Predicting Examination Results using Association Rule Mining

Omprakash Chandrakar
Associate Professor, Dept. of Computer Science
Uka Tarsadia University
Bardoli, India

Jatinderkumar R. Saini, Ph.D.
Director I/C & Associate Professor
Narmada College of Computer Application
Bharuch, Gujarat, India

ABSTRACT

Higher education has changed a lot in the last decade. The use of various innovative techniques and technologies, especially ICT in teaching learning process is increasing day by day. Number of information systems has been developed and successfully implemented to support educational processes. These systems typically capture almost every data regarding a student, right from their enrollment into a course to graduation and placement. If these data are analyzed and visualize properly, can provide valuable knowledge that can be used to enhance their learning skill and to predict threats if any, well in advance so that appropriate measure can be taken to avoid it. Knowledge Discovery in Database (KDD) is usually referred as Data mining. It is a process of extracting new and potential useful information from large databases. Data mining tools are used to identify any pattern or predict future trends and behaviors. This enables decision maker to make proactive and knowledge-driven decisions. This paper presents an application of data mining in higher education. Association rule mining is applied to analyze the performance of students in their examinations and predicts the outcome of the forthcoming examination. This prediction allows student and teacher to identify the subjects which need more attention even before the commencement of semester.

Keywords

Data mining, association rule, rule generation, rules pruning, academic performance.

1. INTRODUCTION

Higher education has changed a lot in the last decade. The infusion and integration of the new information technologies in the teaching and learning have had an immense impact on the educational environment (Chomal, Saini, 2013). Number of information systems has been developed and successfully implemented to support educational processes. These systems typically capture almost every data regarding a student, right from their enrollment into a course to graduation and placement. Also, students' individual data (like name, age and gender), demographic data (like socioeconomic status), academic background (including opted subject(s) and result in school), current academic performance (like daily attendance and marks obtained in each subject) are also captured. Various socio-demographic and study environment variables which influence the performance of the student have been examined (Nidhi Arora, JatinderKumar R. Saini). The information system uses these data to prepare reports in the predefined format only.

If these data are analyzed and visualize properly, can provide valuable knowledge that can be used to enhance students learning skill and to predict threats if any, well in advance so that appropriate measure can be taken to avoid it.

The objective of this study is to determine whether any significant association exists between student's performances in different subjects. If such relationship exists, then this knowledge can be used to guide the students to improve their overall performance.

The rest of the paper is organized as follows. Section II reviews some of the related work. Section III describes the data set used in this study and presents the preprocessing performed on the data set. In section IV, we briefly describe association rule mining and its algorithm. Section V presents the run of association rule mining with different parameter values and criterion. Rules generated by the association rule mining are listed. Finally in section VI we conclude the paper.

2. DATA MINING IN HIGHER EDUCATION & RELATED WORK

Higher education institutions have long been interested in predicting the paths of students and alumni (Luan, 2004), thus identifying which students will join particular course programs (Kalathur, 2006), and which students will require assistance in order to graduate. Another important preoccupation is the academic failure among students which has long fuelled a large number of debates. Researchers (Vandamme et al., 2007) attempted to classify students into different clusters with dissimilar risks in exam failure, but also to detect with realistic accuracy what and how much the students know, in order to deduce specific learning gaps (Piementel& Omar, 2005).

The distance and on-line education, together with the intelligent tutoring systems and their capability to register its exchanges with students (Mostow et al., 2005) present various feasible information sources for the data mining processes. Studies based on collecting and interpreting the information from several courses could possibly assist teachers and students in the web-based learning setting (Myller et al., 2002). Scientists (Anjewierden et al., 2007) derived models for classifying chat messages using data mining techniques, in order to offer learners real-time adaptive feedback which could result in the improvement of learning environments. Researchers (Nidhi Arora, JatinderKumar R. Saini, 2013) proposed a model that allows prediction of students' academic performance based on some of their qualitative observations using Neural Network.

3. DATA PREPROCESSING

Various data preprocessing are usually applied on raw data to make it appropriate for data mining. It converts the data into a format that will effectively processed by the data mining algorithms.

There are a number of data preprocessing techniques. What techniques is to be used and how to perform preprocessing on the data is depends on the particular data mining function we want to apply and more importantly the data mining algorithm we want to implement.

