Mining Association Rules in Student's Assessment Data

Dr. Varun Kumar[1], Anupama Chadha[2]

**[1] Department of Computer Science and Engineering, MVN University
Palwal, Haryana, India**

**[2] Anupama Chadha, Department of Computer Science and Engineering, ITM University
Gurgaon, India**

Abstract

Higher education, throughout the world is delivered through universities, colleges affiliated to various universities and some other recognized academic institutes. Today one of the biggest challenges, the educational institutions face, is the explosive growth of educational data and to use this data to improve the quality of managerial decisions to deliver quality education. In this paper we will perform a case study of a university that hopes to improve the quality of education by analyzing the data and discover the factors that affect the academic results so as to increase success chances of students. In this perspective we use association rules discovery techniques. Also we will show the importance of data preprocessing in data analysis which has a significant impact on the accuracy of the predicted results.

Keywords: *Higher education, Data mining, Knowledge discovery, Data preprocessing, Association rules*

1. Introduction

Education is an essential element for the betterment and progress of a country. It enables the people of a country civilized and well mannered[6]. Today the important challenge that higher education faces, is reaching a stage to

facilitate the universities in having more efficient, effective and accurate educational processes.

To date, higher educational organizations are placed in a very high competitive environment and are aiming to get more competitive advantages over the other competitors. To remain competitiveness among educational field, these organizations need deep and enough knowledge for a better assessment, evaluation, planning, and decision-making.[5] The required knowledge cannot be gained from the tailor made software used now a days. Data mining incorporates a multitude of techniques from a variety of fields including databases, statistics, data visualization, machine learning and others.[7] The data mining technology can discover the hidden patterns, associations, and anomalies from educational data. This knowledge can improve the decision making processes in higher educational systems.

Data mining is considered as the most suited technology appropriate in giving additional insight into the lecturer, student, alumni, manager, and other educational staff behavior and acting as an active automated assistant in helping them for making better decisions on their educational activities.[1] The data mining techniques can help the institutes in extracting patterns like students having similar characteristics, Association of students' attitude with performance, what factors will attract meritorious students and so on. The past several decades have witnessed a rapid growth in the use of data and knowledge mining as a means by which academic institutions extract useful hidden information in the student result repositories in order to improve students' learning processes.[2]

The main objective of this paper is to use data mining methodologies to study students' performance in their courses. In this research, we will be using Association rules discovery techniques to compare the student's performance in the subjects common at Graduation and Post Graduation level and will predict the factors which can explain their success or failure.

This paper is organized as follows:

Section 2: discusses about the motivations of this work and some related works.

Section 3: gives the relevant information about knowledge discovery process along with the data mining and association rule for the discovery of hidden knowledge.

Section 4: discusses the results of the analysis and the rules discovered from the present study.

Section 5: the conclusion discussed in this section.

2. Related Work

There have been done some studies in. the area of data mining in education. Each of them is trying to enhance the educational system by discovering patterns among the great deal of data. In this section, we will analyze the existing works.

C. Romero and S. Ventura (2010) survey the relevant studies carried out in the field of education. They have described the types of users, types of educational environments and the data they provide. Also they have explained in their work the common tasks in the educational environment that have been resolved through data mining techniques.

Hua-long Zhao (2008) has done Multidimensional cube analysis by taking use of OLAP technology and has shown that the curriculum chosen by the students can depend upon many angles like teacher, semester and student. He has used Star model of data warehouse to the analysis of curriculum which can provide certain policy making support for different levels of education policy- maker in the school.

Fadzilah Siraj and Mansour Ali Abdoulha (2009) have used data mining techniques for understanding student enrolment data. They have done comparative study of three predictive data mining techniques namely Neural Network, Logistic regression and Decision tree. The results obtained can be used by the planners to formulate proper plan for the university.

Shaeela Ayesha et al. (2010) discusses data mining technique named k-means clustering is applied to analyze student's learning behavior. Here K-means clustering method is used to discover knowledge that come from educational environment.

