Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques

Mr. M. N. Quadri1

Dr. N.V. Kalyankar[2]

GJCST Computing Classification
H.2.8 & K.3.m

***Abstract-* Students' academic performance is critical for educational institutions because strategic programs can be planned in improving or maintaining students' performance during their period of studies in the institutions. The academic performance in this study is measured by their cumulative grade point average (CGPA) upon graduating. This study presents the work of data mining in predicting the drop out feature of students. This study applies decision tree technique to choose the best prediction and analysis. The list of students who are predicted as likely to drop out from college by data mining is then turned over to teachers and management for direct or indirect intervention.**

Keywords- Intruder; hacker; cracker; Intrusion detection; anomaly detection; verification; validation.

I INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It is defined as "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Due to rapid advancement in the field of information technology, the amount of information stored in educational databases is rapidly increasing. These huge databases contain a wealth of data and constitute a potential goldmine of valuable information. As new courses and new colleges emerge in environment, the structure of the educational database changes. Finding the valuable information hidden in those databases and identifying and constructing appropriate models is a difficult task. Data mining techniques play an important role at each stop of the information discovery process.

Nowadays, higher educational organizations are placing in a very high competitive environment and are aiming to get more competitive advantages over the other business competitions. These organizations should improve the quality of their services and satisfy their customers. They consider students and teachers as their main assets and they want to. improve their key process indicators by effective and efficient use of their assets

The most striking features of data mining techniques are clustering and prediction. The clustering aspect of data mining offers comprehensive characteristics analysis of students, while the predicting function estimate the likelihood for a variety of outcomes of them, such as transferability, persistence, retention and success in classes.

This study makes use of decision tree analysis to analyze the problem of drop outs in any educational institution.

Decision tree analysis is a popular data mining technique that can be used in many areas of education. In this study, decision trees are used to make important design decisions and explain the interdependencies among the properties of drop out students. This study also provides examples of how data mining technique can be used to improve the effectiveness and efficiency of the modeling process.

This study is an extension of the educational model developed and published in the information technology journal[1] . The main contribution in this study is addressing the capabilities and strengths of data mining technology in identifying drop out students and to guide the teachers to concentrate on appropriate features associated and counsel the students or arrange for financial aid to them.

II APPLICATIONS OF DATA MINING IN EDUCATIONAL INDUSTRY

Identify risk factors that predict results: One critical question in any educational institution is the following "What are the risk factors or variables that are important for predicting the results (pass/fail) of students?". Although many risk factors that affect results are obvious, subtle and non-intuitive relationships can exist among variable that are difficult, for not impossible to identify without applying more sophisticated analysis.

Modern data mining models such as decision trees can more accurately predict risk than current models, educational institutions can predict the results more accurately, which in turn can result in quality education.

Student Level Analysis: Successfully training the student requires analyzing the data at the student level. Using the associated discovery data mining technique, educational institutions can more accurately select the kind of training to offer to different kinds of students. With the help of this technique, educational institutions can.

i. Segment the student database to create student profiles.

Manuscript received "Thu, Feb 11, 2010 at 9:05 PM GMT"

Mr. M.N. Quadri is working as Lecturer with Dept. of Computer Science Yeshwant College Nanded
E-Mail : mnq_1977@yahoo.com
Dr. N.V. Kalyankar is working as Principal, Yeshwant College Nanded.
E-Mail: drkalyankarnv@rediffmail.com

ii. Conduct analysis on a single student segment for a single factor. For example, “the institution can perform in-depth analysis of the relationship between attendance and academic achievement”.

iii. Analyze the student segments for multiple factors using group processing and multiple target variables. For example, “What are the characters shared by students who drop out from colleges?”.

iv. Perform sequential (over time) basket analysis on student segments. For example, “What percentage of high attendance holders also achieved in academic side also?”.

