International Journal of Soft Computing 2 (1): 150-156, 2007

Mining Student Data to Characterize Drop out Feature Using Clustering and Decision Tree Techniques

[1]K. Shyamala and [2]S.P. Rajagopalan
[1]Department of Computer Science, Ambedkar Government College, Chennai-600 039, India
[2]Mohammed Sadak Trust, Group of Educational Institutions, Chennai-600 034, India

Abstract: Compared to traditional analytical studies that are often hindsight and aggregate, data mining is forward looking and is oriented to individual students. 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 clustering 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.

Key words: Data mining, decision trees, clustering, drop out

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[1,2].

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 step 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 competitors. 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 estimates 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 and cluster discovery methods 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 techniques can be used to improve the effectiveness and efficiency of the modeling process.
- Cluster analysis is one of the basic techniques that are often applied in analyzing data sets. This study makes use of cluster analysis to segment students into groups and associating a distinct profile with each group, which can help analysis of drop out students.

This study is an extension of the educational model developed and published in the information technology journal[3]. 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.

APPLICATION OF DATA MINING IN EDUCATIONAL INDUSTRY

Identify risk factors that predict results: One critical question in any educational institution is the following.

Corresponding Author: K. Shyamala, Department of Computer Science, Ambedkar Government College, Chennai-600 039, India

"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 variables that are difficult, if not impossible to identify without applying more sophisticated analyses.

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

- Segment the student database to create student profiles.
- 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".
- 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 ?".
- 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.

DROP OUT

Graduation, especially timely graduation is an increasingly important policy issue[4]. College graduates earn twice as much as high school graduates and six times as much as college dropouts[5]. 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[6]. Students drop out due to different reasons; academic trouble, academic preferences, marriage (girls) and their financial position.

- 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.
- Students also drop out due to academic preferences. Generally students choose majors offering the greatest stream of future earnings.
- In Indian society, girls are expected to get married at the age of 18 and they may drop out when they are married.
- 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[7,8].

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.

PREDICTIVE DATA MINING

Decision trees: A tree diagram contains the following items.

- Root node-top node in the tree that contains all observations.
- Internal nodes-non-terminal nodes (including the root node) that contain the splitting nodes.
- Leaf nodes-terminal nodes that contain the final classification for a set of observations.

Decision trees are part of the induction class of data mining techniques[9]. 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.

Tree techniques provide insights into the decision making process[10]. 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 Revision 8) is used for predictive data mining.

Table 1:Shows the variables that influence the drop outs

Variable	Variable type	Description
Gender	Nominal	Male, Female
Attendance	Nominal	Regular, Irregular
Prevsemgrade	Numeric	1..10
Parentedn	Nominal	Educated, Noteducated
Parentincome	Nominal	Low, Medium, High
Scholarship	Nominal	Getting, Notgetting
Firstchild	Nominal	True, False
Working	Nominal	Working, Notworking
Dropout	Nominal	Yes, No

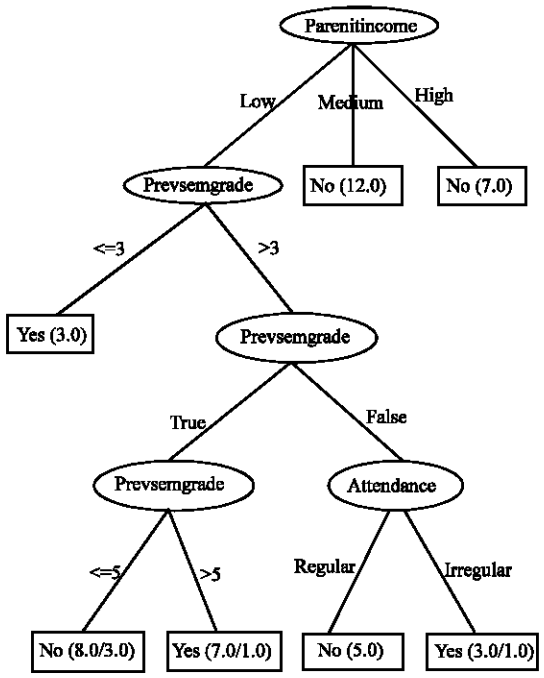

Fig. 1: Decision tree for drop out students

Modeling student drop outs: The modeling process starts by studying the relationship between student drop outs and underlying risk factors including gender, attendance, previous 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 drop outs. After the factors are identified, the logistic regression technique is used to quantify the drop outs and the effect of each risk factor.

