The Application of Improved Decision Tree Algorithm in Data Mining of Employment Rate: Evidence from China

Yuxiang Shao[1] , Qing Chen[2] , Weiming Yin[3]
[1] School of Computer Science and Technology,
[3] Faculty of Mechanical & Electronic information
China University of Geosciences, Wuhan, Hubei,430074, China
[2] School of Computer Science and Technology,
Wuhan Institute of Technology, Wuhan, Hubei, 430073, China
syxcq@163.com

Abstract

In allusion to the limitations of ID3 algorithm which is the core algorithm of building decision tree, this paper presented the improved project which added attribute measure and introduced information gain ratio. The project trained the example set in the data mining of the students' employment rate and created the employment data mining model. The model was tested with the data of the students in school. The test proved that the improved decision tree algorithm has high reference value to the reallocation of the teaching resources.

1. Introduction

At present, the students' employment rate is one of the important factors to assess the higher education institutions and the valve of the regulation to the efficient teaching resources. The employment is the most important issue cared by society and family. In allusion to the limitations of ID3 algorithm which is the core algorithm of building decision tree, this paper presented the improved solution. The improved decision tree algorithm was applied in the field of employment rate forecast and the deep-seated data mining of the teaching information. The improved algorithm is a good guidance for improving teaching management, enhancing teaching quality and distributing teaching resources reasonably[1~2].

2. Decision Tree Algorithm

2.1 Basic Idea

The so-called decision tree is a tree structure similar to flow chart which is composed of decision-making nodes, leaves and branches. Each internal node presents the test on an attribute, each branch presents a testing output, and each leaf presents type or its distribution. The top node of tree is the root. The process of classifying a specific data project is to determine along a branch from the root to the leaf, thus, achieve a decision making. A path from root to leaf is a classification rule. Decision tree can be transformed into classification rules easily which is a very intuitionistic classification model expression. At present, the mostly used and mature decision tree algorithm is ID3 brought forward by Quinlan[3~5].

2.2 ID3 Algorithm

In ID3 algorithm, the test attribute is the largest gain information value. The division of the sample set according to the value of test attribute, the sample set divides into how many sub-sample set in term of how many different value of the test attribute. At the same time, the node corresponding to this sample set on the decision tree outgrows a new node. This method allows the least amount of anticipant test number to classify an object and build a simple but not the simplest tree. The basic principle of the decision tree algorithm based on ID3 is as follows[6~8].

(1) Consider T as a dataset. Category set is $\{C_1, C_2, C_3, \cdots, C_k\}$. Select an attribute V to divide T into multi-subset.

(2) Suppose that the attribute V has n non-overlapped values $\{v_1, v_2, v_3, \cdots, v_k\}$, then T is divided into n subset $T_1, T_2, \cdots, T_n$. The value of all instances in T_i is v_i .

978-0-7695-3604-0/09 $25.00 © 2009 IEEE
DOI 10.1109/DBTA.2009.72

(3) Consider $\left|T\right|$ as the instance number of dataset T , $\left|T_i\right|$ as the instance number of $V = v_i$, $\left|C_j\right| = freq(C_j, T)$ as the instance number of C_j category and $\left|C_{jv}\right|$ as the instance number which has C_j category in the attribute $V = v_i$.

Then, the occurrence probability of category C_j is $\mathrm{P}(\mathrm{C}_j) = \frac{C_j}{|T|} = \frac{freq(C_j, T)}{|T|}$, the occurrence probability of attribute $V = v_i$ is $\mathrm{P}(\mathrm{v}_j) = \frac{|T_i|}{|T|}$, and the condition probability which has C_j category in the attribute $V = v_i$ is $P(C_j \mid v_i) = \frac{|C_j v_i|}{|T_i|}$.

The information entropy of category:

$$H(C) = -\sum_j P(C_j) \log_2 P(C_j)$$
$$= -\sum_j \frac{freq(C_j, T)}{|T|} \times \log_2 \frac{freq(C_j, T)}{|T|} \qquad (1)$$

Hear $\inf o(T) = H(C)$.

The condition entropy of category:

Divide T according to the attribute V, then the condition entropy is:

$$H(C \mid V) = -\sum P(V_i) \sum P(C_j \mid v_i) \log_2 P(C_j \mid v_i)$$
$$= \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \inf o(T_i) \qquad (2)$$

Hear $\inf ov(T) = H(C \mid V)$

The information gain namely mutually information is:

$$I(C,V) = H(C) - H(C \mid V) = \inf o(T) - \inf ov(T)$$
$$= gain(V) \qquad (3)$$

The ID3 algorithm takes $I(C,V)$ as the criterion of test attribute, and divides training example until building a decision tree. Quinlan suggested that choose the attribute with the largest amount of information as extended attribute, because the maximum information is equivalent to minimum uncertainty, namely minimum entropy.