This section describes data set and preprocessing we performed on the given data set.

3.1 Data Set

The data set used in this study was collected from an institute offering BCA course from Gujarat. We collected the marks obtained in each of 6 subjects in semester 1 and 2. The data are available in Excel. We have performed following operations on the data.

3.2 Data Cleaning

Data cleaning involves removing noise and inconsistencies present in the data. In the given data set we found that some of the students not appeared in examination. So to make the data set consistent, we removed all such records from the data set.

3.3 Data integration

Data integration combines data from two or more tables/sources into a coherent data set. In our data set there are two records for each student, one consisting of result in semester 1 and other in semester 2. For each semester, student is assigned an Exam Seat Number. So in the given data set, there are two different Exam Seat Numbers. To integrate these two records, we replaced Exam Seat Number with corresponding Student Identification Number. Each student is assigned a unique Student Identification Number at the time of enrollment in a course and it remains same during the entire course.

Now in our data set there is only one record of each student, consisting of Student Identification Number along with marks obtained in all 12 subjects from semester 1 and 2 both.

3.4 Data transformations

Various data transformation techniques are applied on the data set before mining with an objective to increase the accuracy and efficiency of mining algorithms.

In our data set, it has been observed that marks obtained in each subjects varies from 0 to 70 for theory papers and 0 to 140 for practical papers. The range is very large so it in needs to be normalized.

In university education system, marks are divided into three grades, FAIL (<40), PASS (<70) and DIST (>70). For normalizing the data set, we replaced the marks with corresponding grades applying the above rule.

3.5 Data reduction

Data reduction techniques like aggregation, elimination of redundant data or clustering are used on the data set to bring down the data size.

In our data set, no redundancy is found. Further we have considered only one batch containing 120 records. So no data reduction is required.

Now our final data set contains 119 records, each records contains 13 attributes.

4. ASSOCIATION RULE MINING

Association rule mining is the process of finding interesting pattern or relationships exist in the large data set. Because of the availability of large data, mining association rules from the databases gaining importance day by day. The significant relationship discovered through Association rule mining can be used in the marketing, sales, arranging items in the rack in a supers store, designing catalogs.

Let I = {M11, M12, M13, …, M26} be a set of distinct items (Marks scored by the student in each subject), and D = {T1, T2, …, Tn} be a transactional database. Each transaction T contains values for I. Table 1 shows an example data set.

Table 1. A Sample Data Set

STUD ID	M11	M12	……	M25	M26
1	P	P	…	P	D
2	P	F	....	P	D
4	P	P	…	D	D
5	P	P	…	P	D
6	P	D	..	P	D
7	P	F	…	P	D
8	P	D	…	P	D
9	F	P	…	P	D
10	P	P	…	P	D
.			..		
.			.		
.			.		
119	P	P	…	P	D
120	P	P	…	P	F

Let "X → Y" is an association rule with $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. Appropriateness or strength of an association rule can be measured using following three measures: support, confidence, and lift.

- Mathematically, support is the probability that X and Y both occur in a transaction. Support = $P(X \cap B)$.
- Confidence is the probability that Y occurs in a transaction that X has already occurred.

  Confidence = $P(Y|X)$ or $P(X \cap Y)/P(X)$.
- Lift normalizes the confidence with the probability of B, i.e. $P(X \cap Y)/(P(X) \cap P(Y))$. This is a measure of the importance of the association that is independent of support. The lift is equal to 1 then X and Y are independent of each another.

We used Apriori algorithm for association mining. It does two tasks:

1. Generate item sets that pass a minimum support threshold.
2. Generate rules that pass a minimum confidence threshold.

5. EXPERIMENTS AND RESULTS

Our data set contains STUDENT_ID, which is a row identifier. Before performing association rule mining we removed this column from our data set. We are interested in finding the relationship in following format "STUDENT IS FAILING IN A PARTICULAR SUBJECT IN SEMESTER 1 IS LIKELY TO FAIL IN A PARTICULAR SUBJECT IN SEMESTER 2". So in the given data set we are interested in the result in which students are failing. To find such an association, we replaced the result value for P and D with ?, which represents missing value. Now our final data set looks as shown in the following table.

