Research on Application of C4.5 Algorithm in Performance Analysis

Kai Lu[1, a *], Mingrui Chen[2, b]

[1] Technology department of public safety, Hainan Vocational College of Political Science and Law Haikou, China

[2]College of Information Science & Technology, Hainan University Haikou, China

[a]294870140@qq.com, [b]1607885098@qq.com

Keywords: Data mining; C4.5 algorithm; Classification rules; Information gain rate; Decision tree

Abstract. C4.5 algorithm is one of the ten classic algorithms of data mining, it is a series of algorithms used in machine learning and data mining. This paper mainly studies the application of C4.5 algorithm to the analysis of performance data, to dig out the hidden relationship between various factors and test results, to provide a fair and objective analysis of the quality of the teaching assessment and provide decision support for the teaching improvement in the future. This paper focuses on the C4.5 algorithm, including the division rules, the calculation of the information gain rate and algorithm workflow, and finally through an example to explain the specific application of the C4.5 algorithm.

Introduction

Traditional performance analysis mainly depends on the software comes with the query, the distribution of statistics, ranking, data backup and other primary functions, and for the achievement of the data contains a lot of useful information can not be used, thus easily lead to the waste of data. Data mining is the analysis of large data sets or databases, extraction or mining unknown, potential, valuable resources or information for the user to provide the basis for the user's decision.

C4.5 algorithm is one of the ten classic algorithms of data mining, it is an improvement of the ID3 algorithm, proposed by J.R.Quinlan in 1993 [1]. In the C4.5 algorithm, workflow and ID3 algorithm are basically the same, while improving the ID3 algorithm in the existing problems. Studies using C4.5 algorithm results analysis, dig out the internal relations between the various factors and test results, for teaching quality assessment is to provide objective and fair the full range of deep-seated analysis and can provide decision support for the future teaching improvement is very necessary [2].

Algorithm Principle

The core idea of C4.5 algorithm is that using information entropy theory and takes the attribute of the maximum information gain is selected as the classification attribute, construction of branch of decision tree, and the use of recursive way to construct the whole decision tree, information gain ratio equals to the ratio of the information gain and the segmentation information.

Splitting rule. The splitting rule also called the attribute selection metric, because they determine how the tuple on a given node split. Attribute selection metrics provide the rank of each attribute to describe the given training tuple, and the attribute that has the best measure score is selected as the splitting attribute of the given tuple [3]. At present, the popular attribute selection measures have information gain and gain rate.

Set S is a class labeled tuple training sample data set, class label attribute has m different values, m different classes of Ci (i=1,2,... m), CiS is a set of S classes of Ci, |S| and |CiS| are the number of S and CiS in the tuple.

On the training sample data set S, assuming that A is a different value of $\{a_1, a_2, \ldots, a_v\}$ discrete attributes, the data set S is divided into V subset S_i (i= 1, 2,... V), using A to split the sample data set to get the information gain algorithm with the same ID3.

Computing information gain.

Information gain is actually used in the ID3 algorithm for attribute selection metric. It selects the attribute with the highest information gain as the splitting attribute of the node n. This property makes the amount of information needed for the classification of the results in the classification of the results [4]. The desired information for the tuple classification in D is the following:

$$\text{Info(S)}=-\sum_{i=1}^{m} p_i \log_2(p_i). \tag{1}$$

Info (S) is also called entropy.

It is now assumed that the tuple in S is partitioned according to the attribute A, and the attribute A divides S into V different classes. In this division, in order to get the accurate classification also need to measure information by the following Eq. 2:

$$\text{Info}_A(S)=\sum_{i=1}^{v} \frac{|s_i|}{|s|} \times \text{Info}(S_i) \tag{2}$$

The information gain is defined as the difference between the original information requirement (i.e., only the class ratio) and the new requirement (i.e., the difference between the A Division),that is Eq. 3:

$$\text{Gain(A)}= \text{Info(S)}- \text{Info}_A(S) \tag{3}$$

Calculation Segmentation information, calculated by the following Eq. 4:

$$\text{SplitInfo}_A(S)=-\sum_{i=1}^{v} \frac{s_i}{s} \log_2(\frac{s_i}{s}) \tag{4}$$

This value represents the information generated by dividing the training data set S into a V partition that corresponds to the V output of the attribute A test [5].

Calculate the information gain rate, calculated by the following Eq. 5:

$$\text{GainRatio(A)}= \frac{Gain(A)}{SplitInfo(A)} \tag{5}$$

C4.5 algorithm determines the decision attribute by calculating the information gain rate of each attribute in the training data set, select Eq.5 of maximum value as decision attribute, so as to creating branch node, and construct the decision tree [6].

