ATLANTIS PRESS

Advances in Computer Science Research, (ACSR), volume 52

2016 International Conference on Computer Engineering and Information Systems (CEIS-16)

A New Algorithm of Grading and Classification for Massive Data Processing Based on Decision Tree

Xin Jing, Hong-Da Li

College of Science, Shenyang Jianzhu University, Shenyang, China

E-mail: jingxin@sjzu.edu.cn

***Abstract*-In the paper, the rules of grading and classification for decision tree have been discussed. By using the information entry gain ratio in dealing with classification, the accuracy of the algorithm has been improved. Based on the C4.5 method, an algorithm for massive data processing has been established by the decision tree with the rules. Based on the data CET4, the result shows that the decision tree method is effectively. From the result, the English learning strategies and suggestions have been put forward to the students with different foundation.**

Keywords-massive data; data processing; information entry; grading and classification; decision tree

I. INTRODUCTION

There are many new characteristics of massive data, such as distributed, heterogeneous. Many high requirements are put forward to storage resources, computing resources and network resources. The huge challenge has been brought to the traditional data management [1].

Data is a kind of expression of facts, concepts, or instructions form. Data will become information after explain and give to a certain meaning. The basic purpose of data processing is to extract and deduced valuable and meaningful data for people from a large number of messy and difficult data to understand. Through model analysis, useful information can be found out and then through the validation of online analysis (OLAP),useful market information, or the potential market information combined with the customer registration will also be found out [2].

The algorithm for massive data processing based on decision tree is studied in the paper. The classification of data mining and classification rules have been established.

Through testing the data of CET4, the experiment results show that this algorithm can improve the classification accuracy and provide effective strategies and suggestions to the students English learning.

II. THE DESIGN IDEA OF THE DECISION TREE ALGORITHM

A. The Decision Tree Algorithm

The decision tree algorithm was first proposed by Australia's J.R. Quinlan professor [4]. In the algorithm, the concept of entropy in information theory was first used in calculating the information gain of before and after the split. But the algorithm still exist some disadvantages, such as it can't deal with continuous attributes, it needs to calculate the information gain to choose more attribute values, etc. In order to solve these problems, scholars have put forward a series of improved algorithms [3] [6].

In 2002, Salvatore Ruggieri improved C4.5 algorithm. EC4.5 is proposed. The algorithm adopts the binary search to replace the linear search. EC4.5 also puts forward three different strategies looking for continuous attribute value of improvement of local worshiping [5].

But it is also general and arbitrary in dealing with data properties. In order to overcome the disadvantages of the low accuracy of classification, a new algorithm is put forward.

B. Decision Tree Classification Model

The establishment of a decision tree classification model to predict is actually a kind of induction-deduction process. Classification model by the training set first must be test and meet certain requirements, then it can be used to predict.

Decision tree generation algorithm includes two steps: one is the tree generating. At the beginning all of data is in the root node, then to recursive data fragmentation; Second, tree pruning. That is to remove some of the unusual data may be noise. The stopping segmentation conditions of decision tree is as follows: one node data is belong to the same category, no attribute can be used in the data again divided.

Let T is testing set, $T=(\langle x,c_j\rangle), x=(a_1,a_2,\cdots,a_n)$ is as a training example. There are *n* properties in *x*, they are listed in $(A_1,A_2,\cdots,A_n)$.

$C=\{C_1,C_2,\cdots,C_n\}$ is the result of classification.

There are 5 steps in the decision tree generation algorithm:

- Step1: To Select property A_i from the table As the classification of attributes;
- Step2: If there are k_j properties of A_i, testing set T will be divided into k_j subsets, among them,
- $T_{ij}=\{\langle x,C\rangle|\langle x,c\rangle\}\in T$ and x is of the value for k_j properties of A_i;
- Step3: To delete the property A_i in the table;
- Step4: Let $T=T_{ij}$,for each j;
- Step5: If the property set is not empty, return to step1, otherwise output.

Decision tree algorithm can be used through the training set, among them, and as a training example, having some properties, attributes being listed in table, value of the property. In order to have the classification results, algorithm is as the following.

C. Decision Tree Pruning

Pruning method is used post pruning. It is to cut out the branches from the completely tree. The no pruning nodes become leaves on the bottom of the tree. And to use the most frequently one will be marked. For each no leaf node in the tree, it should be to calculate the expected error rate after the pruning subtree of that node. Then, to use error rate of each branch, error rate can be calculated from that not pruning combined with the weight of each branch evaluation,.

