The Research and Application of Improved Decision Tree Algorithm in University Performance Analysis

Wushi Gao, Yunfeng Dong
Computer Center
Shandong Polytechnic University
Jinan, China
E-mail: gws@spu.edu.cn, dyf@spu.edu.cn

Kan Li
Network Center
Shandong Polytechnic University
Jinan, China
E-mail: Likan@spu.edu.cn

***Abstract*—with the promoting of education informationization process, more and more complex data is accumulated, the data mining techniques used in education. The work, that finding in hidden and useful information to guide education from large number of educational data, will be helpful to the education reform and development. This article proposes one kind of improvement ID3 algorithm, this algorithm simplified information entropy solution, which is the standard of attribute selection and reduced complication of calculation. Its application to analysis of university students performance will find out the influence of implicit factors, improve the quality of teaching.**

Keywords-Decision tree; ID3 algorithm; performance analysis

I. INTRODUCTION

Student performance is a measure of whether students master knowledge, also is the important basis for evaluating teaching quality. Using data mining techniques can convert amounts of data into classification rules, which can accurately show the student performance analysis of multiple aspects, and comprehensively analysis of the test results and the hidden relationship between various factors through a better analysis of these data. The introduction of data mining technology into the student performance analysis will help teachers and teaching departments to formulate the corresponding measures, be helpful to improve the teaching quality and enhance the teaching effect.

II. THE OVERVIEW OF DECISION TREE TECHNOLOGY

Data mining is the process of extract potentially useful, credible information and knowledge from amounts of noisy, fuzzy and random raw-data. Decision tree, an algorithm common used to predict model, can find out some valuable information through huge amounts of data classification.

The decision tree is the basis of learning example inductive learning algorithm, it through the huge amounts of data classification to find some valuable information.

ID3 algorithm is one important method in the technology of decision tree classification and so is widely applied. ID3 algorithm searches through attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n = number of possible values of an attribute) partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.

The central principle of ID3 algorithm is based on information theory.

Given a training dataset S, the information entropy of the set S is defined as:

$$\mathrm{I}(\mathrm{p},n)=\left(-\frac{p}{p+n}\log_2\frac{p}{p+n}-\frac{n}{p+n}\log_2\frac{n}{p+n}\right) \quad (1)$$

Where p is the number of positive example, n is the number of negative example.

Let us suppose that attribute A ∈{A1,A2……Av}, S is divided into a number of disjoint subsets{ S1,S2…Sv },where Si have pi positive examples and ni negative examples, so the desired information entropy with A as the root is defined as:

$$E(A)=\sum_{i}^{v}\frac{p_i+n_i}{p+n}\mathrm{I}(\mathrm{p}_i,n_i) \quad (2)$$

So the gain of example set S on attribute A is:

$$Gain\ (A)=\mathrm{I}(p,n)-E(A) \quad (3)$$

Id3 choose the maximum attribute of Gain (A) as root-node [1].

III. IMPROVED ID3 ALGORITHM

Every time of choosing a split node, the ID3 algorithm is related to multiple logarithm operation that will obviously affect the decision tree generation efficiency in times of amount of data to operate. Therefore we will consider to change selection criteria from data attribute so that reduce the computational cost of saving decision tree and decision tree generation time. In addition, the choice by ID3 often turn to attributes with more values, because it use each attribute information entropy to judge the value of the data of the division of concentrated properties.

According to the basic principle and algorithm of decision tree based on information theory, we converted the formula of information gain, so as to find a new attribute-select criterion. This new standard choosing attributes can not only overcome the ID3 algorithm shortcomings, which easily tend to choose more different values attribute as the test attributes ,but also reduce generation time of the decision tree and calculated cost greatly, so that accelerate construction speed of decision tree, improve the efficiency of decision tree classifier.