Working flow of C4.5 Algorithm. exactly ,C4.5 is not a single algorithm, but a set of algorithm, C4.5 has many features, every feature corresponds to an algorithm that these functions are combined to form a set of is C4.5 algorithm [7]. According to the previous description of the algorithm, it is now given the C4.5 classification tree construction algorithm workflow shown in Fig. 1:

Algorithm 1.1 C4.5(S)
Input: an attribute-valued dataset *S*
1: Tree={} 2: **if** *S* is ”pure” OR other stopping criteria met **then** 3: terminate 4: **end if** 5: for all attribute a $\in$ *S* **do** 6: Compute information-theoretic criteria if we split on a 7: **end for** 8: a_{best}=Best attribute according to above computed criteria 9: Tree=Create a decision node that tests a_{best} in the root

10: S_v=Induced sub-datasets from S based on a_{best}
11: for all S_v **do**
12: $Tree_v$=C4.5(S_v)
13: Attach $Tree_v$ to the corresponding branch of Tree
14: **end for**
15: **return** Tree

Figure 1. Working flow of C4.5 classification tree construction algorithm

Algorithm Application. To the authors of this semester teaching of "web design and production" course as an example, through the medicine before class preparation, classroom learning, time to review after class, interest courses etc. information and according to the final results of the survey results are summarized, obtain the " student achievement is whether good data sheet " as shown in Table 1.

Table 1 Student achievement is whether good data sheet

Preview before class time	Classroom learning	Review after class time	Grades	Whether or not good
2 hours or more	master	2 hour or less	medium	yes
1 to 2 hours	general	2 to 4 hours	medium	no
1 hour or less	master	2 to 4 hours	medium	yes
2 hours or more	master	4 hours or more	good	yes
1 to 2 hours	poor	2 hour or less	poor	no
1 to 2 hours	general	4 hours or more	medium	no
1 hour or less	master	2 hour or less	good	yes
2 hours or more	poor	2 to 4 hours	medium	no
1 hour or less	general	2 hour or less	poor	no
1 to 2 hours	poor	2 to 4 hours	good	yes
1 to 2 hours	general	2 to 4 hours	poor	no
…	…	…	…	…

50 records listed in the table (student number), of which 16 positive cases (excellent), 34 counterexamples (not good). Using C4.5 algorithm to analyze the data in the table, and calculate the information gain rate each attribute in the table, according to the calculation results, the root node of the decision tree selection "grades" attribute of the maximum information gain rate, after calculated "grades" attribute of the "good", "medium", "bad" three attribute value information gain rate, obtain "student achievement is whether good decision tree", as shown in Fig. 2 [8].

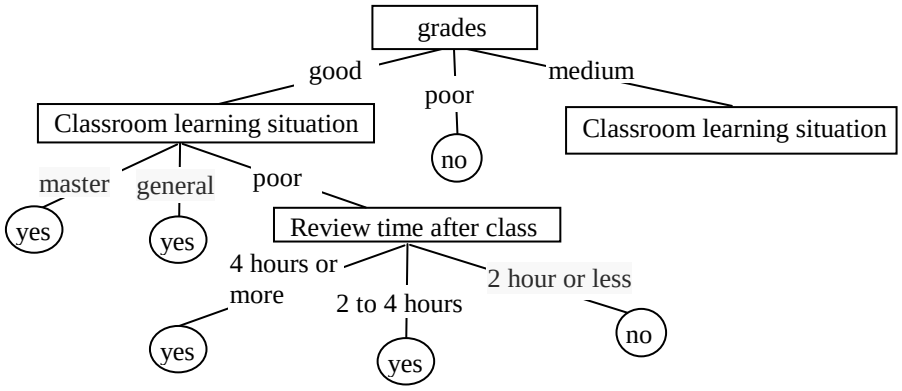

Figure 2. Student achievement is whether good decision tree

When we build that “student achievement whether to pass the decision tree” model, in this paper, the "preview before class time", "classroom learning", "review after class time", "grades", "final grade “ five the attribute of the close relation with student achievement as property field of decision tree, to facilitate establishing decision tree model, to " whether to pass "as a class attribute, obtain " student achievement whether to pass the data table ", as shown in Table 2.