W.M.R. Tissera et al. (2006) presents a real-world experiment conducted in an ICT educational institute in Sri Lanka. A series of data mining tasks are applied to find relationships between subjects in the udergraduate syllabi. This knowledge provides many insights into the syllabi of different educational programmes and results in knowledge critical in decision making that directly affects he quality of the educational programmes.

Hongjie Sun (2010) conducts a research on student learning result based on data mining. It is aimed at putting forward a rule-discovery approach suitable for the student learning result evaluation and applying it into practice so as to improve learning evaluation skills and finally better serve learning practicing.

Qasem A. Al-Radaideh et al.(2006) worked on student data to study the main attributes that may affect the student performance in courses. Also S. Anupama Kumar and Dr. Vijayalakshmi M.N(2011) applied decision tree algorithm on student's internal assessment data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to fail or pass.

3. Knowledge Discovery Process

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The sequences of steps identified in extracting knowledge from data are:[8]

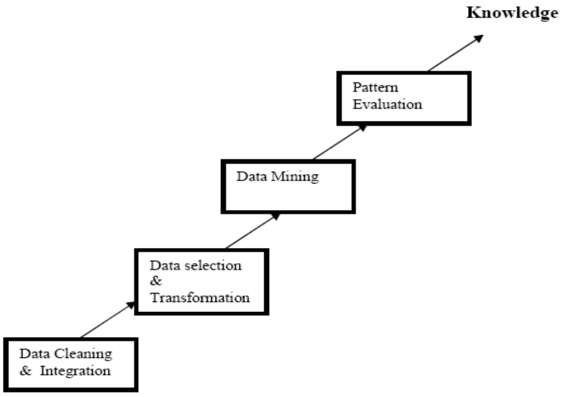


Fig. 1 The steps of extracting knowledge from data

3.1 Selecting Mining Frequent Patterns and Associations

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes' value conditions that occur frequently together in a given dataset.[8]

The preliminaries necessary to understand for performing data mining on any data are discussed below.
Let {I1,I2,…..Im} be set of items. Let *D*, the task relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called transaction identification (TID). Let *A* be a set of items. A transaction *T* is said to contain *A* if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \varnothing$

Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule $A \Rightarrow B$ means that A and B exist together in 2% of all the transactions under analysis. The rule $A \Rightarrow B$ having confidence of 60% in the transaction set D means that 60% is the percentage of transactions in D containing A that also contains B.

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. If the relative support of an itemset I satisfies a prescribed minimum support threshold, then I is a frequent itemset. The association rule mining can be viewed as a two-step process:

***1)* Find all frequent itemsets: Each of these itemsets will occur at least as frequently as a predetermined minimum support count.**

***2)* Generate strong association rules from the frequent itemsets: The rules must satisfy minimum support and confidence. These rules are called strong rules**. [8]

3.2 Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agarwal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. The following lines state the steps in generating frequent itemset in Apriori algorithm.[3]

Let *Ck* be a candidate itemset of size *k* and *Lk* as a frequent itemset of size *k*. The main steps of iteration are:

- Find frequent set *Lk-1*
- Join step: *Ck* is generated by joining *Lk -1* with itself (cartesian product *Lk-1 x Lk-1*)
- Prune step (apriori property): Any (*k* − *1*) size itemset that is not frequent cannot be a subset of a frequent *k* size itemset, hence should be removed
- Frequent set *Lk* has been achieved [3]

3.3 Data Collection

The data required for our study was taken from MCA course of a renowned university in Haryana. The data which we utilized in our study consists of the results of the subjects which are common in their Graduation and Post graduation degree programs. The data also contains the stream from which the students have done their graduation.