According (1) (2), we can get

$$E(A)=\sum_{i}^{v}\frac{1}{(p+n)\ln 2}(-p_i\ln\frac{p_i}{p_i+n_i},-n_i\ln\frac{n_i}{p_i+n_i}) \quad (4)$$

Because (p+n) ln2 is a constant in training set, so we can assume that the function e(A) satisfies the following equation:

$$e(A)=\sum_{i}^{v}(-p_i\ln\frac{p_i}{p_i+n_i},-n_i\ln\frac{n_i}{p_i+n_i}) \quad (5)$$

Using McLaughlin formula:

$$f(x)\approx f(0)+f^{1}(0)x+\frac{f^{11}(0)}{2!}x^2+\ldots\frac{f^{(n)}(0)}{n!}x^n \quad (6)$$

When f(x)=ln(1+x) and x is very small, we can get: ln(+x)≈x

So we simplify e(A) as：

$$\ln\frac{p_i}{p_i+n_i}=\ln(1-\frac{n_i}{p_i+n_i})\approx-\frac{n_i}{p_i+n_i}$$

$$\ln\frac{n_i}{p_i+n_i}=\ln(1-\frac{p_i}{p_i+n_i})\approx-\frac{p_i}{p_i+n_i}$$

Put top two equations into e (A), we can get:

$$e(A)=\sum_{i}^{v}(p_i\frac{n_i}{p_i+n_i}+n_i\frac{p_i}{p_i+n_i})=\sum_{i}^{v}\frac{2p_in_i}{p_i+n_i} \quad (7)$$

Assuming that each of the number of attributes is N, so the improved attribute information entropy formula is:

$$e(A)=(\sum_{i}^{v}\frac{2p_in_i}{p_i+n_i})N \quad (8)$$

At last, we show the flow chart of information entropy calculation as the following Fig.1, which is the key part of whole improved ID3 algorithm.

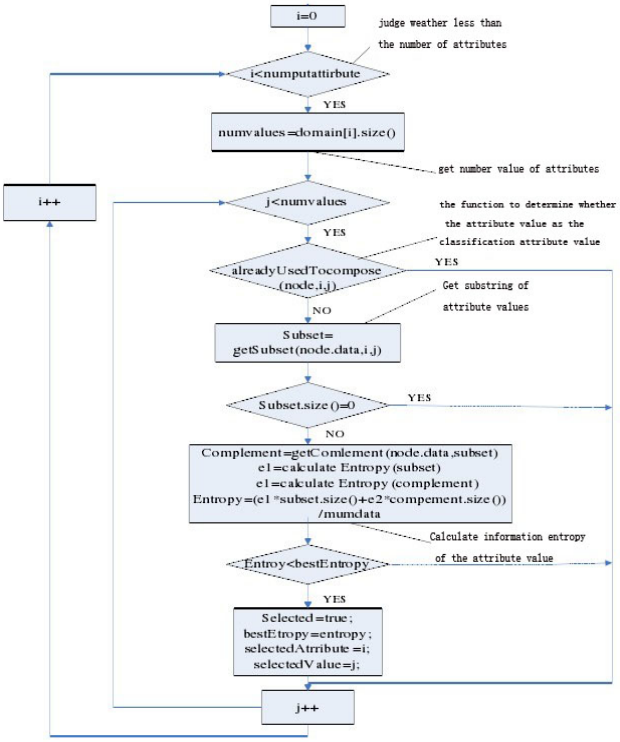

Fig1. flow chart of improved ID3 algorithm

IV. THE APPLICATION AND PERFORMANCE ANALYSIS OF ID3 ALGORITHM IN UNIVERSITY PERFORMANCE ANALYSIS

A. *Data mining with improved ID3 and traditional ID3*

We choose the students ' final results of grade 2009 as a data object. After data preprocessing, the formation of students' examination results the training set, and then respectively by using the traditional ID3 and the improved ID3 for data mining in order to construct decision tree.

