Educational Data Mining: An Exploration of students' Academic performance Through Association Rule Mining set

Karan Sukhija
Research Scholar
Panjab University, Chandigarh

Abstract: Data Mining is a technology to explore data, analyse the data and finally discovering patterns from large data repository. Currently there is growing interest in data mining and educational systems leading to the development of educational data mining to monitor the progress of student's academic performance. The knowledge is hidden among the educational data set and it is extractable through data mining techniques. This paper discusses the capabilities of data mining techniques in context of school education system to monitor the progression of academic performance of students for the purpose of making an effective decision by the academic planners. The entire process is composed of different phases that begin with the collection of the educational data from relevant sources. After data collection phase, pre-processing of dataset is done to prepare the data for execution and remove all anomalies. Further, the implementation of powerful association rule mining is done for analysing student's result data in order to discover the association rules based on educational dataset. Finally the paper is concluded with the research outcomes for different cases of association rule sets viz. mining from multiple variables and mining from single variable.

Index Terms— Data Mining, Educational Data Mining, Knowledge Discovery in Database (KDD), Association Rule Set, Association Rule Mining.

I. Introduction

The process of analyzing the data and summarizing it into useful information is known as data mining. Technically, data mining is the procedure of discovering useful patterns among dozens of fields existing in large relational databases. It works as analytical tool that allows users to analyse data from different dimensions, classify it, and summarize the relationships identified from data [6].

Educational data mining (EDM) is the latest advancement in education field by applying the data mining techniques. EDM deals with the development of techniques for exploring and analysing the huge data from various educational databases. It involves analysing the results in depth in order to monitor the student's academic activities closely. For educational institutions, promoting students success is a vital need as competition among the institutions is all-time high forcing management to focus on increasing enrolment and registration whilst controlling costs [7].

The activities involved in this complex process do not take place in many education institutions due to the lack of appropriate practices and an adequate technological support that sustain these practices. Data mining, which is defamed as the process of extracting previously unknown knowledge, and detecting the interesting patterns from a massive set of data, offers a way of dealing with this problem in educational institutions. The educational data mining for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the setting which they learn in [8].

The student's academic success is a subject of great importance in education context. Several studies have been performed throughout the globe in Educational field, in order to identify and analyze the student's failure and to propose the measures against this problem. One of the measures frequently pointed out to increase the success promotion is associated to the students closely monitoring and with the approximation of the teacher/tutor to the student's day-by-day academic activities. To monitor the progress of student's academic performance is a critical issue to the academic community of higher learning. A system for analysing the students results based on cluster analysis uses standard statistical algorithms in order to arrange their scores data, according to the level of their performance is described in this paper [9].

II. Related Work:

Association rule mining [1], basically describes relationships between data items in data sets. It helps in finding out the items, which would be selected provided certain set of items have already been selected. An improved algorithm for fast rule generation has been discussed Agrawal et. al (1994).

The online mining of data is performed by pre-processing the data effectively in order to make it suitable for repeated online queries. An online association rule mining technique discussed by [2] Charu C Agrawal at al (2001) suggests a graph theoretic approach, in which the pre -processed data is stored in such a way that online processing may be done by applying a graph theoretic search algorithm.

Khan [3] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream.

Hijazi and Naqvi [4] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed.

Galit [5] gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams.

III. Data Mining Methods:

Data mining, also popularly known as Knowledge Discovery in Database (KDD), refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The techniques and methods in data mining mentioned below to have better understanding.

a. Classification of Data Mining Techniques

By the time many DM techniques and systems have been designed and implemented. These techniques can be classified based on the database, the knowledge to be discovered, and the techniques to be utilized [8]. In this section, we review one of the classification schemes proposed by Chen et al. (1996).

- **Based on the Database**: Different organizations uses different types of databases as per their need and requirements, such as relational database, transaction database, object- oriented database, spatial database, multimedia database, legacy database, and Web databases. A DM system can be classified based on the type of database it is designed to be used for. For example, if the system discovers knowledge from relational database then it is called a relational DM system and if the system finds knowledge from object-oriented database, it is an object-oriented DM system [9] [17].