The Table 1 shows the variables that influence the drop outs.

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. Though 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. Figure 2 shows a bar

chart showing logistic regression. The vertical axis represents the absolute value for the effect. In this example, the variable parent income has the positive value and all the other variables have negative values.

CLUSTERING

Data clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. A cluster indicates a number of similar objects, such that the members inside a cluster are as dissimilar as possible (heterogeneity)[11].

A simple distance measure i.e., Euclidean distance can be used to express dissimilarity between every two patterns. The clustering criterion can be specified to measure the distance between data observations and seeds. The observations are divided into clusters such that every observation belongs to at most one cluster.

Unlike data classification, data clustering does not require category labels or predefined group information. The aim is to segment the cases into disjoint classes that are homogenous with respect to the inputs. Clustering studies have no dependent variables. The database can be segmented by clustering methods such as partitioning methods and hierarchical methods. K-means clustering method is used in the model.

K-means clustering: The K-means algorithm is the simplest and most commonly used clustering algorithm employing a square error criterion[12,13]. It is computationally fast and iteratively partitions a data set into k disjoint clusters, where the value of k is an algorithmic input. K-means algorithm attempts to minimize the sum of squares clustering function given by

$$J = \sum_{j=1}^{k} \sum_{n \in sj} || xn - \mu j ||2$$

where μ_j is the mean of the data points in cluster S_j and is given by

$$\mu_j = \frac{1}{N} \sum_{n \in s_j} x^n$$

The training is carried out by assigning the points at random to k-clusters and then computing the mean vectors μ_j of the N_j points in each cluster. Each point is re-assigned to a new cluster according to which is the nearest mean vector. The mean vectors are then recomputed.

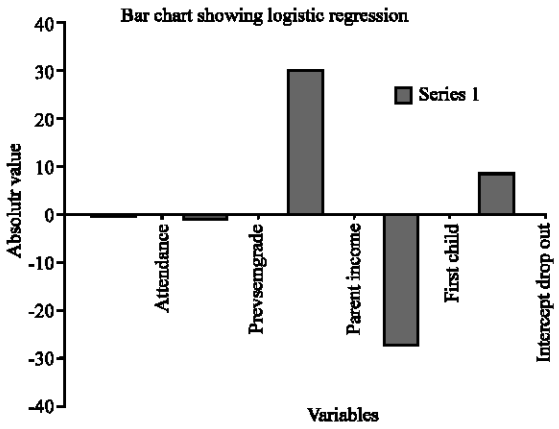

Fig. 2: Bar chart showing logistic regression

Clustering students: The Drop out frequency depends on demographic factors. Students are segmented into groups that are similar to each other with respect to these attributes. After the students are segmented, samples from each segment will be used to estimate the frequency. The results of this test estimate will allow the teachers to evaluate the potential profit of prospects from the list, both overall as well as for specific segments.

The model evaluates the drop out feature by considering the various factors such as gender, attendance, previous semester grade, parent education, parent income, scholarship, first child and whether the student is working or not. Table 1 shows the variables that affect the dropouts. K-means clustering is used to form the clusters. It segments the students into groups that are similar to each other with respect to these attributes.

Figure 2-5 shows cluster presentation for variables attendance, parent income, parent education and first child. Blue colour indicates the drop out variable with the value 'Yes' and red colour indicates the drop out variable with the value 'No'. Cluster 0 and cluster 2 have more number of dropout students than any other clusters. Figure 3 and 4 shows that the variables "parent income" and "parent education" are key inputs that help differentiate the students in cluster 0 from all of students in data set. Investigation of the data reveals that the parents of the students in this cluster are not educated and belong to the low-income category. Students in cluster 0 are the "first children" in the family.