If cutting off the node can lead to the expectation of higher error rate, then keep the subtree;

Otherwise, the subtree should be cut out.

After the trees have been pruned gradually, a separate test set can be using to evaluate the accuracy in each tree. The decision tree with least expected error rate can be gotten.

The complexity and classification accuracy of the decision tree are the two most important factors of considering. The new algorithm is based on the quantitative evaluation standard as follows:

1) *the forecast accuracy:* In massive data, it should be in accordance with the requirements of the user. The useful information can be found by classifying data.

2) *simplicity description of:* The model is described simplicity will be easy to understand, as well as the more popular.

3) *computation complexity:* In data mining, the space and time complexity of the problem will be a very important factor, it will directly affect the cost of generating and calculation.

4) *the model robustness:* Robustness is a supplement of the prediction accuracy. It is the ability to accurately classify requested data in the presence of noise and defect data.

5) *the processing scale:* Processing scale is the ability in modeling in the case of vast amounts of data and the accuracy of tree constructing classification model.

III. DECISION TREE ALGORITHM DESCRIPTION:

Function Tree = Decision_Tree_Create (T, A, Y)

Input: training set T, the condition attribute sets A, target attribute Y.

Input: the decision Tree.

The Tree=Create_Node (T); // generated node of the Tree

If the T has the same target attribute values of all samples, use the attribute values to identification of the current node category, back to the Tree.

If there are not separable properties, to use target attribute value with T appear in the highest frequency of the current node type, back to the Tree.

(x,values)=Attribute-section (T, A, Y); //select the best attributes X and splitting point Values

For each V in Values do // according to the test (x, Values) divided sample set, generate the nodes

SubT = subset of V meeting the sample x test conditions;

Node=Decision_Tree_Create (subT, A – {x}, Y); //of child nodes recursive operations, usually x in subtrees generated no longer as splitting attributes

Create_Branch (Tree, the Node); // A branch of generated of T

End

Back to the Tree

Let T represent the current sample set, the current candidate attributes with T.

The algorithm is as follows:

Generate_decision_tree can be produced by a decision tree based on the given training data

Input: the training sample: samples; the set of candidate attributes ,attributelist

Output: a decision tree

- *Step*1: Create a root node N;
- *Step*2: IF T belong to the same class C, return N as leaf nodes, marked as class C;
- *Step*3: IF an attributelist is empty OR T sample left in less than a given value, Return N as leaf nodes, marked N is appeared of T in most classes:
- *Step*4: FOR each attribute in an attributelist information gain rate is to be Computed ;
- *Step*5: Test N's testing attributetest. Attribute=attributelist, the attribute with the highest information gain rate;
- *step*6: IF the test attribute is continuous, the attribute threshold segmentation can be found;
- *Step*7: For each new leaf node by node N

{

If the sample subset of T ' of leaf node corresponding is empty, breaking the leaf node to generate new leaves, to mark it as in most classes of T

else

On the leaf nodes perform C4.5 formtree (T 'T', an attributelist),

Continue to split;

}

- *Step*8: The classification error of each node is calculating.

by the tree pruning.

Here, using the information gain rate (information gain ratio) as a standard of choice testing attributes. Information gain rate is equal to the information gain ratio of segmentation information (Split information). *A* is a hypothesis of sample set T with different values of discrete attributes, divided *A* into a total of a subset $S_1, S_2, \cdots, S_n$ of segmentation information given by the following formula:

$$Split(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2 \left(\frac{s_i}{s} \right) \tag{1}$$

Information gain rate is given by the follows:

$$Gain - Ratio(S, A) = \frac{Gain(S, A)}{Split(S, A)} \tag{2}$$

Choosing the biggest attribute *A* as a branch of attributes, it can not only deal with discrete attributes, but also continuous attributes.

When calculating the expected information entropy, only need to treat the samples that is known testing value of an attribute and then multiplied the ratio as the expectations of the whole training set information entropy by the proportion of these samples in the training of the current node. When calculating the division of information entropy, the loss of the test attribute value samples treated as a new category, separate by the expectations of the samples information entropy. In

dividing into training set, first it is to be in accordance with the general algorithm to divide samples of the normal into several subsets, then put the missing test sample according to certain probability distribution of attribute values to each subset.