- **Based on the Knowledge**: DM systems are powerful in discovering different types of knowledge, including association rules, characteristic rules, classification rules, clustering, evolution, and deviation analysis and anomaly detection. Based on abstraction level of the discovered knowledge the DM system can be classified into general knowledge, primitive-level knowledge, and multiple- level knowledge [10] [18].

- **Based on the Techniques:** DM systems can also be classified by nature of DM techniques used to extract information [16]. For example, a DM system can be classified according to the data mining driven force method, such as autonomous knowledge mining, web interaction mining, data- driven mining, user activity over web mining, and query-driven mining. Alternatively, it can be categorized according to its underlying mining approach used, such as generalization-based mining, pattern mining, and statistical data mining or mathematical-based mining and ensemble approaches [11].Some of the mostly used data mining techniques are mentioned as below:

  - **Statistics:** It is an important component in data selection, sample selection and evaluation of extracted knowledge. Statistical analysis can be used to evaluate the outcomes of DM to separate the relevant from the irrelevant. During cleaning of data, statistics offer the different techniques to detect "outliers", to smooth data when necessary, and to identify noise in data. It is with the use of Statistics missing values can be dealt using estimation techniques. For exploratory data analysis techniques like clustering and experimental design can be employed [12] [15].

  - **Mining from Transactional or Relational Database:** Association Rule Mining (ARM) plays important part in mining useful information from the transactional or relational database. The task is to formulate a set of strong association rules in the form of

"A1 ^ . . . ^ Am B1 ^ . . . ^ Bn"

Where, Ai (for i=1 to m) and Bj (for j=1 to n) are attribute-value sets, from the associated data sets in a database. For example, in a database related to market basket one may identify the association rule: if a customer buys one brand of coffee in mall, he/she usually buys another brand of milk in the same transaction. This is because association rule mining requires scanning through a huge transaction database repeatedly; for this the enormous processing power could be required [13] [14].

IV. Educational Data Mining Process:

In this research work educational data mining is performed on school examination system. The subject wise association is performed on student database to find the association or relationship between different subjects. The entire process is composed of below mentioned phases:

I. Data Collection:

The approximate 80,000 records of matriculation standard of the session 2014-15 has been collected from state school education department (* name is hidden subject to confidentiality constraint) based on the sampling method. Table 1 depicts various variables of data set based on parameters variable name, description and possible values.

Variable Name	Description	Possible Values
Sex	Gender information	Male, Female
Area	Location	Rural, Urban
Eng_Res	English Subject Result	Pass, Fail
Eng_Grade	English Subject Grade	A+,A,B,C,D,E
Math_Res	Math Subject Result	Pass, Fail
Math_Grade	Math Subject Grade	A+,A,B,C,D,E
Punjabi_Res	Punjabi Subject Result	Pass, Fail
Punjabi_Grade	Punjabi Subject Grade	A+,A,B,C,D,E
Hindi_Res	Hindi Subject Result	Pass, Fail
Hindi_Grade	Hindi Subject Grade	A+,A,B,C,D,E
Sci_Res	Science Subject Result	Pass, Fail
Sci_Grade	Science Subject Grade	A+,A,B,C,D,E
SS_Res	SS Subject Result	Pass, Fail
SS_Grade	SS Subject Grade	A+,A,B,C,D,E
Final Result	Overall Result	Pass, Fail, Reappear
Final_Grade	Overall Grade	A+,A,B,C,D,E

II. Data Selection and Transformation:

The selection and transformation of dataset has performed after the collection of dataset. Initially only required fields are taken into consideration for mining process. Initially, some derived variables are also considered and some of the information for the variables has been extracted from the database. Furthermore, the all dataset files are converted into .csv format which is required for mining process. Moreover, so many techniques are applied to remove the anomalies in the dataset and prepare them into transaction based format. Three different transaction based files have prepared from database. The following view described the each one in detail.