Table 2. Final Data Set

STUDENT ID	M11	M12	…	M25	M26
1	?	?	…	?	?
2	?	F	..	?	?
4	?	?	..	?	?
5	?	?	..	?	?
6	?	?	..	?	?
7	?	F	..	?	?
8	?	?	..	?	?
9	F	?	..	?	?
10	?	?	..	?	?
.					
.					
.					
119	?	?	?	?	?
120	?	?	?	?	F

Rest of this section describes the experiments we performed on the data set and results.

Run 1.
We run the association rule mining against RESULT09BATCH data set with following parameters.

Table 3. Parameters

Parameter	**Value**	**Description**
Minimum confidence value	.5	Criteria for the rule generation
Delta	.05	Iteratively decrease support by this factor until min support is reached or required number of rules has been generated.
NumRules	20	Number of rules to discover.
LowerBoundMinSupport	0.1	Lower bound for minimum support.
UpperBoundMinSupport	1.0	Upper bound for minimum support. Start iteratively decreasing minimum support from this value.

Total 17 rules are generated. We are interested in such rules in which result in semester 1 are premises and result in semester 2 are consequences. Out of these 17 rules 5 rules are in the desired format.

4. M13=F M14=F 19 ==> M24=F 14 conf:(0.74)

8. M12=F 18 ==> M24=F 12 conf:(0.67)

9. M12=F 18 ==> M25=F 12 conf:(0.67)

11. M14=F 33 ==> M24=F 21 conf:(0.64)

13. M13=F 39 ==> M24=F 23 conf:(0.59)

By replacing the subject code with subject, the above rules can be rewritten as

4 If a student is failing in Introduction to computer and Computer programming & programming methodology in semester 1, he is likely to fail in Programming Language C (conf: 0.74).

8 If a student is failing in Mathematic-1 in semester 1, he is likely to fail in Programming Language C (conf: 0.67).

9 If a student is failing in Mathematic-1 in semester 1, he is likely to fail in Database Management System (conf: 0.67).

11 If a student is failing in Computer programming & programming methodology in semester 1, he is likely to fail in Programming Language C (conf: 0.64).

13 If a student is failing in Introduction to computer in semester 1, he is likely to fail in Programming Language C (conf: 0.59).

6. CONCLUSION

This paper presented an application of data mining in higher education. Association rule mining is applied to analyze the performance of students in their examinations. The rules discovered through association rule mining are used to predict the outcome of the forthcoming examination. This prediction may be used to guide the student at the very beginning of their semester by identifying the subjects in which they need to focus more.

7. REFERENCES

[1] Vikas Sitaram Chomal, Dr. Jatinderkumar R. Saini, "A study and analysis of paradigm shifts in education triggered by technology", IJRESS Volume 3, Issue 1 (January 2013) ISSN: 2249-7382

[2] http://www.dkms.com/papers/dmkdd.pdf.

[3] I. H. Witten, E. Frank: Data Mining, Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann Publishers, San Francisco, 2005.

[4] Nidhi Arora, JatinderKumar R. Saini, "A Fuzzy Probabilistic Neural Network for Student's Academic Performance Prediction", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 9, September 2013 ISSN: 2319-875.

[5] Y. Ma, B. Liu, C. K. Wong, P.S. Yu, S. M. Lee, "Targeting the Right Students Using Data Mining", Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining, Boston 2000.

[6] Nidhi Arora, Jatinderkumar R. Saini, "Predicting Student Academic Performance using Fuzzy ARTMAP Network", International Journal of Advances in Engineering Science and Technology, V3N3, Page no.187-192, ISSN: 2319-1120

[7] Hideko Kitahama, "Data Mining through Cluster Analysis Evaluation on Internationalization of Universities in Japan".

[8] Jing Luan, PhD Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories "Data Mining Applications in Higher Education".

[9] Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 406-415.

[10] Thulasi Kumarthulasi.kumar@uni.edu, University of Northern Iowa " Theoretical Basis for Data Mining Approach to Higher Education Research".

[11] Amershi, S., Conati, C. (2006) Automatic Recognition of Learner Groups in Exploratory Learning Environments. Proceedings of ITS 2006, 8th International Conference on Intelligent Tutoring Systems.

[12] Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008) Developing a Generalizable Detector of When Students Game the System. User Modeling and User-Adapted Interaction,18, 3, 287-314.