3.4 Need of Data Preprocessing

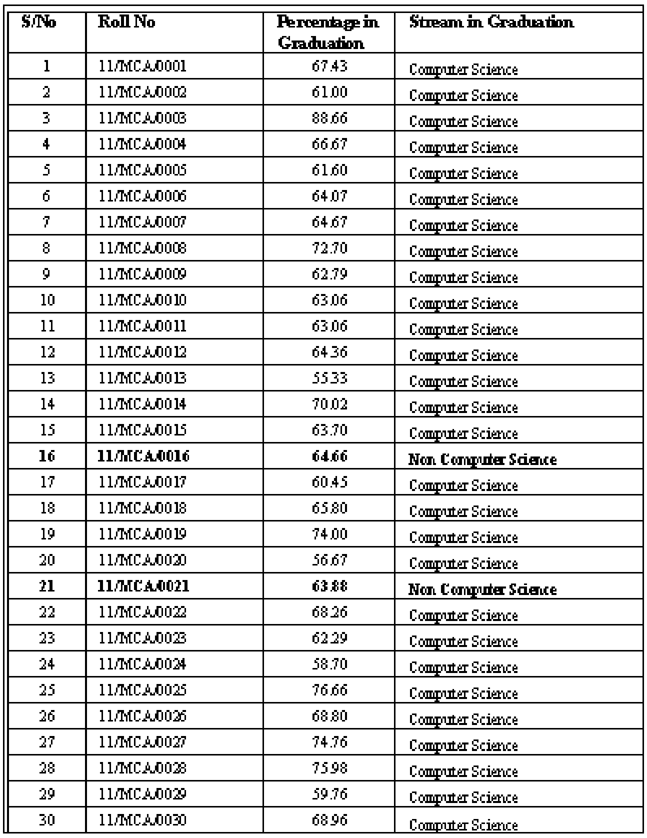
Data preprocessing is an important step in the knowledge discovery process, because quality decisions must be based on quality of data. Real-world data tend to be noisy (containing errors or outlier values that deviate from the expected), incomplete (lacking attribute values or certain attributes of interest) and inconsistent (containing discrepancies). Data preprocessing involves Data cleaning and Data transformation. Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. Data transformation operations like Normalization and Aggregation are additional data preprocessing procedures that would contribute toward the success of the mining process.[8]

3.5 Data Cleaning

The data collected needs to undergo cleaning and transformation, the need for which is already discussed above. The following measure was taken to cleanse our data:

- Remove all the records of the students who are from Non Computer Science background. Removing these kinds of records will give us the data that contains only those students whose results in Graduation and Post Graduation can be compared.

Table 1: The highlighted records will be removed in data cleaning process



S/No	Roll No	Percentage in Graduation	Stream in Graduation
1	11/MCA/0001	67.43	Computer Science
2	11/MCA/0002	61.00	Computer Science
3	11/MCA/0003	88.66	Computer Science
4	11/MCA/0004	66.67	Computer Science
5	11/MCA/0005	61.60	Computer Science
6	11/MCA/0006	64.07	Computer Science
7	11/MCA/0007	64.67	Computer Science
8	11/MCA/0008	72.70	Computer Science
9	11/MCA/0009	62.79	Computer Science
10	11/MCA/0010	63.06	Computer Science
11	11/MCA/0011	63.06	Computer Science
12	11/MCA/0012	64.36	Computer Science
13	11/MCA/0013	55.33	Computer Science
14	11/MCA/0014	70.02	Computer Science
15	11/MCA/0015	63.70	Computer Science
16	**11/MCA/0016**	**64.66**	**Non Computer Science**
17	11/MCA/0017	60.45	Computer Science
18	11/MCA/0018	65.80	Computer Science
19	11/MCA/0019	74.00	Computer Science
20	11/MCA/0020	56.67	Computer Science
21	**11/MCA/0021**	**63.88**	**Non Computer Science**
22	11/MCA/0022	68.26	Computer Science
23	11/MCA/0023	62.29	Computer Science
24	11/MCA/0024	58.70	Computer Science
25	11/MCA/0025	76.66	Computer Science
26	11/MCA/0026	68.80	Computer Science
27	11/MCA/0027	74.76	Computer Science
28	11/MCA/0028	75.98	Computer Science
29	11/MCA/0029	59.76	Computer Science
30	11/MCA/0030	68.96	Computer Science

3.6 Data Transformation

he evaluation of the students in their respective courses differ from University to University. Some Universities have Marking system while others have opted for Grading system. In our data which consist of students who have done their graduation from different universities have their results in both the forms, so we have streamlined the data by converting all the results in Grade system. The Grading system that is being followed by the University is shown in Table II.