Cluster 2 shows that parents are educated though they belong to low-income category. Cluster 2 has only girl students. Cluster 1, 3 and 4 are more or less similar and the drop out students are also very less. Parents from

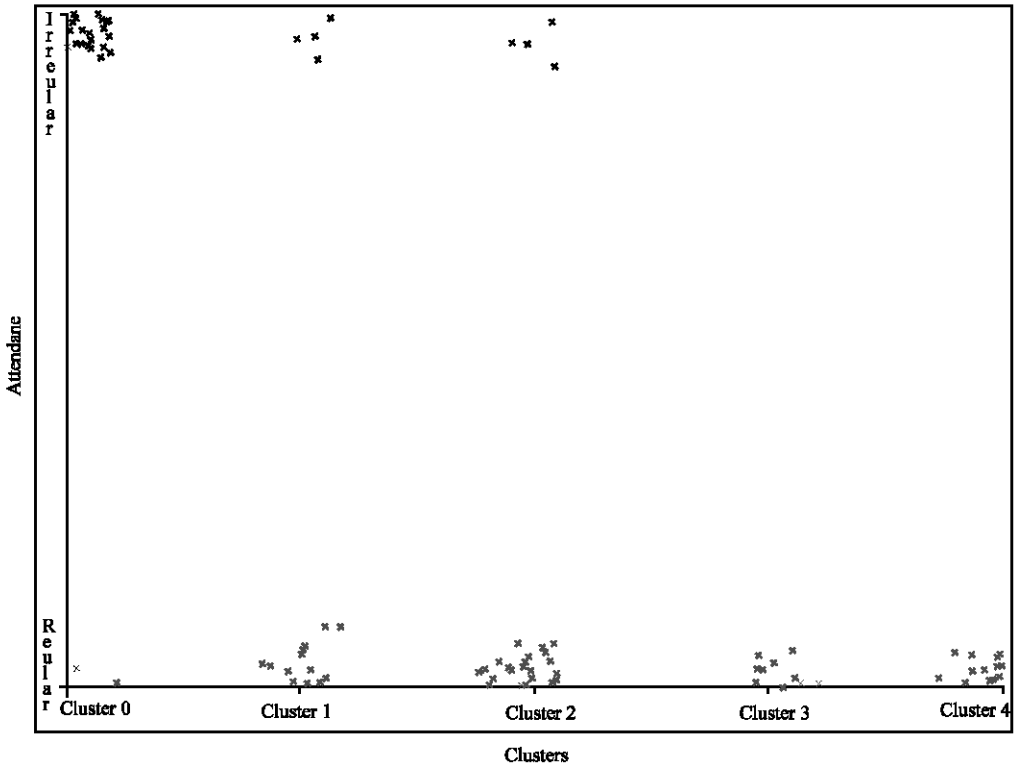

Fig. 3: Cluster repersentretion for attendance

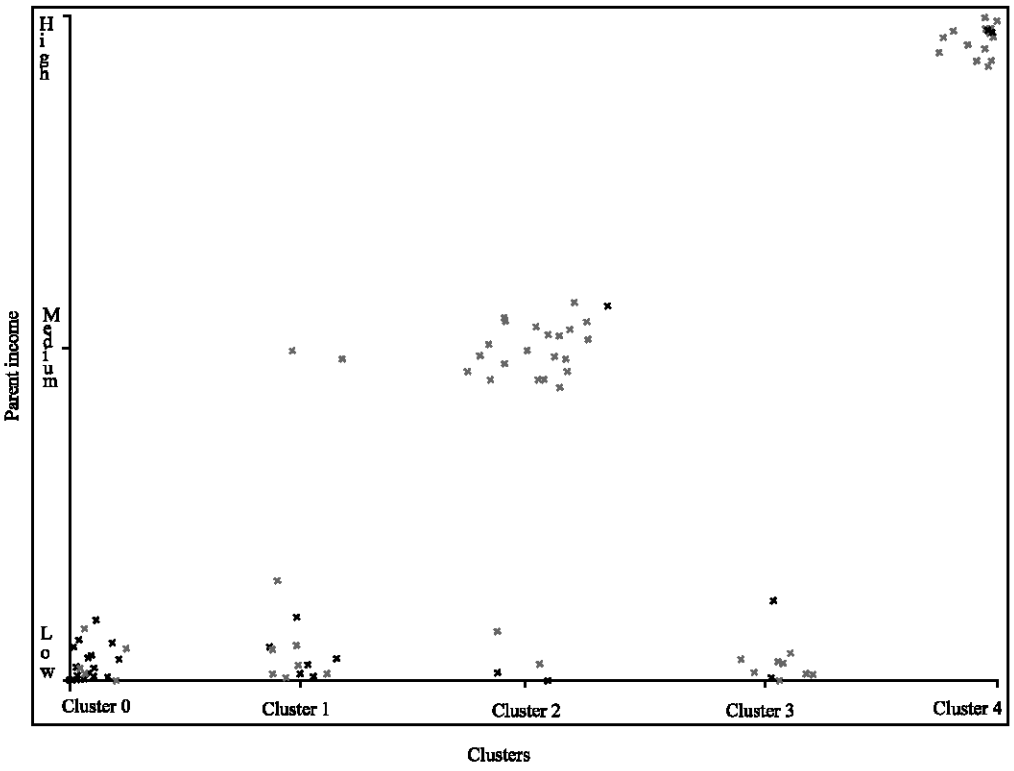

Fig. 4: Cluster repersentretion for parent income

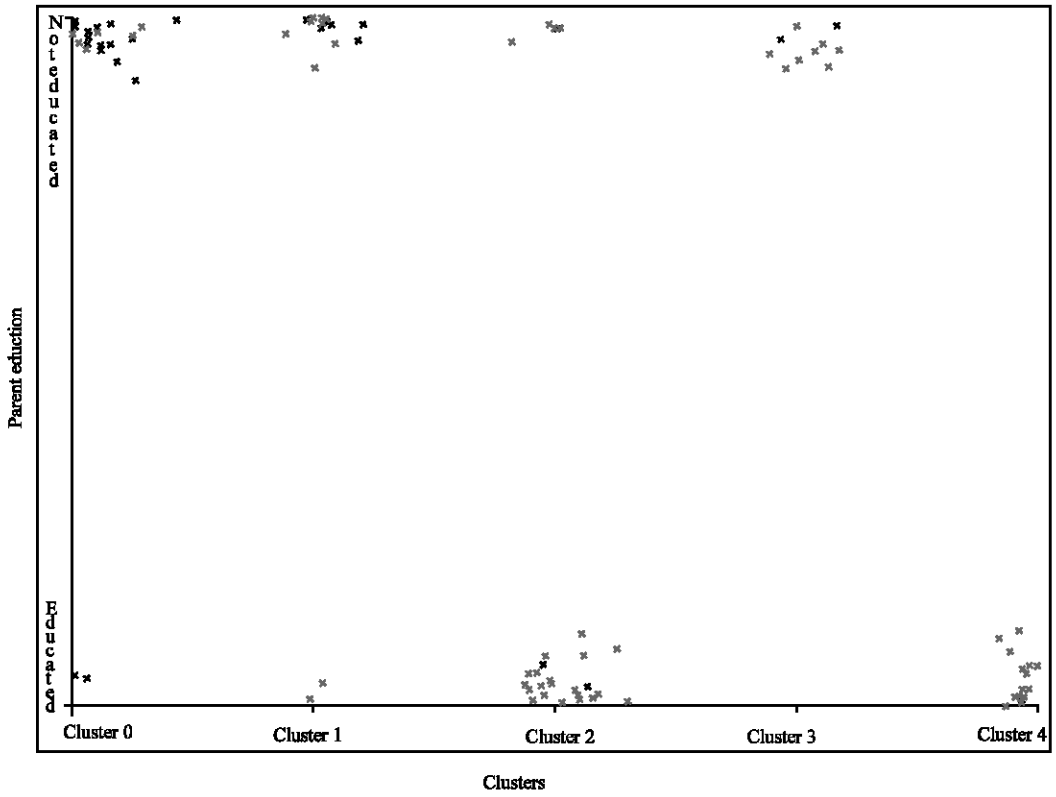

Fig. 5: Cluster repersentretion for parent education

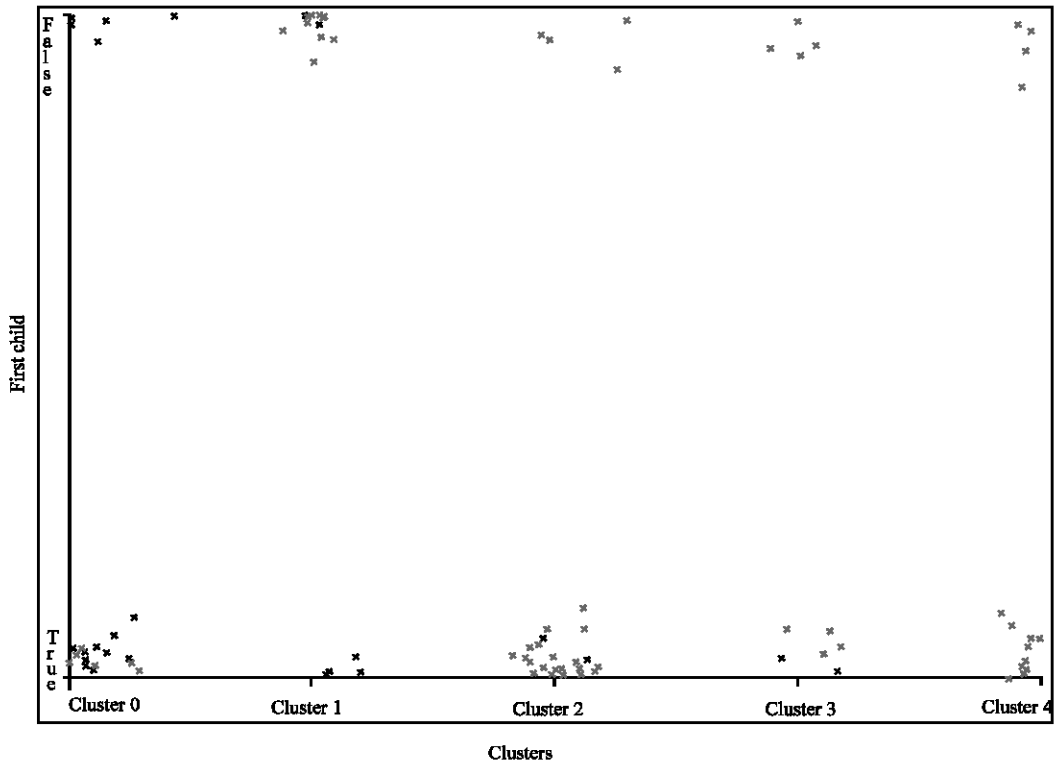

Fig. 6: Cluster repersentretion for first child

these clusters are also educated, belong to medium and high-income categories and the student's attendance is also regular. Cluster 0 seems to be an interesting cluster from a teacher's perspective and it requires immediate action. Cluster 2 also draws the attention of teachers.