The sample in subset of the loss testing attribute value with test samples is to keep a proportion of attribute values: when the test case classification attribute values of the unknown and the case through all branch, then the result can be gotten by calculating it probability of distribution on the class instead of a certain class. In the end, the result can be obtained from that classes with maximum probability.

IV. DECISION ANALYSIS BASED ON THE DATA OF CET4

Based on the grading and classification of the decision tree algorithm, an analysis on the student CET4 is given.

CET4 is a national English test that is executive director of the ministry of education. The purpose is to accurate measure students' practical English ability for objective and provide assessment services for college English teaching. College English test is a large-scale standard test. Ii is a "standard related norm criterion-referenced test".

CET4 of performance evaluation of the current is used mostly by manual calculation method, usually organized by teachers to evaluate students by educational administration department after the result come out. The educational administration department according to the evaluation results makes corresponding arrangements based on statistics. Under this evaluation method it can only obtain a simple evaluation results, but can't analyze the assessment data, unable to implement performance evaluation giving full play to the effect of teaching.

Because the CET4 levels of achievement is a kind of information, performance management process to information acquisition, storage, transmission, processing, output, process, most of school in grades, reporting, classification, registration custody after work is over, and no in-depth analysis from a large number of achievements, to capture the message is beneficial to teaching. Student performance analysis of the limitations of traditional difference in the average, methods, significance test, etc. It is often based on considering the teaching itself. In fact, there are some subtle factors in teaching, these are all need further analysis to come to the conclusion. Based on these conclusions can make corresponding decision.

As an example to "students' basic information table" and "students' CET4 score table", the factors that affect students' CET4 grades have been analyzed.

First the decision tree model is to be generated based on the correlation attribute decision algorithm.

A. Class Attribute Information Entropy

First, the information entropy of category attributes is calculated as follows:

$$I = -\sum p_i \log p_i = 0.7834$$

B. The Category Attribute Information Gain Rate

Example of the category attribute "listening scores". The values of A, B, C, D have been given in statistics. The information entropy of grades is as follows:

$$E(\text{listening}, T) = 0.4766$$

The amount of information is divided as:

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right) = 1.6306$$

$$Gain, ratio(listening) = \frac{Gain(listening)}{Split(listening)} = \frac{I(listening) - E(listening)}{Split(listening)} = 0.7077$$

If "reading scores" is selected as test attributes values of A, B, C, D, the information entropy of grades is as follows:

$$E(\text{listening}, T) = 0.5745$$

Dividing the amount of information for:

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right) = 1.667$$

$$Gain, ratio(reading) = \frac{Gain(reading)}{Split(reading)} = \frac{I(reading) - E(reading)}{Split(reading)} = 0.6552$$

If "writing achievement" is selected as test attributes values of A, B, C, D, the information entropy of grades is as follows:

$$E(\text{writing}, T) = 0.6112$$

Dividing the amount of information for:

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right) = 1.8551$$

$$Gain, ratio(writting) = \frac{Gain(writting)}{Split(writting)} = \frac{I(writting) - E(writting)}{Split(writting)} = 0.6705$$

If "comprehensive test" is selected as test attributes values of A, B, C, D, the information entropy of grades is as follows:

$$E(Comprehensive, T) = 0.7167$$

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right) = 1.9088$$

$$Gain, ratio(comprehensive) = \frac{Gain(comprehensive)}{Split(comprehensive)} = \frac{I(comprehensive) - E(comprehensive)}{Split(comprehensive)} = 0.6245$$

C. The Comparative Information Gain Rate

By compared with the four results of information gain ratio:

a) Listening

$$Gain, ratio(listening) = 0.7077$$

b) Reading

$$Gain, ratio(reading) = 0.6552$$

c) Writing

$$Gain, ratio(writting) = 0.6705$$

d) Comprehensive

$$Gain, ratio(comprehensive) = 0.6245$$

From the above calculation results can be seen that the biggest one is 0.7077. It is namely listening information contained in this classification largest. So, it should be to choose listening performance to be attributes as the root of the decision tree.

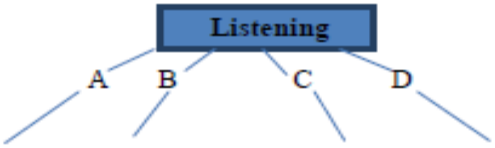

Figure 1. Listening as the root of the decision tree

D. Using a Recursive Method to Establish the Decision Tree

According to the listening ability for the root node, the training set can be divided into four subsets A, B, C, D, four leaf node generated. Of the other three leaf nodes using the recursive method to calculate each attribute of conditional entropy and information gain.