Table 2: Grading system

Grade	Range of Marks in percent
O	95-100
A	85-94
B	75-84
C	65-74
D	55-64
E	40-54

Table 3: Data Preprocessed (Transformed) to contain Grades only

S/No	Roll No.	Marks Obtained	Max. Marks	Marks in percentage	Grade in Graduation
1	11/MCA/0001	69	100	69	C
2	11/MCA/0002	20	50	40	E
3	11/MCA/0003	51	100	51	E
4	11/MCA/0004	76	125	61	D
5	11/MCA/0005	48	100	48	E
6	11/MCA/0006	47	75	63	D
7	11/MCA/0007	64	100	64	D
8	11/MCA/0008	57	100	57	D
9	11/MCA/0009	88	200	44	E
10	11/MCA/0010	65	100	65	C
11	11/MCA/0011	74	100	74	C
12	11/MCA/0012	55	100	55	D
13	11/MCA/0013	64	100	64	D
14	11/MCA/0014	20	50	40	E
15	11/MCA/0015	52	100	52	E
16	11/MCA/0017	50	100	50	E
17	11/MCA/0018	105	200	53	E
18	11/MCA/0019	70	100	70	C
19	11/MCA/0020	78	100	78	B
20	11/MCA/0022				B
21	11/MCA/0023				B
22	11/MCA/0024				D
23	11/MCA/0025	71	100	71	C
24	11/MCA/0026				C
25	11/MCA/0027				D
26	11/MCA/0028	82	100	82	B
27	11/MCA/0029				C
28	11/MCA/0030	53	100	53	E

4. Generating And Analyzing Association Rules

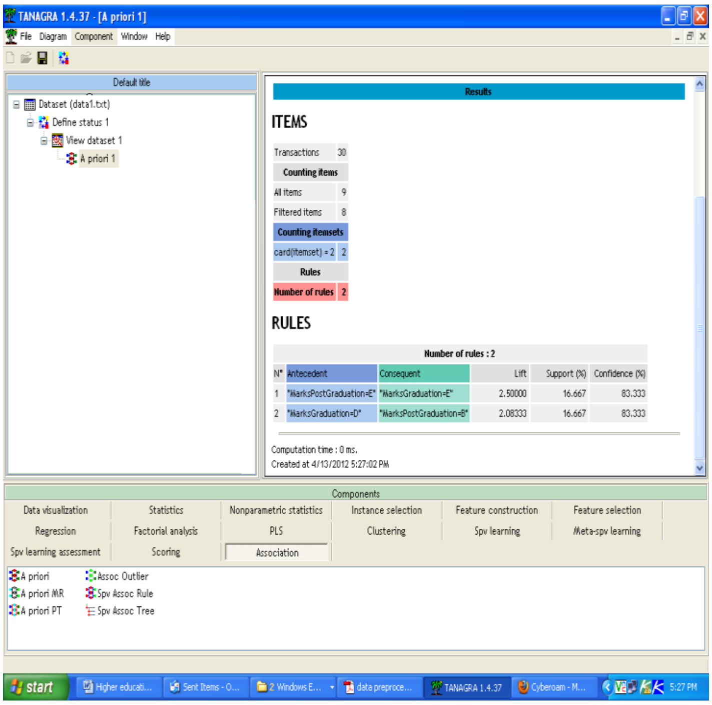
In our study we have used the data mining tool Tanagra to mine Association rules. A snapshot of the data that is utilized to generate the rules is shown in Table IV.