Developing new strategies: Teachers can increase the pass percentage by identifying the most lucrative student segments and organize the training sessions accordingly. The results may be affected, if teachers do not offer the “right” kind of training to the “right” student segment at the “right” time. With data mining operations such as segmentation or association analysis, institutions can now utilize all of their available information for betterment of students.

III DROP OUT

Graduation, especially timely graduation is an increasingly important policy issue[2]. College graduates earn twice as much as high school graduates and six times as much as college dropouts[3]. In addition to the financial rewards, the spouses of college graduates are more educated and their children do better in schools and colleges. Graduation rates are considered as one of the institutional effectiveness[4]. Students drop out due to different reasons; academic trouble, academic preferences, marriage (girls) and their financial position.

i. Students are unable to get into the major they prefer when they matriculate and therefore they find it difficult to carry on with the course and may leave the institution due to academic trouble.

ii. Students also drop out due to academic preferences. Generally, students choose majors offering the greatest stream of future earnings.

iii. In Indian society, girls are expected to get married at the age of 18 and they may drop out when they are married.

iv. Financial position of the students plays an important role in drop out percentage.

It is important to understand the determinants of successful and timely degree completion. Most studies of student departure focus on the characteristics of students as determinants of success. The study considers the features such as gender, attendance, previous semester grade, parent education, parent income, scholarship, first child, and part time job.

Parental income: Is an important determinant of the demand for education. Students from higher-income families are less likely to have to drop out to work to finance their education and are most likely to have aspirations that promote persistence. Empirical studies indicate a strong positive correlation between family income and other family background measures on educational attainment enrollment, persistence and graduation[5,6].

Parental education: Plays an important role. Children of college graduates fare well in their exams and are less likely to drop out. A student’s previous semester grade and attendance are also included in the study. Grades and attendance may have some tangible value that can be used for future educational and career mobility. Grades may also be considered as an indication of realized academic potential.

Financial aid/scholarship plays an important role in higher education by lowering the costs of attendance. The study measures the effect of financial aid/scholarship on student departure. The study also investigates about other information such as whether the student is the ‘first child’ in the family and he/she is doing part time job to support the family. Both these variables are expected to be positively correlated with graduation.

IV PREDICTIVE DATA MINING

Decision Trees: A tree diagram contains the following

i. Root node-top node in the tree that contains all observations

ii. Internal nodes-non-terminal nodes (including the root node) that contain the splitting nodes.

iii. Leaf nodes-terminal does that contain the final classification for a set of observations.

Decision trees are part of the induction class of data mining techniques[7]. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modeling, the decision is simply the predicted value.

The tree techniques provide insights into the decision making process[8]. The decision tree is efficient and is thus suitable for large/small data sets. They are perhaps the most successful exploratory method for uncovering deviant data structure. Trees recursively partition the input data space in order to identify segments where the records are homogeneous.

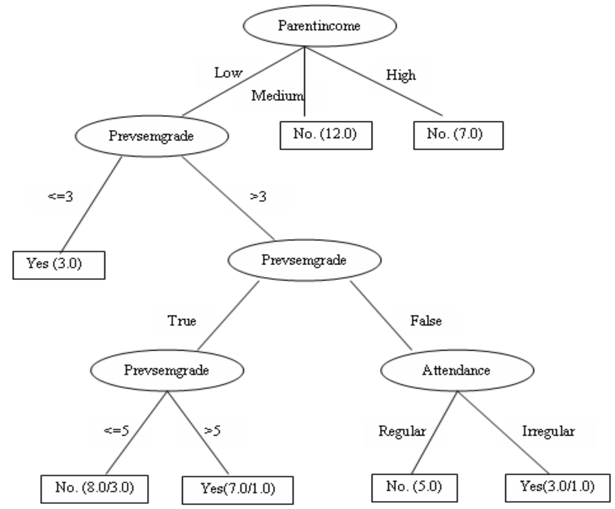
This model, make use of the software Weka The J4.8 algorithm(J4.8 implements a later and slightly improved version called C4.5) is used for predictive data mining.