- **Transaction file based on Pass, Fail, Re-appear**: A Transaction file contain records that are based on the different subjects in the form of pass, fail and reappear have failed is shown below:

Fig.1. Transaction File 1

- **Transaction file based on Fail**: A Transaction file contains records that are based on the different subjects in which students have failed is shown below:

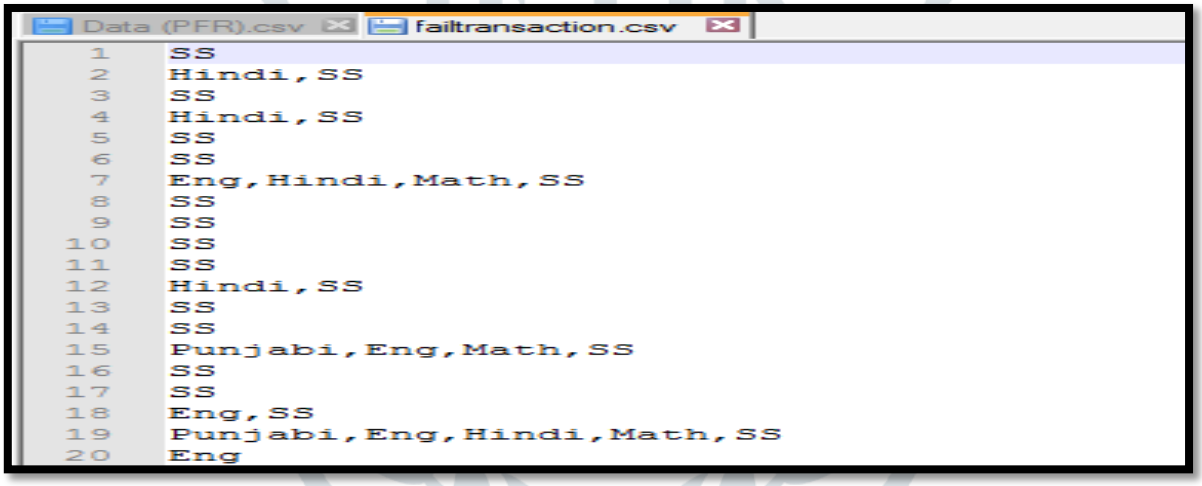

Fig.2. Transaction File 2

- **Transaction file based on Pass**: A Transaction file contains records that are based on the different subjects in which students have passed is shown below:

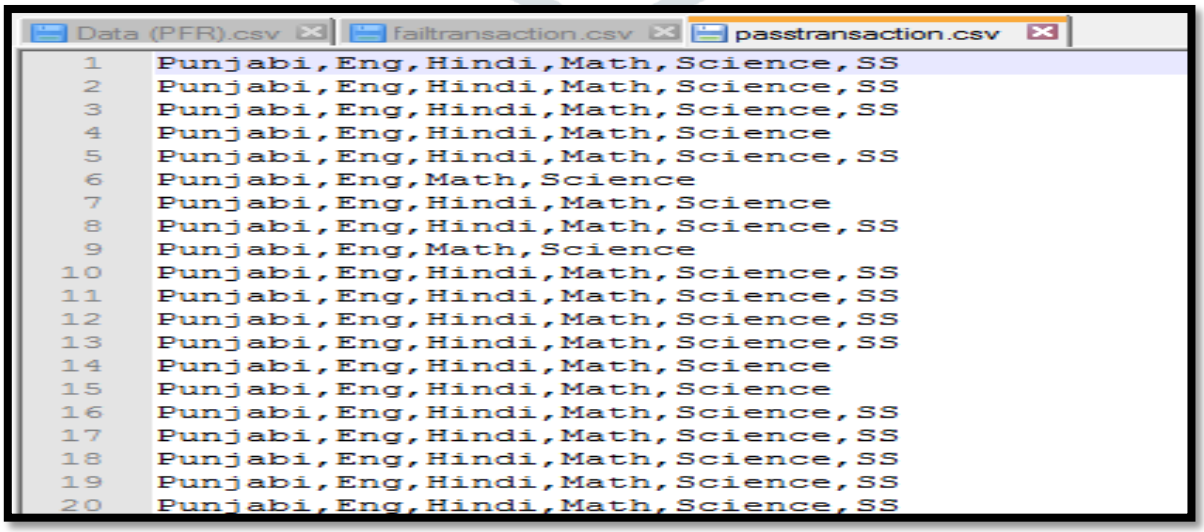

Fig.3. Transaction File 3

V. Association Rule Mining Process:

Aforementioned data set represents the analysis of real data collected from several institutions and a sample study has been conducted to depict how the Apriori Algorithm can be used in educational field and their corresponding results have been observed. Once the frequent item sets from transactions based dataset have been found, it is straightforward to generate strong association rules from them where strong association rules satisfy both minimum support and minimum confidence. R-language and weka tool the powerful open source software is used for implementation of association types of algorithm. Here, it is used for perform mining on educational data. In this manuscript, I tried to find the subjects wise association and find the relationship on among different subjects.