Clustering analysis can be used by the educational industry to improve the predictive accuracy by segmenting databases into more homogeneous groups. Then the data of each group can be explored, analyzed and modeled. Segments based on types of variables that associate with risk factors or behaviors often provide sharp contrasts, which can be interpreted more easily. As a result teachers can more accurately predict the likelihood of the drop out feature and its frequency.

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.

REFERENCES

1. Frawley, W., G. Piatetsky-shapiro and C. Matheus, 1992. Knowledge discovery in databases: An overview, AI magazine, Fall, pp: 213-228.
2. Fayyad, U.M., G. Pitateskey-shapiro, P. Smyth and R. Uthurasamy, 1996. Advances in knowledge discovery and data mining, AAAI/MIT Press.
3. Shyamala, K. and S.P. Rajagopalan, 2006. Data mining model for a better higher educational system. Information Tech. J., 5: 560-564.
4. 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.
5. Murphy, K. and F. Welch, 1993. Inequality and relative wages. Ameri. Economic review, 83: 104-109.
6. Murtaugh, P.A., L.D. Burns and J. Schuster, 1999. Predicting the retention of university students. Higher Education, 4: 355-371.
7. 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.
8. Manski, C. and D. Wise, 1983. College choice in America, Cambridge, MA: Harvard University Press.
9. Quinlan, J.R, 1983. Induction of Machine learning. Machine learning, 1: 81-106.
10. Han, J. and M. Kamber, 2003. Data mining: Concepts and techniques. Morgan Kaufmann Publishers, New Delhi.
11. Hoppner, F., F. Klawn, R. Kruse and T. Runkler, 2000. Fuzzy Cluster analysis: "Methods for classification, data analysis and image recognition", John wiley and Sons, Inc., New York NY.
12. McQueen, J.B., 1967. Some methods of classification and analysis of multivariate observations. Proceedings of fifth Berkeley symposium on mathematical statistics and probability, pp: 281-297.
13. Adriaans, P. and D. Zantinge, 2000. Data mining, Addison Wesley.