the relative label of each attribute in condition attribute *C* to 0.

(2) Compute with formula (1) to (4) the lower approximation, upper approximation, and negative region and boundary region of each attribute in condition attribute *C*.

(3) Choose the condition attribute with the smallest boundary as the node of the present branch, as well as the current testing node, and set the relative label to 1.

(4) Construct decision tree at the present branch according to the possible values of the chosen attribute in the data set.

(5) For each branch of the chosen attribute of the tree, if the leaf node isn't reached, the recursive transfer is carried out.

(6) Otherwise, set the relevant label to 0, and the algorithm is over.

This algorithm is a recursive one. The input is a data set to be classified, while the output is a decision tree.

Next, it is shown that during the process of the construction of a decision tree, each tuple of the data set is partitioned to different classifications according to the four condition attributes.

First, partition according to the attribute *A*. Using formula (1) and (2), compute the upper and lower approximation about whether the result attribute *d* is 1, corresponding to condition attributes. The results are as follow:

$$A_o(d=1)=\phi$$
$$A^o(d=1)=\{1,2,3,4,5,6,7,8,9,10,11,12\}$$
$$B_o(d=1)=\phi$$
$$B^o(d=1)=\{1,2,3,4,5,6,7,8,9,10,11,12\}$$
$$C_o(d=1)=\{8\}$$
$$C^o(d=1)=\{1,2,3,4,5,6,7,8,9,10,11,12\}$$
$$D_o(d=1)=\{3,12\}$$
$$D^o(d=1)=\{1,2,3,4,5,6,7,8,9,10,11,12\}$$

Using formula (4), compute the boundary attribute of each attribute. Because the boundary of condition attribute D i.e. $D^o(d=1)-D_o(d=1)$ has the least tuples, *D* is chosen as the root node of the decision tree. Part of the constructed decision tree is shown in Fig. 2.

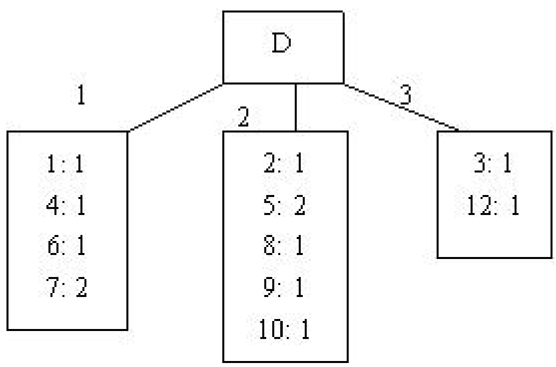

Fig. 2: Part of the decision tree

The other three attributes *A*, *B* and C are tested respectively on branch *D*=1. The lower and upper approximation computed by formula (1) and (2) are as follow:

$$A_o(d=1)=\{4\}$$
$$A^o(d=1)=\{1,4,6,7\}$$
$$B_o(d=1)=\{6\}$$
$$B^o(d=1)=\{1,4,6,7\}$$
$$C_o(d=1)=\{1,4,6\}$$
$$C^o(d=1)=\{1,4,6\}$$

Compute the negative region corresponding to attribute *C* with formula (3).

$$NEG_C(d=1)=\{7\}$$

The boundary of attribute *C* has less tuples than the other two attributes, therefore attribute *C* is chosen as the node of branch *D*=1. Branch *D*=2 is handle in the same process; attribute *A* is chosen as the node, then attribute *C* is to be chosen as the node of branch *A*=2. Because branch *D*=3 doesn't need to be further partitioned, the constructed decision tree is shown in Fig. 3.

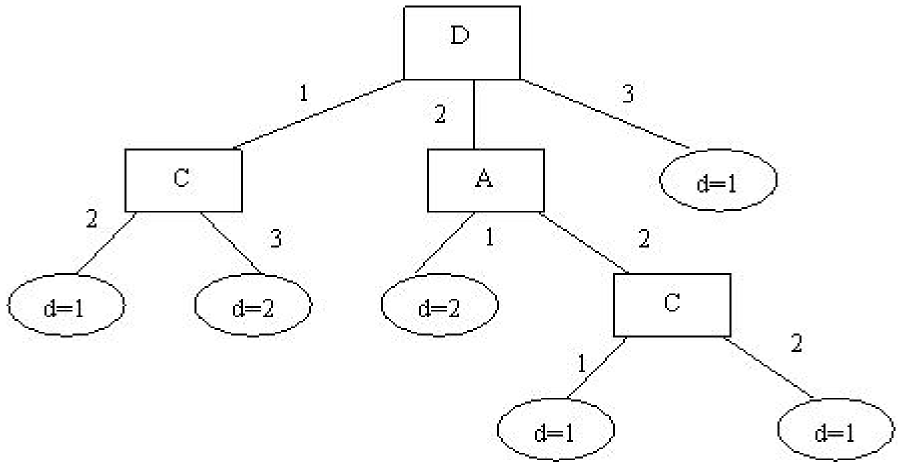

Fig. 3: Decision tree based on rough set theory

From the Fig. 3, we can see that the number of nodes of the tree is 10, which are 2 nodes less than the tree constructed by ID3 algorithm. Correspond to 6 leaf node can gained 6 classify rule. So decision trees constructed based on rough set theory is simpler than those constructed by information entropy.

5. Conclusion

With the increase of data set, the decision tree led out from it will also increase rapidly, which makes it more difficult for people to understand the decision. The main difficulty of decision tree algorithm in data mining lies in how to choose a good branch value. The growth of the tree speed up if the rate of information gains is used as the standard to choose attribute. Using the method mentioned in this paper to construct decision trees, the complexity of trees are decreased, the structure of the tree is optimized and better information regulation can be mined.

Acknowledgements

This paper is supported by Beijing Educational Committee under Grant No. KM200410028013.

References

[1] Shaowen Zhu, "Technique and Development of Decision Tree Mining", Computer Engineering, Vol. 26, No. 10, pp.1-3, 2000. (in Chinese)

[2] J.R. Quinlan, "Induction of decision trees", Machine Learning, No. 1, pp. 81-106, 1986.

[3] J.R. Quinlan, "C4.5: Programs for Machine Learning", San Mateo, CA: Morgan Kaufmann, 1993.

[4] M. Mehta, R. Agrawal, and J. Rissanen, " SLIQ: A fast scalable classifier for data mining", In Proc. 1996 Int. Conf. Extending Database Technology (EDBT'96), Avignon, France, Mar.1996.

[5] J. Shafer, R. Agrawal, and M.Mehta, " SPRINT: A scalable parallel classifier for data mining", In Proc. 1996 Int. Conf. Very Large Data Bases (VLDB'96), pp. 544-555, Bombay, India, Sept. 1996.

[6] J. Gehrke, R. Ramakrishnan, and V. Ganti, " Rainforest: A framework for fast decision tree construction of large datasets", In Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98), pp.416-427, New York, Aug. 1998.

[7] Pawlak Z, "Rough sets: theoretical aspects of reasoning about data", Netherlands: Kluwer Academic Publishers, 1991.

[8] Jiawei Han, Micheline Kamber, " Data Mining Concepts and Techniques", Beijing. China Machine Press, Aug. 2000. (in Chinese)

[9] Tongming Liu, " Data Mining Techniques and Its Application", Beijing. Publishing House of National Defence Industry, Sept. 2001. (in Chinese)