I. **Mining from Multiple Variables**: The following association rule set retrieved from dataset which is based on passed based transaction file.

	lhs		rhs	support	confidence
1	{Eng_res=P,Sci_result=P}	=>	{Hindi_res=P}	0.8133	0.9830775
2	{Eng_res=P,Hindi_res=P}	=>	{Sci_result=P}	0.8133	0.9799976
3	{Hindi_res=P,Sci_result=P}	=>	{Eng_res=P}	0.8133	0.9036667
4	{Eng_res=P,Sci_result=P}	=>	{Math_result=P}	0.8235	0.9954067
5	{Eng_res=P,Math_result=P}	=>	{Sci_result=P}	0.8235	0.9817597
6	{Math_result=P,Sci_result=P}	=>	{Eng_res=P}	0.8235	0.9013792
7	{Eng_res=P,Sci_result=P}	=>	{Pun_res=P}	0.8258	0.9981869
8	{Pun_res=P,Eng_res=P}	=>	{Sci_result=P}	0.8258	0.9804108
9	{Eng_res=P,Hindi_res=P}	=>	{Math_result=P}	0.8241	0.9930112
10	{Eng_res=P,Math_result=P}	=>	{Hindi_res=P}	0.8241	0.9824750
11	{Eng_res=P,Hindi_res=P}	=>	{Pun_res=P}	0.8274	0.9969876
12	{Pun_res=P,Eng_res=P}	=>	{Hindi_res=P}	0.8274	0.9823103
13	{Eng_res=P,Math_result=P}	=>	{Pun_res=P}	0.8373	0.9982117
14	{Pun_res=P,Eng_res=P}	=>	{Math_result=P}	0.8373	0.9940639
15	{Sci_result=P,SS_result=P}	=>	{Hindi_res=P}	0.8389	0.9781950
16	{Hindi_res=P,SS_result=P}	=>	{Sci_result=P}	0.8389	0.9730890
17	{Hindi_res=P,Sci_result=P}	=>	{SS_result=P}	0.8389	0.9321111
18	{Sci_result=P,SS_result=P}	=>	{Math_result=P}	0.8513	0.9926539
19	{Math_result=P,SS_result=P}	=>	{Sci_result=P}	0.8513	0.9749198
20	{Math_result=P,Sci_result=P}	=>	{SS_result=P}	0.8513	0.9318082

Fig.4. Association rule set having multiple variables

- **Mining from single variable**: The following association rule set retrieved from dataset which is based on fail based transaction file.

	lhs		rhs	support	confidenc
1	{Eng}	=>	{SS}	0.8691309	0.9571324
2	{SS}	=>	{Eng}	0.8691309	0.9494764
3	{Eng}	=>	{Hindi}	0.8927580	0.9831518
4	{Hindi}	=>	{Eng}	0.8927580	0.9385102
5	{Eng}	=>	{Science}	0.8941174	0.9846489
6	{Science}	=>	{Eng}	0.8941174	0.9395127
7	{Eng}	=>	{Math}	0.8990742	0.9901075
8	{Math}	=>	{Eng}	0.8990742	0.9345224
9	{Eng}	=>	{Punjabi}	0.9057179	0.9974239
10	{Punjabi}	=>	{Eng}	0.9057179	0.9253194
11	{SS}	=>	{Hindi}	0.8975929	0.9805696
12	{Hindi}	=>	{SS}	0.8975929	0.9435929
13	{SS}	=>	{Science}	0.8974103	0.9803701
14	{Science}	=>	{SS}	0.8974103	0.9429728
15	{SS}	=>	{Math}	0.9046628	0.9882930
16	{Math}	=>	{SS}	0.9046628	0.9403314
17	{SS}	=>	{Punjabi}	0.9120747	0.9963900
18	{Punjabi}	=>	{SS}	0.9120747	0.9318137
19	{Hindi}	=>	{Science}	0.9272956	0.9748178
20	{Science}	=>	{Hindi}	0.9272956	0.9743754

Fig.5. Association rule set having single variable

VI. Research Outcomes:

The above mentioned both cases varied in terms of their results. Case 1 is concerned with association rule set for multiple variables i.e. majority of the students who passed in English course(s) have also passed in many other subjects, including Hindi and Science subjects. Case 2 is concerned with association rule set for single variable i.e., consider only single variable and different association rules are retrieved as result. Those students who have failed in English also failed in Social Study in majority of cases and the same relationship in case of Hindi also.

