http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (3): 560-564, 2006
ISSN 1812-5638

Data Mining Model for a Better Higher Educational System

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

Abstract: The main objective of any higher educational institution is to impart quality education. One way to reach the highest level of quality in higher education systems is by improving the decision making procedures on various processes such as assessment, evaluation, counseling and so on which requires knowledge. The knowledge is hidden among the educational data set and it is extractable through data mining technology. This paper is designed to present and justify the capabilities of data mining in the context of higher education by offering a data mining model for higher educational system in the colleges. It presents an approach to classifying students in order to predict their final grade based on certain features extracted from educational data bases. It helps earlier in identifying the dropouts and students who are below average and allow the teacher to provide appropriate counseling/advising in appropriate time.

Key words: Data mining, classification, association, higher education

INTRODUCTION

Data mining extracts previously unknown, valid, novel potentially useful and understandable patterns in educational data of large databases. The discovered hidden patterns enable the higher educational system in making better decisions and having more advanced plan in directing students (Zaiane, 2001). The education domain offers a fertile ground for many interesting and challenging data mining applications. These applications can help both educators and students to improve the quality of education.

Literature study in this area (Beck, 2004; Minaei *et al*., 2004; Silva and Vieira, 2001; Thomas and Loay, 2005) reveals many web based educational systems with different capabilities and approaches have been developed. But the requirement of the model which cater to the needs of the Indian educational system was felt and was developed. The data mining model monitors each student's progress by capturing the variables such as previous semester grade, test mark, assignment grade and attendance. The students performance is also analyzed based on the features of interpersonal peer groups such as degree aspirations, intellectual self confidence, scoring pattern and time spent with peer groups. The system also identifies the students who are likely to drop out.

The data mining model presented in this study aims to answer the following questions.

- How data mining concepts can be used to improve the quality of the decision making process in higher education?
- How data mining techniques can be used to assess student's performance?
- How can data mining be used to classify the result of students? Can we associate the grades of previous semester with the test marks of existing semester to predict the result?
- How can data mining be used to identify the students who are likely to drop out? Can we help the students in providing appropriate counseling in a timely manner?
- How can data mining be used to find the influence of friendship groups on education of students?
- Can we help teachers to identify the students at risk of failure and provide additional support services such as extra coaching, academic counseling and financial aid helping based on early prediction regarding semester grades?

The goal of this model is to find similar patterns from the data gathered and to make predictions about student's performance. The system also points out the

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

influence of friendship groups on performance of students and the number of students likely to drop out. Based on the performance of the students, teachers can make suggestions to improve and design/modify the curriculum effectively.

Clementine, data mining software by SPSS Inc. was chosen for this data mining model. The algorithm C5.0 was used for this model. Using the decision tree classification technique, students are classified into groups of successful and unsuccessful students. By associating the previous semester grades with the present semester marks, the result is predicted.

DATA MINING IN HIGHER EDUCATIONAL SYSTEM

Education is viewed as a critical factor in contributing to the long term economic well being of the country. Today the important challenge that higher education faces is reaching a stage to facilitate the universities/colleges in having more efficient, effective and accurate educational processes. Data mining is considered as the most suited technology appropriate in giving additional insight into the student, lecturer, alumini and other educational staff behavior and acting as an active automated assistant in helping them for making better decisions on their educational activities.

Data mining is the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information (Adriaans and Zantinge, 2000). Rather than starting from a question or a hypothesis, data mining simply finds patterns that are already present in the data.

Data mining is the process of discovering 'hidden images', patterns and knowledge within large amount of data and of making predictions for outcomes or behaviors (Han and Kamber, 2003). Lack of deep and enough knowledge in higher educational system may prevent system management to achieve quality objectives. Data mining methodology can help bridging this knowledge gaps in higher educational system. Therefore the hidden patterns, association and anomalies, which are discovered by some data mining techniques can be used to improve the effectiveness, efficiency and speed of the processes.

This improvement may bring a lot of advantages to higher educational system such as maximizing educational system efficiency, decreasing student's drop out rate, increasing student's promotion rate, increasing student's success, increasing student's learning outcome, increasing education improvement ratio and reducing the cost of system processes. In order to achieve the above quality improvement, we need a data mining system that can provide the needed knowledge and insights for the decision makers in the higher educational system.

PROPOSED MODEL

In present day's educational system, a student's performance in any college is determined by the combination of internal assessment and external mark. The internal assessment is carried out by the teachers based upon the student's performance in various evaluation methods such as tests, assignments, seminars, attendance and extension activities (NCC/NSS/Sports). The external mark is the one that is scored by the student in semester examination. Each student has to get minimum pass mark in internal and as well as in external examination.

The current educational system does not involve any prediction about fail or pass percentage based on the performance. The system doesn't deal with dropouts. There is no efficient method to caution the students about the deficiency in attendance. It does not identify the weak students and inform the teachers. Since the proposed model identifies the weak students, the teachers can provide academic help for them. It also helps the teachers to act before a student drops or to plan for resource allocation with confidence gained from knowing how many students are likely to pass or fail.