Table 4. Shows the data to be used in generating Association Rules

S/No	Roll No.	Grade in Graduation	Grade in Post Graduation
1	11/MCA/0001	C	A
2	11/MCA/0002	E	E
3	11/MCA/0003	E	E
4	11/MCA/0004	A	A
5	11/MCA/0005	E	B
6	11/MCA/0006	D	B
7	11/MCA/0007	D	B
8	11/MCA/0008	A	A
9	11/MCA/0009	E	B
10	11/MCA/0010	C	A
11	11/MCA/0011	C	A
12	11/MCA/0012	D	B
13	11/MCA/0013	D	B
14	11/MCA/0014	E	C
15	11/MCA/0015	E	C
16	11/MCA/0017	E	B
17	11/MCA/0018	E	B
18	11/MCA/0019	E	E
19	11/MCA/0020	B	B
20	11/MCA/0022	C	B
21	11/MCA/0023	B	B
22	11/MCA/0024	B	B
23	11/MCA/0025	A	C
24	11/MCA/0026	A	C
25	11/MCA/0027	C	A
26	11/MCA/0028	D	E
27	11/MCA/0029	A	C
28	11/MCA/0030	A	C

4.1 Association Rules generated using Tanagra

Below are the snapshots showing different Association Rules with various Support and Confidence factors.



Results

ITEMS

Transactions	30
Counting items	
All items	9
Filtered items	8
Counting itemsets	
card(itemset) = 2	2
Rules	
Number of rules	2

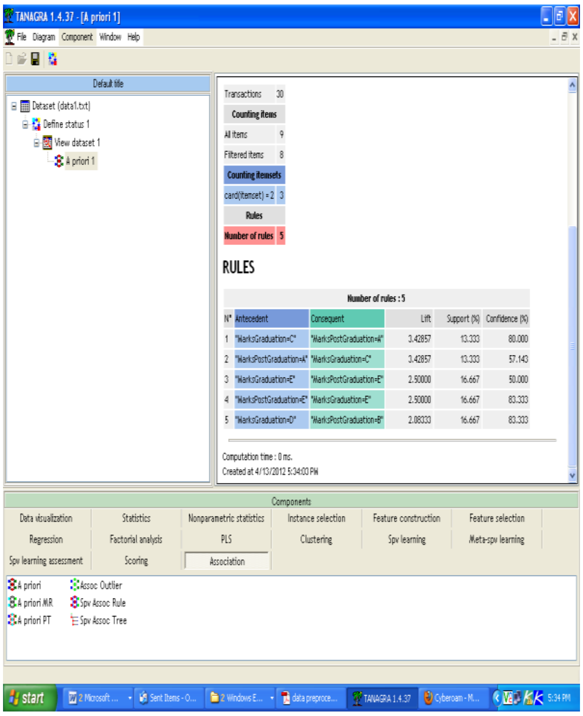
RULES

Number of rules : 2

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"MarksPostGraduation=E"	"MarksGraduation=E"	2.50000	16.667	83.333
2	"MarksGraduation=D"	"MarksPostGraduation=B"	2.08333	16.667	83.333

Computation time : 0 ms.
Created at 4/13/2012 5:27:02 PM

Fig. 2. Snapshot 1. Association Rules mined



Transactions	30
Counting items	
All items	9
Filtered items	8
Counting itemsets	
card(itemset) = 2	3
Rules	
Number of rules	**5**

RULES

Number of rules : 5

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"MarksGraduation=C"	"MarksPostGraduation=A"	3.42857	13.333	80.000
2	"MarksPostGraduation=A"	"MarksGraduation=C"	3.42857	13.333	57.143
3	"MarksGraduation=E"	"MarksPostGraduation=E"	2.50000	16.667	50.000
4	"MarksPostGraduation=E"	"MarksGraduation=E"	2.50000	16.667	83.333
5	"MarksGraduation=D"	"MarksPostGraduation=B"	2.08333	16.667	83.333