Table 2 Student achievement whether to pass the data table

Preview before class time	Classroom learning	Review after class time	Regular grade	Whether to pass
2 hours or more	master	2 hour or less	good	yes
1 to 2 hours	general	2 to 4 hours	good	yes
1 hour or less	master	2 to 4 hours	medium	yes
2 hours or more	master	4 hours or more	good	yes
1 to 2 hours	poor	2 hour or less	poor	no
1 to 2 hours	general	4 hours or more	medium	yes
1 hour or less	master	2 hour or less	good	yes
2 hours or more	poor	2 to 4 hours	medium	yes
1 to 2 hours	general	2 to 4 hours	good	yes
1 to 2 hours	general	2 hour or less	poor	no
1 to 2 hours	poor	2 to 4 hours	good	yes
2 hours or more	general	4 hours or more	poor	yes
1 to 2 hours	general	4 hours or more	medium	yes
2 hours or more	general	2 to 4 hours	medium	yes
…	…	…	…	…

50 records listed in the table (student number), of which 46 positive cases (pass), 4 counterexamples (fail). Also by using C4.5 algorithm to analyze the data in the table, and calculate the information gain rate each attribute in the table, "grades" attribute information gain ratio of the maximum value is obtained. Will "grades" attribute as the root node of the decision tree, according to the "grades" attribute of the "good", "medium", "bad" three attribute values, respectively calculator information gain rate, get "student achievement whether to pass decision tree" is shown in Fig. 3 [9].

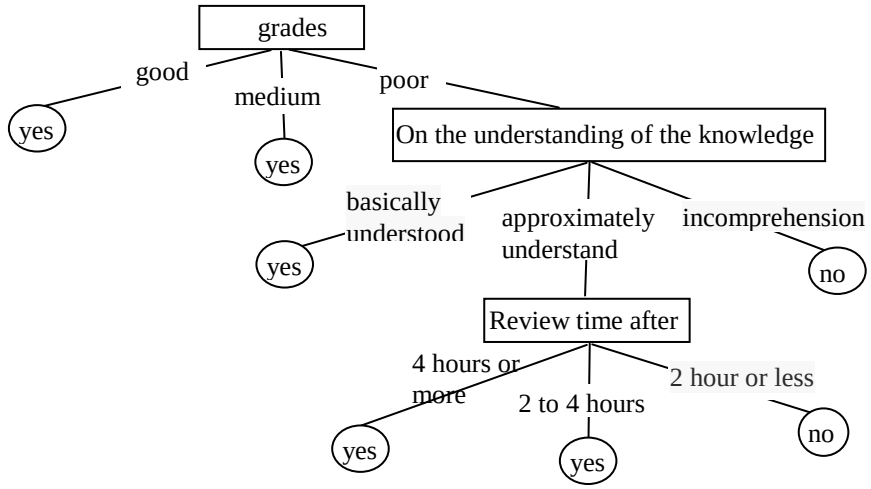

Figure 3. Student achievement whether to pass decision tree

Through the above example shows that data mining technology is applied to the analysis of student achievement, to the intrinsic link between the comprehensive analysis of test scores and all kinds of

influencing factors, effectively guide the students to learning, which is not available in the traditional performance evaluation methods [10].

Summary

This article introduced the C4.5 algorithm principle and working process in detail, and the C4.5 decision tree in data mining technology is applied to the analysis of student achievement, by extracting the characteristic information of student data, a comprehensive analysis of the internal relations between students' performance and various factors, through the analysis of the system, teachers can adjust teaching strategies to improve the teaching quality, students can adjust learning methods to improve the learning efficiency.

Acknowledgements

This paper is written for the natural science foundation of Hainan Province. The project name is “The Research and Development of Intelligent Examination System Based on Improved ACO and C4.5 Algorithm”, which was applied in 2014 and the project number is 614243.

References

[1] B. Yang: Application of C4.5 algorithm in the analysis of higher vocational school [J]. Journal of biotechnology and computer science, Vol.12 (2015), p.127.
[2] W.B Li: *The Top the Algorithms in Data Mining* (MS., Tsinghai university press, China 2013), P.34.
[3] A.X. He, X.S.YUAN: Application of C4.5 decision tree algorithm in the employment management of application oriented Undergraduate Colleges and universities [J]. Journal of Chuzhou University, Vol.5 (2012), p.87.
[4] H.M. Liu: *Application of data mining in the analysis of students' performance in Higher Vocational Colleges* (MS., Anhui University, China 2011), p.25.
[5] K. Tao: The application of decision tree technology in the academic performance analysis [J]. Journal of Xinxiang College, Vol.2 (2011), p.217.
[6] Y. Chen: *Data mining technology and application* (MS., Tsinghai university press, China 2011), p.16.
[7] H.P. Bain: Based on the improved C4.5 algorithm application in performance analysis [J], computer knowledge and technology, Vol. 11(2015) No. 27, p.263.
[8] X.F. Li, etc: *The data warehouse and data mining* (mechanical industry publishing house, China 2013), p.209.
[9] H.Q. tang, etc: Mobile phone terminal design based on C4.5 decision tree algorithm for weather warning system [J], computer application, Vol. 33(2013) No. 5,p.160.
[10] Y.L. Fu: Application of data mining C4.5 algorithm in performance analysis [J]. Journal of Chongqing University of Technology, Vol. 11(2013), p.148.