2.3 The Improved Strategy of ID3 Algorithm

The ID3 algorithm with clear theory, simple method and strong learning ability, however, has some shortcomings. Firstly, ID3 takes information gain as classification evaluation function to select optimal attribute. However, this selection criterion tends to choose the attribute with more values which is not always the most important attribute $gain(V)$. Secondly, ID3 can only deal with the discrete value and helplessly with the continuous value, and does not consider the lack value problem in the training set.

For the above disadvantages and shortcomings of ID3, the improvement is as following:

(1) Add the attribute measure

ID3 taking the attribute V as test attribute is according to the principle of minimum entropy and maximum information that is the largest gain information principle. This algorithm tends to choose the attribute with more values which is not always the optimal attribute. Moreover, in the discussion of different issues, each attribute has different importance degree which can be its measurement S. The revised formula of classification condition entropy introduced attribute measure is as following[9]:

$$H(C \mid V) = -(\sum P(V_i) + S) \sum P(C_j \mid v_i) \log_2 P(C_j \mid v_i) \qquad (4)$$

The information gain formula is：

$$gain(V) = I(C,V) = H(C) - H(C \mid V)$$
$$= \inf o(T) - \inf ov(T) \qquad (5)$$

Hear, the attribute measure value is in [0， 1], of which the size is computed by the data in the training dataset according to the method of probability statistics. Assume that A and B are events, $P(B / A)$ is the condition probability of the event B occurrence when event A is occurring. Moreover, $P(B / A)$ is the effect degree of the certain category in the training set from event B after the occurrence of event A, namely the measurement of event B. The formula is

$$P(B / A) = P(AB) / P(A) \qquad (6)$$

For each attribute in the example set, firstly, compute all the attribute measurement, then, compare the values and it is known that the attribute with larger measure value provides more information to the classification. Followed by analogy, the effect degree of each attribute namely measurement is confirmed. According to the measure value, the information gain will be recomputed.

(2) Introduce the rate of information gain

Information gain rate equals to the ratio of information gain and the information divided.

$$GainRatio(A) = Gain(A) / SplitInfo(A) \qquad (7)$$

For the sample set T, suppose that A is the discrete attribute with s different values. The information gain algorithm using A to divide the sample set is like ID3. The formula of division information is as follows[10]:

$$SplitInfo(A) = -\sum_{i=1}^{s} \left|\frac{T_i}{T}\right| \times \log_2 (p \left|\frac{T_i}{T}\right|) \qquad (8)$$

The choice of the attribute with the largest $GainRatio(A)$ as the branch property can be a very good solution to the ID3 multi-value problem.

3. The Application of Employment DM

3.1 Data Mining Flow

```
Step1: Create the root N
Step2: If T belongs to the same category C, then return N as the leaf node marked category C.
Step3: If T_Attributelist is NULL, then return N as the leaf node and mark N as the mostly appeared category.
Step4: For each attribute in T_Attributelist
    Compute the attribute Measure
    Compute the information Gain
    Compute the $SplitInfo(A)$
    Compute the $GainRatio(A)$
Step5: The Test_Attribute of N equals to the attribute with tiptop $GainRatio(A)$ in T_Attributelist.
Step6: For each value in T_Attributelist
{
Create a new leaf node from node N;
If the sample set $T'$ corresponding to the new leaf node is NULL.
    Then this leaf node can't be split and mark it as the mostly appeared category in T
Else Keep execute
$FunctionID3(T', T'\_Attributelist)$
on this leaf node and split it
}
```

3.2 Data Source

The employment forecast is to analyze the function relationship between the basic information attribute of the students in school and the employment rate, needing the students' information and the employment situation data, which can be acquired from the student management system. At present, the database accumulated a mass of graduate information. 2000 items were draw from all the graduate information to ensure the credibility of data mining. Table 1 is part items of the selected data.

Table 1 The part items of the selected data

Number	Gender	Age	Department	Scholarship	Associator	Employed
1	F	27	Automation	False	True	False
2	F	27	Computer	False	False	False
3	M	24	Automation	True	True	True
4	M	24	Computer	False	True	True
5	F	23	Automation	False	True	True
6	M	25	Computer	True	True	True
7	M	24	Computer	False	True	True
8	M	25	Computer	True	False	True
9	F	25	Automation	False	False	False
10	F	23	Computer	True	False	True

3.3 The Model of DM

Input Attributes：Gender、Agent、Department、Scholarship、Associator. PredictAttribute：Employed.

