



RESEARCH ARTICLE

An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education

¹Mahendra Tiwari,

Research Scholar, (UPRTOU, Allahabad)

²Randhir Singh,

Assistant Professor, (UIM, Allahabad)

³Neeraj Vimal,

Assistant Professor, (Azad Inst.of Eng. & Technology, Lucknow)

Abstract— Educational institutions are important parts of our society and playing a vital role for growth and development of nation and prediction of student's performance in educational environments is also important as well. Student's academic performance is based upon various factors like personal, social, psychological etc. Educational data mining concerns with developing methods for discovering knowledge from data, that comes from educational domain. The Data Mining tool has accepted as a decision making tool which is able to facilitate better resource utilization in terms of students performance. In this paper a student data from an engineering college has been taken and various data mining methods have been performed. This paper addresses the applications of data mining in educational institution to extract useful information from available data set and providing analytical tool to view. The result of study is aimed to develop a faith on data mining techniques so that present education system may adopt this as a strategic management tool.

Indexed Terms: - Academic performance, Data mining, Data classification, Clustering, Student's result database.

I. INTRODUCTION

Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process [7].

Evaluating students' performance is a complex issue, which can't be restricted for the grading. Reasons of good or bad performances belong to the main interests of teachers, because they can plan and customize their teaching program, based on the feedback. Data mining is one of the approaches, which can provide an effective assistance in revealing complex relationships behind the grades [5].

Data mining consists of a set of techniques that can be used to extract relevant and interesting knowledge from data. Data mining has several tasks such as association rule mining, classification and prediction, and clustering. Classification techniques are supervised learning techniques that classify data item into predefined class label. It is one of the most useful techniques in data mining to build classification models from an input data set. The used classification techniques commonly build models that are used to predict future data trends.

There are several algorithms for data classification such as decision tree and Rule. With classification, the generated model will be able to predict a class for given data depending on previously learned information from historical data [3].

There are increasing research interests in education field using data mining. Application of Data mining techniques concerns to develop the methods that discover knowledge from data and used to uncover hidden or unknown information that is not apparent, but potentially useful[8],The data can be personal or academic which can be used to understand students behavior to assist instructors, to improve teaching, to improve curriculums and many other benefits.

This study investigates and compares the educational domain of data mining from data that come from student's behavior. It showed that various kind of data could be collected, how could we preprocess the data, how to apply methods of data mining on the data, and finally how can we benefited from the discovered knowledge. In this study, university students were predicated his/her final grade by using classification and grouped the students according to their similar characteristics, forming clusters.

II. RELATED WORK

Data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, and proactively manage student outcomes. The management system can improve their policy, enhance their strategies and thereby improve the quality of that management system [1].

Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. It showed how useful data mining can be in higher education in particularly to predict the final performance of student [2], on working on performance, many attributes have been tested, and some of them are found effective on the performance prediction. The job title was the strongest attribute, then the university type, with slight effect of degree and grade [3].

Data Mining could be used to improve business intelligence process including education system to enhance the efficacy and overall efficiency by optimally utilizing the resources available. The performance, success of students in the examination as well as their overall personality development could be exponentially accelerated by thoroughly utilizing Data Mining technique to evaluate their admission academic performance and finally the placement [4]. The results provide the information on which mandatory subjects are essential in determining the success of students. These two classifiers were used to differentiate the students. ID3 algorithm is better than CART algorithm when tested for model performance with cross validation and tested on evaluation data set (20% number of data, randomly selected)[5]. Students' result repository is a large data bank which shows the students raw scores and grades in different courses they enrolled for during their years of attendance in the institution. Student performance score is basically determined by the sum total of the continuous assessment and the examination scores [6].

III. DATA COLLECTION AND PREPARATION

In our case study we collected the student data of B.Tech second year (CS & IT branch) from database management system course held at the United College of Engineering and Research Naini Allahabad (Affiliated to GBTU) in fourth semester of 2011/2012 and we used questionnaire to collect the real data that describing the relationships between learning behavior of students and their academic performance.