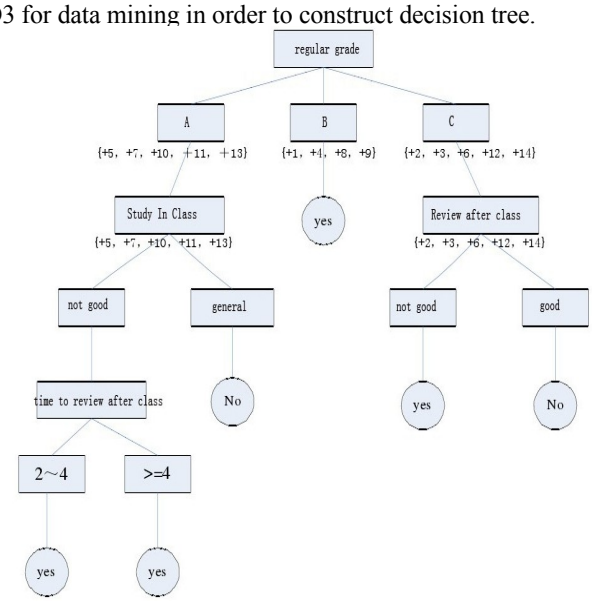

Fig2. decision tree of the traditional ID3 algorithm

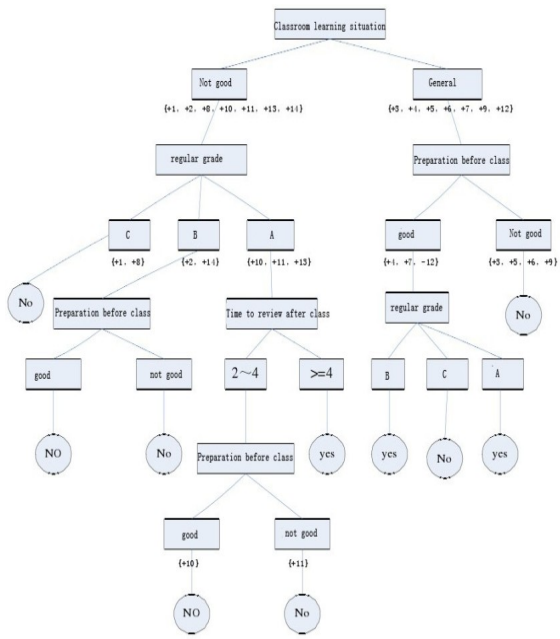

Fig3. decision tree of the improved ID3 algorithm

It is obvious from the Fig.2 and Fig.3, the improved ID3 algorithm overcomes the defects of the traditional ID3 algorithm easy to select attributes with more value as splitting attribute node, but also fully in line with the [IF—THEN] sentence criterion, the scale of improved ID3 algorithm in decision tree is smaller than traditional ID3 algorithm, and improved ID3 algorithm is relatively simple, easy to understand.

B. Classification performance analysis

We select 4 datasets to test the traditional ID3 algorithm and improved ID3 algorithm. Comparative analysis of the improved ID3 algorithm and the ID3 differences from the nodes-number, regular-number, accuracy and cost time. Each dataset is conducted 20 experiments, and then calculate the average value, so the experimental data with more generality.

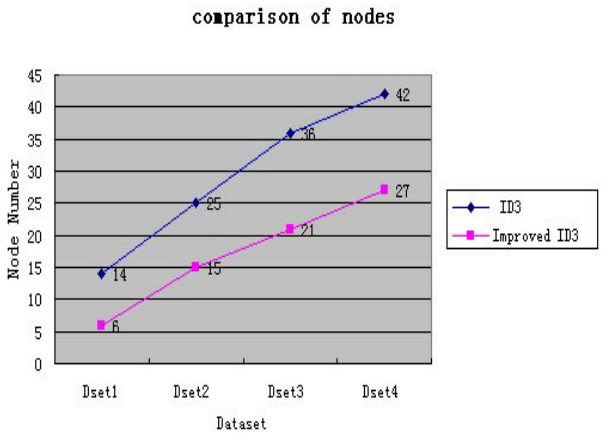

Fig4. comparison of number nodes-number

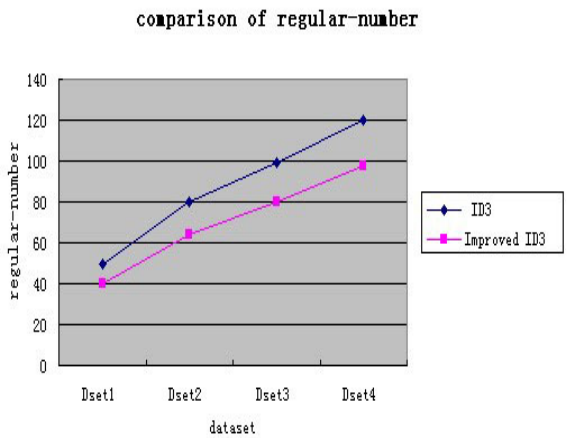

Fig5. comparison of regular-number

From Fig.4 and Fig.5, we can see the scale and regular-number of improved ID3 algorithm in decision tree is smaller than traditional ID3 algorithm. This advantage is more obvious when the collection of examples is bigger.