Computation time : 0 ms.
Created at 4/13/2012 5:34:03 PM

Fig. 3 Snapshot 2. Association Rules mined

4.2 Analysis of Association Rules

After analyzing the generated Association Rules it is observed that the students who have scored badly in their Graduation have done relatively well in their Post Graduation in the subjects which are common in both Graduate and Post Graduate courses. We can see that the Rule no.2 in snapshot 1(Support 16% and Confidence 83%) and Rule no.1 in Snapshot 2(Support 13% and Confidence 80%) are strong rules in support of this observation. Also from the above generated Association Rules it is observed that there are students whose performance remained poor at both the levels and this can be the matter of concern for the instructors, curriculum planners, academic managers, and other stakeholders. Rule no 1 in snapshot 1(Support 16% and Confidence 83%) is a strong rule favoring this observation. The above observation can be due to various factors i) Students not interested in that particular subject ii) the instructors provided to the students are not able to satisfy their queries iii) the time slot of that particular subject in the Time Table is not favoring their conformability i.e. the time slot is in the afternoon when the students are exhausted and not able to grasp the technicalities of that subject iv) the curriculum of that particular subject is not designed well i.e. it does not contain the basics and starts from the advanced topics directly, that's why the students who are weak in this subject are not able to link themselves with the subject v) students are not interested in the particular course itself .

The generated association rules are of great help for the curriculum planners and academic mangers. The stakeholders of the academic institution can certainly use the hidden knowledge and patterns discovered in the present study for redesigning the curriculum, changing teaching and assessment methodologies, changing the time slot in the Time Table to ensure that the students are fully equipped with the technicalities of the subject and make them capable to perform better in the Post Graduation level. This would not only benefit students but the concerned academic institute in improving the quality of their students so that they can be better placed in their jobs which indirectly will help the institute in better intake of the students.

5. Conclusion

The paper analyzed the potential use of one of the data mining technique called association rule mining in enhancing the quality of students' performances at Post Graduation level.. The mined association rules reveal various factors like student's interest, curriculum design; teaching and assessment methodologies that can affect students who have failed to attain a satisfactory level of performance in the Post Graduation level.

References

[1] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol 40., No. 6, November 2010

[2] M. A. Anwar and Naseer Ahmed, "Knowledge Mining in Supervised and Unsupervised Assessment Data of Students' Performance", *2011 2nd International Conference on Networking and Information Technology IPCSIT vol.17 (2011)*

[3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. "The KDD process for extracting useful knowledge from volumes of data", *CACM* 39 (11), pp. 27-34, 1996.

[4] Fadzilah Siraj and Mansour Ali Abdoulha, " Uncovering hidden Information within University's Student Enrollment Data using Data Mining", *Third Asia International Conference on Modelling and Simulation*, 2009

[5] Fangjun Wu,"Apply Data Mining to student's choosing Teachers under complete Credit Hour", *Second International*

Workshop on Education Technology and Computer Science, 2010

[6] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar and M. Inayat Khan, "Data Mining Model for Higher Education System", *Europen Journal of Scientific Research*, Vol.43, No.1, pp.24-29, 2010

[7] W.M.R. Tissera, R.I. Athauda and H. C. Fernando "Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining", *IEEE International Conference on Information Acquisition*, 2006

[8] Han Jiawei and Micheline Kamber, *Data Mining: Concepts and Technique*, Morgan Kaufmann Publishers, 2000

[9] Hongjie Sun, "Research on Student Learning Result System based on Data Mining", *IJCSNS International Journal of Computer Science and Network Security*, Vol.10, No. 4, April 2010

[10] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, Mustafa I. Al-Najjar, "Mining Student Data Using Decision Trees", *ACIT' 2006: The International Arab Conference on Information Technology*