The conditional entropy is:

$$E(reading, T) = 1.1259$$

Dividing the amount of information for:

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right) = 1.6591$$

$$Gain, ratio(reading) = \frac{Gain(reading)}{Split(reading)} = \frac{I(reading) - E(reading)}{Split(reading)} = 0.3214$$

If choose "writing" as the test attribute, the conditional entropy is:

$$E(\text{writing}, T) = 0.7099$$

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right) = 1.6423$$

Dividing the amount of information for:

$$Gain, ratio(writting) = \frac{Gain(writting)}{Split(writting)} = \frac{I(writting) - E(writting)}{Split(writting)} = 0.5677$$

If choose "comprehensive" as the test attribute, the conditional entropy is:

$$E(Comprehensive, T) = 1.1621$$

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right) = 1.8663$$

Dividing the amount of information for:

$$Gain, ratio(comprehensive) = \frac{Gain(comprehensive)}{Split(comprehensive)} = \frac{I(comprehensive) - E(comprehensive)}{Split(comprehensive)} = 0.3773$$

From the results above, the largest gain ratio is 0.5677, when grades is to choose medium, the comprehensive interest should be as the test attribute.

THE RESULT ANALYSIS

By extracting the classification rules, the results have been come to the conclusion as follows:

- In all students, the "excellent" ones are more than the "average" students in the English four levels of achievement (in accordance with the law of normal distribution);
- The normal impact mainly for listening and reading in English four levels of achievement, and for the information data, listening ability is the main influencing factors. students are all the fine if listening scores is "A" .If listening scores is "B", writing achievement can dominate weather the student is outstanding. If listening scores is "C", "D", reading scores will be as main influence factors;
- Weather male and female, students listening ability is generally weak, fewer proficiency;
- The proportion of girls than boys of excellent students in CET4.(generally speaking, girls language ability is better than boys learn seriously, conform to the normal rule);
- The girl reading ability is generally better than boys, boys ability of comprehensive test is generally better than girls;
- Writing is the weak link of all students, fewer proficiency.

In view of the above situation, in the subsequent mainly strengthen listening learning in English teaching, in order to improve the proficiency. For boys and girls at the same time for training and comprehensive training scheme respectively, all students should be to establish the hearing training classes, to strengthen the hearing training. Training classes for all students should be to add writing in order to improve the level

of writing. For the girl it should strengthen the cultivation of comprehensive ability. For boys should add reading training to improve the reading ability. For a specific student should be carry on the weak links of strengthening practice to the purpose of improving the English level 4 grades.

The algorithm have been given based on decision tree in the paper. By generating data mining classification rules, the data of students' CET4 as an example to generate the decision tree model and find out the factors that influence the English four levels of achievement. By using the information entry gain ratio in dealing with classification, the accuracy of the algorithm has been improved. . At the same time, to solve the problem of how to achieve good grades for s specific student on English learning. The numerical result shows that the decision tree algorithm is effectively.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of Liaoning Province of China under grant No.2013020013.

REFERENCE

[1] Backus P, Janakiram M, Mowzoon S, Runger G C, Bhargava , “Factory cycle-time prediction with a data-mining approach”, IEEE Transactions on Semiconductor Manufacturing, 19(2),2006, pp. 252-258.

[2] Lu Qiu,Cheng Xiao zhui, “Parallelization of decision tree algorithm based on MapReduce”, Journal of Computer Applications, vol. 32. No.9, 2012, pp.2463–2465,2469.

[3] XU Jian-Xin HOU Zhong-Sheng, “Notes on data-driven system approaches”, ACTA Automatica Sinica, Vol. 35, No. 6,2009, pp.668-674.

[4] Hou Zhong-Sheng, Xu Jian-Xin. “On data-driven control theory: the state of the art and perspective”,Acta Automatica Sinica, 35(6) 2009, pp.650-667.

[5] Xu Jun, “The study of the discretization method in decision tree”, Journal of Hebei Institute of Technology, Vol. 29, No.2, 2007, pp.71-74.

[6] Han Xi-Xian,Li jian-Zhong,Gao Hong. “A efficient Top-k Dpminating Algorithm on Massive Data Title”,Chinese Journal of Computers, Vol. 36, No. 10, 2013,pp.2132-2145.