An Improved ID3 Decision Tree Algorithm

Chen Jin, Luo De-lin*

School of Information Science and Technology
Xiamen University
Xiamen, 361005, China
chenjin524ok@163.com, luodelin602@163.com

Mu Fen-xiang

Tsingtao Huanghai Vocational College
Tsingtao, 266427, China

***Abstract*—Decision tree is an important method for both induction research and data mining, which is mainly used for model classification and prediction. ID3 algorithm is the most widely used algorithm in the decision tree so far. Through illustrating on the basic ideas of decision tree in data mining, in this paper, the shortcoming of ID3's inclining to choose attributes with many values is discussed, and then a new decision tree algorithm combining ID3 and Association Function(AF) is presented. The experiment results show that the proposed algorithm can overcome ID3's shortcoming effectively and get more reasonable and effective rules**

***Index Terms*—data mining, decision tree, ID3, association function(AF), variety bias**

I. INTRODUCTION

With the development of computer technology and computer network technology, the degree of informationization is getting higher and higher, people's ability of using information technology to collect and produce data is substantially enhanced. How can we not be drowned by the sea of information, and from which discovering useful knowledge and improving the effectiveness of information utilization are problems need to be addressed urgently. It was under this background that Data Mining (DM) technology came into being and developed. Data mining is a process to extract information and knowledge from a large number of incomplete, noisy, fuzzy and random data. In these data, the information and knowledge are implicit, which people do not know in advance, but potentially useful. At present, the decision tree has become an important data mining method. The basic learning approach of decision tree is greedy algorithm, which use the recursive top-down approach of decision tree structure. Quin lan in 1979 put forward a well-known ID3[1,4,5] algorithm, which is the most widely used algorithm in decision tree. But that algorithm has a defect of tending to use attributes with many values. Aiming at the shortcomings of the ID3 algorithm, in the paper, an association function is introduced to improve ID3 algorithm. The result of experiment shows that the presented algorithm is effective.

II. ID3 ALGORITHM

In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. In this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. That is to say, the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. Therefore, the use of such an information theory approach will effectively reduce the required dividing number of object classification [6].

Set S is set including s number of data samples whose type attribute can take m potential different values corresponding to m different types of C_i *i (1,2,3, ..., m)*. Assume that s_i is the sample number of C_i . Then, the required amount of information to classify a given data is

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log(p_i) \qquad (1)$$

where $P_i = S_{ij} / |S_j|$ is the probability that any subset of data samples belonging to categories C_i.

Suppose that A is a property which has v different values $\{a_1, a_2, \cdots, a_v\}$. Using the property of A, S can be divided into v number of subsets $\{S_1, S_2, \cdots, S_v\}$, in which S_j contains data samples whose attribute A are equal a_j in S set. If property A is selected as the property for test, that is, used to make partitions for current sample set, suppose that S_{ij} is a sample set of type C_i in subset S_i , the required information entropy is

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s} I(S_{1j}, \cdots, S_{mj}) \qquad (2)$$

Such use of property A on the current branch node corresponding set partitioning samples obtained information gain is:

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \qquad (3)$$

ID3 algorithm traverses possible decision-making space using top-down greedy search strategy, and never trace back and reconsider previous selections. Information gain is exactly

*Corresponding author.

Authorized licensed use limited to: INSTITUTO TECNOLOGICO DE AERONAUTICA. Downloaded on March 21,2010 at 20:18:53 EDT from IEEE Xplore. Restrictions apply.

the metrics for selecting the best attribute in each step of the growth tree in ID3 algorithm.

2.1 algorithm for generating a decision tree[2] according to a given data sets

Input: training samples, each attribute taking discrete value, a candidate attribute set available for induction is attribute_list.

Output: a decision tree.

Deal flow:

1) Create a node *N*;
2) If all samples of the node are of the same category *C*, then return *N* as a leaf node and mark with category *C*, the beginning root node corresponds to all the training samples.
3) If attribute_list is empty, then return N as a leaf node and mark the node as a type whose samples contain the largest number of categories;
4) select a test_attribute with the largest information gain from arrtibute_list, and mark node *N* with test_attribute;
5) For each given value a_i of test_attribute, the sample set contained in node N is portioned.
6) According to the condition of test_attribute = a_i , a corresponding branch is generated from the node N to indicate the test conditions;
7) Set s_i is the obtained sample set under the condition of test_attribute = a_i . If s_i is empty, then mark the corresponding leaf node with category of including the most number of sample types. Otherwise, It will be marked with a the return value:

$$Generate\_decision\_tree(s_i, attribute\_list - test\_attribute);$$

This is a greedy algorithm which use recursive manner of top-down, divide and conquer to construct a decision tree. The termination condition of recursion is : all samples within a node are of the same category. If no attribute can be used to divide current sample set, then voting principle is used to make it a Compulsory leaf node, and mark it with the category of having the most number of sample types. If no sample satisfies the condition of test-attribute = a_i , then a leaf node is created, and mark it with the category of having the most number of sample types.