VII. Conclusion:

This research work will help to analyze the student's performance that will assist teachers to improve the result of the student's. It will also work to identify those students which needed special attention to reduce dropout rate and make strategies for the next examination system. The entire process is composed of different phases that begin with the collection of the educational data from relevant sources. After data collection phase, pre-processing of dataset is done to prepare the data for execution and remove all anomalies. Further, the implementation of powerful association rule mining is done for analysing student's result data in order to discover the association rules based on educational dataset.

VIII. References:

[1] *Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.*

[2] *Aggarwal, Charu C., and Philip S. Yu. "A new approach to online generation of association rules." Knowledge and Data Engineering, IEEE Transactions on 13.4 (2001): 527-540.*

[3] *Z. N. Khan, "Scholastic achievement of higher secondary students in science stream",Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.*

[4] *S. T. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student's performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.*

[5] *Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.*

[6] *Singh, Archana, et al. "Online Mining of data to generate association rule mining in large databases." Recent Trends in Information Systems (ReTIS), 2011 International Conference on. IEEE, 2011.*

[7] *Khan, Muhammad Asad, WajebGharibi, and Santanu Kumar Pradhan. "Data mining techniques for business intelligence in educational system: A case mining." Computer Applications and Information Systems (WCCAIS), 2014 World Congress on IEEE, 2014.*

[8] *A. F. D. Costa and 1. T Lopes. OsEstudantes e osseusTrajectos no Ensino Superior: Sucesso e Insucesso, Factores e processos, Promocao de Boas Praticas. 2008 Retrieved July 2009.*

[9] *Man Wai Lee, Sherry Y. Chen, Kyriacos Chrysostomou, Xiaohui Liu, "Mining student's behavior in web-based learning programs" Expert Syst. Appl. 36(2): 3459-3464 (2009).*

[10] *Marquez-Vera, Carlos, Cristobal Romero Morales, and Sebastian Ventura Soto. "Predicting school failure and dropout by using data mining techniques." Tecnologias delAprendizaje, IEEE RevistaIberoamericana de 8.1 (2013): 7-14.*

[11] *Guleria, Pratiyush, Manish Arora, and Manu Sood. "Increasing quality of education using educational data mining." Information Management in the Knowledge Economy (IMKE), 2013 2nd International Conference on IEEE, 2013.*

[12] *Abdullah, Zailani, et al. "Mining Least Association Rules of Degree Level Programs Selected by Students." International Journal of Multimedia and Ubiquitous Engineering 9.1 (2014): 241-254.*

[13] *Jiménez-Gómez, Manuel Ángel, et al. "Discovering clues to avoid middle school failure at early stages." Proceedings of the Fifth International Conference on Learning Analytics and Knowledge ACM, 2015.*

[14] *Wook, Muslihah, Zawiyah M. Yusof, and MohdZakree Ahmad Nazri. "The Acceptance of Educational Data Mining Technology among Students in Public Institutions of Higher Learning in Malaysia." International Journal of Future Computer and Communication 4.2 (2015).*

[15]Ryan S. Baker, ―Educational Data Mining:An Advance for Intelligent Systems in Education", IEEE Society, 2014.

[16]Nor BahiahHj Ahmad, SitiMariyamShamsuddin,"A Comparative Analysis of Mining Techniques forAutomatic Detection of Student's Learning Style", IEEE, 2010 , 978-1-4244-8136.

[17]Agarwal, R., &Srikant, R., ―Mining sequential patterns," In Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan, 2005, (pp. 3– 14).

[18]Jain, A. K., Murty, M. N., & Flynn, P. J., ―Data clustering: A Review," ACM Computing Surveys, 31(3), 1999, (pp. 264–323).