The decision tree model is shown as Figure 1.

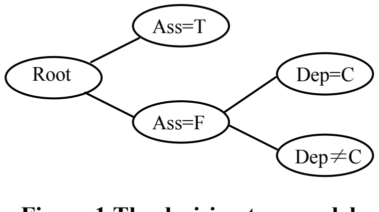

Figure 1 The decision tree model

3.4 The Data Mining Results Analysis

In 100 items of training samples, 20 items are unemployed and the unemployed rate is 21.84 percent. 80 items are employed and the employment rate is 73.56 percent. In 58 items with social experience of 100 training samples, 3 items are unemployed and the unemployed rate is 7.45 percent. 55 items are employed and the employment rate is 89.65 percent. In 42 items without social experience of 100 training samples, 17 items are unemployed and the unemployed rate is 39.67 percent. 25 items are employed and the employment rate is 56.57 percent. In 16 items majored in computer of 42 items without social experience, 1 item is unemployed and the unemployed rate is 8.33 percent. 15 items are employed and the employment rate is 89.10 percent. In 26 items majored in non-computer, 16 items are unemployed and the unemployed rate is 57.78 percent. 10 items are employed and the employment rate is 37.78 percent. The social experience and the specialty are closely related to the employment situation. The most relevant attribute is the social experience and other attributes are weakly related.

Use the model to forecast the employment rate of the undergraduate. The result is shown in table 2.

Table 2 The undergraduate data and forecast result

Number	Gender	Age	Department	Scholarship	Associator	Employed
1	F	21	Automation	True	True	89.64%
2	F	23	Automation	False	False	57.77%
3	M	24	Computer	False	True	89.65%
4	F	26	Communication	True	False	57.77%
5	M	26	Computer	False	False	89.10%

4. Conclusion

Similarly, through experiment and analysis using the above student data, we found that in different scale dataset there is great similarity between the decision trees constructed by ID3 and improved ID3 algorithm in a few aspects such as “amount of root node”, “amount of leaf node”, “levels of tree”, “amount of rules” etc. The time of constructing decision tree according to improved ID3 is less than ID3. The similarity proves that there is generally the same classification accuracy using the decision tree constructed by ID3 and improved ID3. The time difference proves that there is greater superiority using improved ID3 than ID3 to construct decision tree from large scale dataset. Thus, the decision tree model based on improved ID3 can not only succeed the advantages of the traditional ID3, but also overcome the disadvantages of selecting multi-value attribute. Building the optimal forecast model and implementing data mining on the student employment rate information in the teaching field is one of the effective means to adjust teaching resources reasonably, increase the employment chance and improve the teaching quality.

5. Acknowledgement

This work is supported by the National Science Foundation of Hubei Province (No.2005ABA219), The Education Research Project of China University of Geosciences (No.2008023).

6. References

[1] X.B.Yang, J.ZHang. Decision Tree and Its Key Techniques. Computer Technology And Development, 2007,17:1 43~45

[2] Q.LI A Comparative Study on Algorithms of Constructing Decision Trees. Journal of Gansu Sciences, 2006,18:4 84~87

[3] A.D.Yin, L.Q.Xie Researches on Dynamic Algorithm of Decision trees. Computer Engineering and Applications 2004 33 103~105

[4] G.Gong, H.Zhang. Application of AIgorithm of Decision Tree in Weather Assessment. Microcomputer Information 2007 12 245~246

[5] X.M.Dong, CH.Lin Decision tree algorithm based on compacted rule-space. Application Research of Computes 2007 24:11 222~224

[6] Y.ZHang , T.D.Liu Decision Tree Algorithm Based 0n the Information Theory. Control Theory and Applications. 2006 25:1 4~7

[7] R.ZHao, H.Li Algorithm of Multi-valued Attribute and Multi-labeled Data Decision Tree. Computer Engineering 2007 33:13 87~89

[8] J.Xu The Study of the Discretization Method in Decision Tree. Journal of Hebei Institute of Technology 2007 29:2 71~74

[9] X.SH.Li Q.H.Zhang Improved Decision Tree Algorithms and Its Application in Enterprise Resource Planning. Transactions of Beijing Institute of Technology 2006 26:2 139~142

[10] Y.G.Zhu, S.C H An Algorithm s to Construct Decision Tree Based on Information Entropy. Journal of ChangZhou Institute of Technology 2006 19:1 55~58