The test data was selected from Dr. Ambedkar Government College, Chennai from Zoology course. About 180 students were initially enrolled in the course. However some of the students dropped the course after a couple of months. After removing those students, there remained 163 students. Though there is various software available, Clementine, data mining software by SPSS Inc. was chosen for this data mining model due to its advantages. The algorithm C5.0 was used for this model.

The test mark plays a dominant role in internal assessment. The test mark is assigned based on the periodical tests conducted internally by the institution. The model also associates grades of previous semester with present test marks to predict the semester results. Some of the rules produced by C5.0 are given:

If pre_sem_grade≤2.5 and test_mark≤3.0 then
result = 'Fail' (conf = 0.87%)
If pre_sem_grade≤2.5 and test_mark>3.0 then
result = 'Pass' (conf = 0.59%)

Usage of these rules may identify weak students and hence the teachers may concentrate on these students and take necessary steps to improve their performance in semester examination. For I semester students, marks obtained in higher secondary examination may be

Table 1: Assignment grades and corresponding marks

Submission	Correct Responses	No. of tries	Assignment grade	Assignment mark
Yes	Yes	1	A	4
Yes	Yes	2	B	3
Yes	Yes	3	C	2
Yes	No	-	D	1
No	-	-	-	0

Table 2: Distribution of students according to various grades

Grade	No. of students	Percentage
1.0	2	1.20
2.0	5	3.10
3.0	9	5.50
4.0	12	7.40
5.0	15	9.20
6.0	31	19.20
7.0	36	22.10
8.0	22	13.50
9.0	20	12.30
10.0	11	6.80

considered for previous semester grades. Due to early intervention, prediction regarding result is done during the middle of the semester itself. So students at risk may be alerted by the teachers and take up additional support services such as extra coaching, academic counseling and financial aid helping.

Secondly the model deals with assignment grades. The assignment grades are allotted based on various features such as:

- Correct responses
- Incorrect responses
- Submission of assignments
- Non-submission of assignments
- Total number of tries

Table 1 shows the assignment grades and the corresponding marks. The assignment rules may take the following forms.

If submission = 'Yes' and correct_responses = 'Yes' and total_no_tries = 3 then assgmt_grade = 'C'.

Assignment grades may also be used to predict semester results. For example, the feature correct_responses and total_no_tries may be used to predict the semester result. The rule is given:

If correct_responses = 'Yes' and total_no_tries = 1 then result = 'Pass' (conf = 70%)

Students with good assignment grades tend to score good results in the examination. This can be seen from Fig. 1. So students with high assignment grades may

Table 3: Distribution of students according to various class labels

Result	Grade	No. of students	Percentage
Pass	Grade>4.0	135	82.80
Fail	Grade≤ 4.0	28	17.20

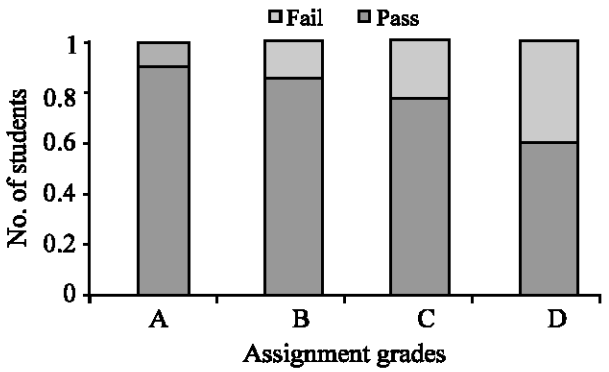

Fig. 1: Assignment grade Vs result

interpret this as they are doing well in their exams. It also shows that the students with lower grades have to work hard to obtain good results.

Attendance mark also plays an important role as internal mark. The attendance may be classified as regular or irregular based on the attendance percentage. The attendance may be relaxed up to certain percentage with a penalty fees. The attendance may be associated with test mark and assignment grade to predict semester result. The student may fail in the exam if his/her previous semester grade is low and assignment grade is also low and irregular in attendance. Grade may be allotted to seminar based on the presentation and subject content of the students. The extension activities (NCC/NSS/Sports) also play an important role and grade is allotted for them. Students are expected to take up any one of the extension activities.

We can group the students regarding their final grades using several ways. The proposed model involves grouping the students according to various grades. Table 2 shows the distribution of students according to various grades.

The students may also be distributed according to various class labels such as pass and fail as show in Table 3.