Modeling student drop outs: The modeling process starts by studying the relationship between student drop outs and underlying risk factors including gender, attendance, previosus semester grade, parent education, parent income, scholarship, first child and whether the student is working or not.

A hybrid method is developed for this study-the modeling process is a combination of the decision tree techniques and logistic regression. First, the decision tree algorithm is used to identify the factors that influence dropouts. After the factors are identified, the logistic regression technique is used to quantify the dropouts and the effect of each risk factor.

The Table 1 shows the variable that influence the drop outs

Variable	Variable Type	Description
Gender	Nominal	Male, Female
Attendance	Nominal	Regular, Irregular
Prevsemgrade	Numeric	1..10
Parentedn	Nominal	Educated, Not educated
Parentincome	Nominal	Low,Medium,High
Scholarship	Nominal	Getting, Not getting
Firstchild	Nominal	True,False
Working	Nominal	Working, Not working
Dropout	Nominal	Yes,No



The Fig. 1 shows the tree diagram for analysis.

The drop out frequency varies with the most important risk factor parent income (in this study) among all the other variables. The low-income level has great influence on the drop out feature, than the medium and high-income levels.

The fact that whether the child is a first child in the family also has an influence on the drop out feature. Through the previous semester grade is above 5, the drop out feature seem to be high due to the responsibility of the student as a first child (for male). Girl students face the problem of getting married as a 'first child' of the family. Based on tree analysis, gender, parent education, scholarship, part time job are the irrelevant factors. They should not be included in the claim frequency model.

Based on tree analysis, logistic regression is used to estimate the probability of drop out feature based on the factors under consideration. Logistic regression attempts to predict the probability of drop out feature as a function of one or more independent inputs.

V Analysis

The analyzer component incorporates a number of machine learning methods for automatically analyzing the data in the log database. In addition to getting a better insight into the underlying relationship in the data, this also allows for prediction and classification of future sessions. Many machine-learning methods provide their output in an intelligible, human readable form. For instance, methods for generating decision trees from data such as C4.5 [9], allow for a tree-shaped representation of the learning results.

The aim of using the decision tree method was to characterize the students' motivation in terms of the other attributes that are automatically generated from the log data, in order to provide an abstracted view on the underlying data as well as to allow for predicting motivational aspects in other students.

VI Conclusion

This study introduced the data mining approach to modeling drop out feature and some implementation of this approach

The key to gaining a competitive advantage in the educational industry is found in recognizing that student databases, if properly managed, analyzed and exploited, are unique, valuable assets. Data mining uses predictive modeling, database segmentation, market basket analysis and combinations to more quickly answer questions with greater accuracy. New strategies can be developed and implemented enabling the educational institutions to transform a wealth of information into a wealth of predictability, stability and profits.

VII References

1) Shyamala, K and S.P. Rajagopalan, 2006. Data mining model for a better higher educational system. Information Tech. J.,5 :560-564
2) DesJardins, S.L., D.A. Ahlburg and B.P. McCall, 2002. A temporal investigation of factors related to timely degree completion. J. Higher Education, 73:555-581.
3) Murphy, K and F. Welch, 1993. Inequality and relative wages. Ameri. Economic review, 83: 104-109.

4) Murtaugh, P.A., L.D. Burns and J. Schuster, 1999 Predicting the retention of university students. Higher Education, 4: 355-357.
5) Kane, J., 1994. College entry by blacks since 1970. The role of college costs, family background and the returns to education. J. Political Econo., 102: 878-911
6) Manski, C. and D. Wise, 1983. College choice in America, Cambridge, MA: Harvard University Press.
7) Quinlan, J.R., 1983. Induction of Machine learning Machine learning, 1:81-106.
8) Han, J. and M. Kambar, 2003. Data mining: Concepts and techniques. Morgan Kaufmann Publishers, New Delhi.
9) Quinlan, R. (1993). C4.5 : Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.