2.2 The shortcoming of ID3 algorithm

The principle of selecting attribute A as test attribute for ID3 is to make E (A) of attribute A, the smallest. Study suggest that there exists a problem with this method, this means that it often biased to select attributes with more taken values[3,6], however, which are not necessarily the best attributes. In other words, it is not so important in real situation for those attributes selected by ID3 algorithm to be judged firstly according to make value of entropy minimal. Besides, ID3 algorithm selects attributes in terms of information entropy which is computed based on probabilities, while probability method is only suitable for solving stochastic problems. Aiming at these shortcomings for ID3 algorithm, some improvements on ID3 algorithm are made and a improved decision tree algorithm is presented.

III. THE IMPROVED OF ID3 ALGORITHM

To overcome the shortcoming stated above, attribute related method is firstly applied to computer the importance of each attribute. Then, information gain is combined with attribute importance, and it is used as a new standard of attribute selection to construct decision tree. The conventional methods for computing attribute importance are sensitivity analysis (SA)[7], information entropy based joint information entropy method(MI)[8], Separation Method(SCM)[9], Correlation Function Method(AF)[10,11], etc. SA needs not only to compute derivatives of output respect to input or weights of neural network, but also to train the neural network. This will increase computational complexity. MI needs to compute density function and it is not suitable for continuous numerical values. SCM computes separation property of input-output and the correlation property of input and output attributes and is suitable for both continuous and discrete numerical values, but computation is complex. AF not only can well overcome the ID3's deficiency of tending to take value with more attributes, but also can represent the relations between all elements and their attributes. Therefore, the obtained relation degree value of attribute can reflect its importance.

AF algorithm: Suppose *A* is an attribute of data set *D*, and *C* is the category attribute of *D*. the relation degree function between *A* and *C* can be expressed as follows:

$$AF(A) = \frac{\sum_{i=1}^{n} |x_{i1} - x_{i2}|}{n} \tag{4}$$

Where x_{ij} ($j = 1, 2$ represents two kinds of cases) indicates that attribute *A* of *D* takes the *i*-th value and category attribute *C* takes the sample number of the *j*-th value, *n* is the number of values attribute *A* takes.

Then, the normalization of relation degree function value is followed. Suppose that there are *m* attributes and each attribute relation degree function value are

$AF(1), AF(2), \cdots, AF(m)$, respectively. Thus, there is

$$V(k) = \frac{AF(k)}{AF(1) + AF(2) + \cdots + AF(m)} \tag{5}$$

Which $0 < k \le m$. Then, equation (3) can be modified as

$$Gain'(A) = (I(s_1, s_2, \cdots, s_m) - E(A)) \times V(A) \tag{6}$$

$Gain'(A)$ can be used as a new criterion for attribute selection to construct decision tree according to the procedures of ID3 algorithm. Namely, decision tree can be constructed by selecting the attribute with the largest $Gain'(A)$ value as test

978-1-4244-3521-0/09/$25.00 ©2009 IEEE

Authorized licensed use limited to: INSTITUTO TECNOLOGICO DE AERONAUTICA. Downloaded on March 21,2010 at 20:18:53 EDT from IEEE Xplore. Restrictions apply.

attribute. By this way, the shortcomings of using ID3 can be overcame.

it construct the decision tree, this tree structure will be able to effectively overcome the inherent drawbacks of ID3 algorithm.

IV. EXPERIMENTAL RESULTS

A customer database of some shopping mall is shown in Table 1 (a training sample set). The category attribute of the sample set is "buying-computer", which can take two different values: buying-computer or No buying-computer

Table 1. Shopping mall customer database

Case	Age	Color-cloth	Income	Student	Buy-computer
1	>40	Red	High	No	No
2	<30	Yellow	High	No	No
3	30--40	Blue	High	No	Yes
4	>40	Red	Medium	No	Yes
5	<30	White	Low	Yes	No
6	>40	Red	Low	Yes	No
7	30—40	Blue	Low	Yes	Yes
8	<30	Yellow	Medium	No	Yes
9	<30	Yellow	Low	Yes	No
10	>40	White	Medium	No	No

In order to illustrate the effectiveness of our present algorithm, the improved ID3 algorithm and ID3 algorithm are applied on this example to construct decision trees and comparison is made. Figure 1 and figure 2 show the generated decision trees using the ID3 algorithm and the improved ID3 algorithm, respectively.