Graduation is an increasingly important policy issue in any educational system. Failure of a student is considered as the failure of an institution or the entire educational system (Murtaugh *et al*., 1999). As graduation rates are one of the institutional effectiveness measures, it is necessary for every college to totally reduce the dropout rate. Therefore one of the tasks important to any institution and its teachers is to identify

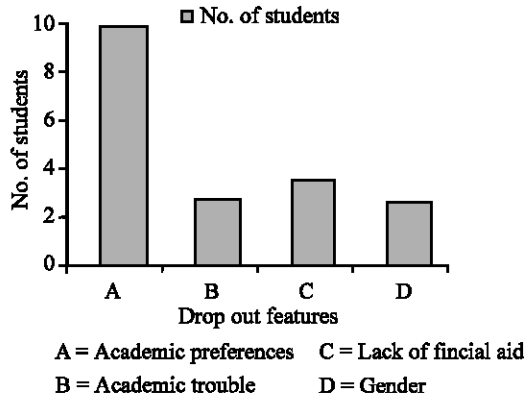

Fig. 2: Distribution of drop out students

the students who are likely to drop out. The students may leave the institution due to various reasons. The model identifies the dropouts based on various features (Desjardins *et al*., 2002). Some of the features are listed below.

- Academic trouble
- Academic preferences
- Lack of financial aid
- Previous semester grade
- Parental education
- Gender
- Attendance

The data set shows 17 drop out students among 180 students. Most of the students left the institution due to academic preferences. Other students left the course due to other reasons. Figure 2 shows the distribution of drop out students. Teachers can observe the various features involved and can take preventive measures to avoid the problem.

The model also deals with influence of friendship groups on education of students. A college student's peer acts as a reference group or an environmental source of socio cultural norms in the midst of which a student grows and develops (Antonio, 2004; Webb, 2003). To understand the role played by interpersonal peer groups in student learning and development, the model studies the various features, which are given below.

- Degree aspirations
- Intellectual self confidence
- Scoring pattern
- Time spent

For example, the degree aspirations may be considered based on the intention of the student to obtain a highest academic degree (Medical/Engineering/ Law/Doctorate/ Bachelor degree/Master degree etc). The intellectual self confidence and degree aspirations can be combined to predict the result. The rule may take the following form.

If int_self_con = A and deg_aspn = A then
Result = 'Pass'
(A = Above average, B = Below average)

After careful examination of the features tested above, the model gives out the following findings.

- Since the model can predict the success rate/graduation rate even during the middle of the semester, teachers can take steps to improve the performance of the students for that semester itself.
- As the model predicts the number of students likely to drop out, teachers can concentrate on appropriate features associated and counsel the students or arrange for financial aid to them.
- The model shows that the influence of friendship groups on student's education is important in predicting student's performance. Hence teachers can advise the students accordingly for their improvement.

Based on the prediction, the colleges may take immediate remedial action to improve the situation and alter the policies that are set up. There is no doubt that the proposed model helps the colleges to better understand the need and requirements of the students to get higher grades and successfully complete the course.

CONCLUSIONS

Among the recent technology innovations, data mining is making sweeping changes in the field of higher education. It has tremendous applications in higher education institutional research alone. In this study, we have shown that data mining can be useful in predicting student outcomes. As degree attainment is necessary for every student, the presented model is aimed at helping move the students along the pipeline with fewer dropouts as possible. This unquestionably translates into increased efficiency, higher graduation rates and lower cost to society as a whole.

The proposed model helps us to predict which students are less likely to perform well in that specific course or those who are less likely to be successful in it.

By identifying such students the college may take necessary action and may provide extra coaching,

academic counseling and financial aid helping. Such activities will definitely lead to improved decision making procedures and will improve the quality of the instructions. We are further developing the model to obtain more complex and interesting rules.

REFERENCES

Adriaans, P. and D. Zantinge, 2000. Data Mining. Addison Wesley, 2000.

Antonio, A.L., 2004. The influence of friendship groups on intellectual self-confidence and educational aspirations in college. J. Higher Education, 75: 446-470.

Beck, J., 2004. Analyzing student-tutor interaction logs to improve educational outcomes. Proc. ITS 2004 Workshop, Brazil, pp: 60-67.

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.

Han, J. and M. Kamber, 2003. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, New Delhi.

Minaei, B., G. Kortemeyer and W.F. Punch, 2004. Association analysis for an online education system. IEEE International Conference on Information Reuse and Integration (IRI-2004), Las Vegas, Nevada, USA.

Murtaugh, P.A., L.D. Burns and J. Schuster, 1999. Predicting the retention of university students. Res. Higher Education, 4: 355-371.

Silva, D.R. and M.T.P. Vieira, 2001. An ongoing assessment model in distance learning. In Proceedings of Internet and Multimedia Systems and Applications, Honolulu, USA.

Thomas, C.G. and M.L. Elbasyouni, 2005. Student online assessment behaviors. IEEE Transactions on Education, 48: 400-401.

Webb, N.M., 2003. Peer interaction and learning in small groups. Intl. J. Educational Res., 13: 21-29.

Zaiane, O.R., 2001. Webusage mining for a better web_based learning environment. In Proceedings of Conference on Advanced Technology for Education (CATE'01, Alberta, pp: 60-64.