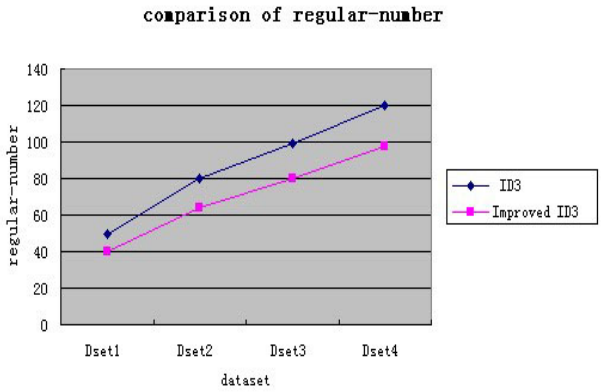

Fig6. comparison of accuracy

Fig6 can be seen that the improved ID3 algorithm accuracy is higher than the original ID3 algorithm. And time difference increasing linearly with the increase of data quantity, but the improved ID3 algorithm with the increase tendency of data quantity decline a little bit compared with traditional ID3 algorithm.

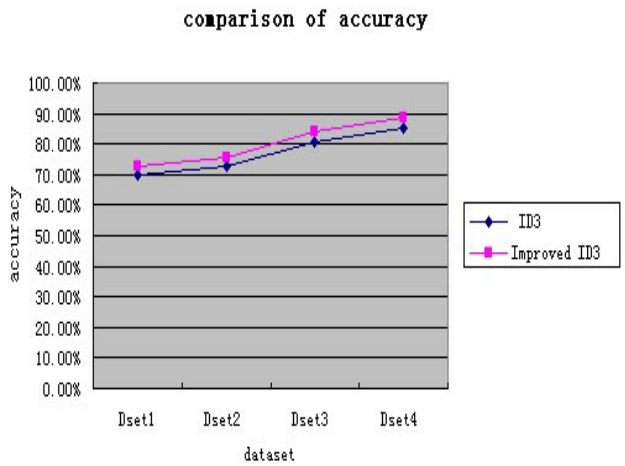

Fig7. comparison of cost time

From fig 7,we can see that improved ID3 algorithm spending less time than traditional ID3 algorithm. Time difference increase linearly with the increase of data quantity, this is fully indicate that the improved ID3 algorithm in efficiency and performance has much more superiority than the original ID3 algorithm in the treatment of the large dataset in the process of constructing decision trees.

V. CONCLUSION

The improved ID3 algorithm can construct more concise decision tree classification model, it's time complexity and cost time in creating decision tree is superior to traditional ID3.The improved ID3 algorithm overcome the traditional ID3 algorithm's shortcomings which easily tend to choose more different values attribute as the test attributes, that can

make the structure of decision tree more compact, get a good classification effect and performance

REFERENCES

[1] Jiawei Han,Micheline Kamber . Data mining:concepts and techniques. MorganKaufmann. 2006. 58-61

[2] Ian H.Witten,Eibe Frank. Data mining:practical machine learning tools and techniques. secondedition. San Francisco, Morgan Kaufmann Publishers, 2005：85-90

[3] Mehmed Kantardzic, Jozef Zurada. Next Generation of Data-Mining Applications. New York：Wiley-IEEE Press, 2005. 3.

[4] Jearanaitanakij K Classifying continuous data set by ID3algorithm . Proc of Fifth InternationalConference on Information, Communica-tions and Signal.2005, （4）：1048-1051

[5] WANG Hong-rui,ZHAO Li-ming,PEI Jian. Equilibrium Modified K-Means Clustering Method. Journal of Jilin University(Information Science Edition) 2006.2： 171-176

[6] Wedel M,Kanmakura W.Marketing data,models anddecision[J].Research in Marketing,2000,17:203-208.

[7] J.MacQueen. Some methods for classification and analysis of multivariate observations. In Proc.of the 5th Berkeley Symp.On Mathematical Statistics and Probability, University of California Press, 1967:281-297

[8] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining. New Jersey：Addison Wesley/Pearson, 2005：135-165

[9] Zhang Gui-Jie,Wang Shuai. Decision Tree Classification. Jilin Normal University Journal(Natural Science Edition). 2008.3

[10] WU Tong ,Wang Xiu-kun. Decision tree algorithm use to forecast and analyse for students' marks . Microcomputer Information. 2010.3