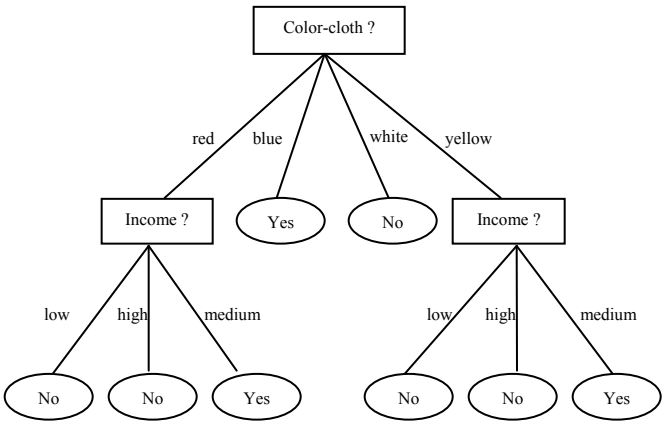

Figure 1. The obtained decision tree using ID3 algorithm

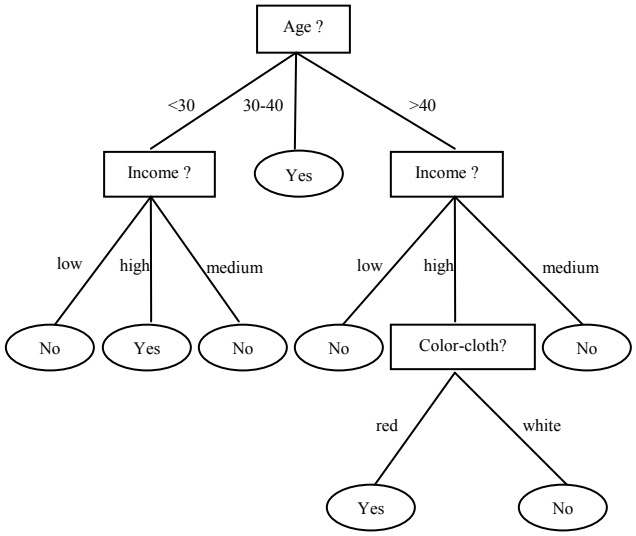

Figure 2. The obtained decision tree using improved ID3 algorithm

The two results of the experiment shows that ID3 algorithm choose attribute color-cloth as root node to generate decision tree, but the importance of attribute color-cloth is lower than the other attributes, and it is just the shortcoming of ID3 which tends to take attributes with many values. However the improved ID3 algorithm decreases the importance of attribute color-cloth in classification and comparatively enhanced the importance of attributes such as age, income, and student, etc. in classification. It well solves the problem that ID3 algorithm tends to take attributes with many values and it can obtain more reasonable and effective rules.

V. CONCLUSION

In this paper, an improved ID3 algorithm is presented to overcome deficiency of general ID3 algorithm which tends to take attributes with many values. The presented algorithm makes the constructed decision tree more clear and understandable. Because it needs to compute the relation degree function value for each attribute based on ID3 algorithm, it unavoidably increases computational complexity. But with the rapid development of computer technology, the operating speed of computer gets faster and faster, the increased computational complexity can be neglected. Generally speaking, the improved ID3 algorithm takes the advantages of ID3 and AF algorithms and overcomes their disadvantages. Experiment results show that the improved ID3 can generate more optimal decision tree than general ID3 algorithm.

ACKNOWLEGMENT

The authors would like to thank the anonymous reviewers of this paper for their helpful suggestions to improve this paper. The work of this paper is supported by the National Natural Science Foundation of China under the grant number 60674100 and Aeronautical Science Foundation.

REFERENCES

Authorized licensed use limited to: INSTITUTO TECNOLOGICO DE AERONAUTICA. Downloaded on March 21,2010 at 20:18:53 EDT from IEEE Xplore. Restrictions apply.

[1] I. H. Witten, E. Frank, Data Mining Practical Machine Learning Tools and Techniques, China Machine Press, 2006.

[2] Y. T. Zhang, L. Gong, Principles of Data Mining and Technology, Publishing House of Electronics Industry.

[3] D. Jiang, Information Theory and Coding [M]: Science and Technology of China University Press, 2001.

[4] S. F. Chen, Z. Q. Chen, Artificial intelligence in knowledge engineering [M]. Nanjing: Nanjing University Press, 1997.

[5] Z. Z. Shi, Senior Artificial Intelligence [M]. Beijing: Science Press, 1998.

[6] M. Zhu, Data Mining [M]. Hefei: China University of Science and Technology Press ,2002.67-72.

[7] A. P. Engelbrecht., A new pruning heuristic based on variance analysis of sensitivity information[J]. IEEE Trans on Neural Networks, 2001, 12(6): 1386-1399.

[8] N. Kwad, C. H. Choi, Input feature selection for classification problem [J],IEEE Trans on Neural Networks, 2002,13(1): 143- 159.

[9] X. J. Li, P. Wang, Rule extraction based on data dimensionality reduction using RBF neural networks [A]. ICON IP2001 Proceedings, 8th International Conference on Neural Information Processing [C]. Shanghai, China, 2001.149- 153.

[10] S. L. Han, H. Zhang, H. P. Zhou, correlation function based on decision tree classification algorithm for computer application in November 200.

[11] S. Y. Zhang, Z. Y. Zhu, Study on decision tree algorithm based on autocorrelation function. Systems Engineering and Electronic Jul. 2005 Vol.27 No.7.

 978-1-4244-3521-0/09/$25.00 ©2009 IEEE

Authorized licensed use limited to: INSTITUTO TECNOLOGICO DE AERONAUTICA. Downloaded on March 21,2010 at 20:18:53 EDT from IEEE Xplore. Restrictions apply.