The variable for judging the learning and academic behavior of students was used in questionnaire are *Assignment, Attendance, Sessional marks, GPA (grade point average for general performance in lab or extracurricular), and Final_ grade (current semester)*. We grouped all grades in to five possible values (Excellent, Good, Average, Poor, Fail).

Table 1: Attributes and its possible values

Attribute	Description	Possible values
Assignment	Online exercise given by teacher	Good, Poor
Attendance	Attendance in one semester	Excellent, Good, Average, Poor
Sessional marks	Percentage of marks obtained in internal exam	Excellent, Good, Average, Poor, Fail
GPA	Grade point average for general performance i.e. in lab or extra-curricular	Good, Poor
Final grade	Percentage of marks obtained in current semester exam	Excellent, Good, Average, Poor, Fail

IV. APPLICATION OF DATA MINING TECHNIQUES TO STUDENTS DATASET: RESULT AND DISCUSSION

There is a work methodology which governs our work. The methodology start from the problem definition, then Data collection from student database are discussed, Data is organize so there is no need of preprocessing, then we come to data mining methods that are association, classification, and clustering followed by evaluation of result.

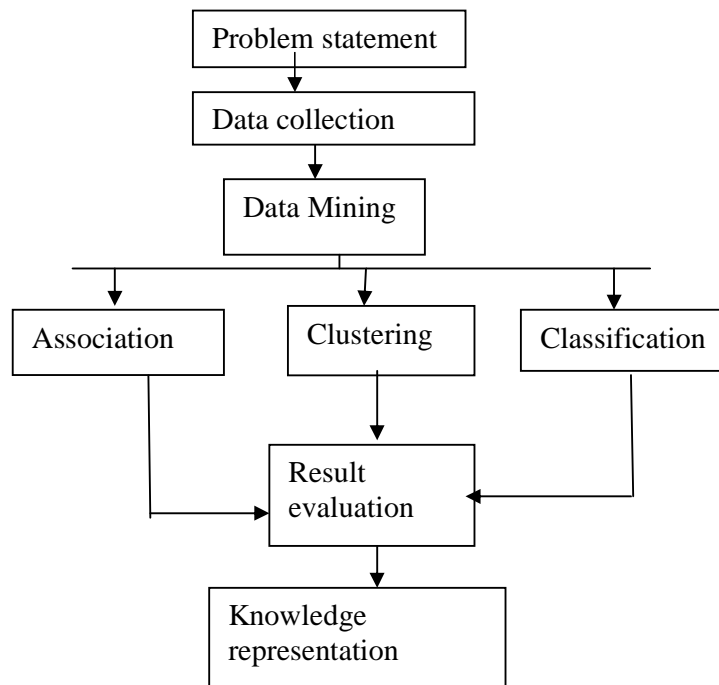


Figure 1: Work Methodology

Association Analysis:

This area of data mining aims at analyzing data to identify consolidated occurrence of events and uses the criteria of support and confidence. It is known to be applied in student behavior [7].

Mining association rules searching for interesting relationships among items in given data set. In our data set association rule mining is used to identify possible grade values i.e. Excellent, Good, Average, Poor, Fail.

[Attendance=poor, Assignmet=poor,GPA=poor]→ [Grade=poor]

(Support: 0.196, confidence: 0.757)

[Attendance=poor,Sessional=poor,GPA=poor]→ [Grade=poor]

(Support: 0.166, confidence: 0.657)

[Assignmet=poor,Final_grade=poor,GPA=poor]→ [Grade=poor]

(Support: 0.176, confidence: 0.737)

[Attendance=poor,Sessional=poor,Final_grade=poor]→ [Grade=poor]

(Support: 0.296, confidence: 0.747)

The resulting Association rule depicts a sample of discovered rules from data for student with poor grade along with their support and confidence.

To interpret the rules in association rules model, the first rule means that of engineering students under study, 19%(support) are poor in attendance, poor in assignment, having poor in GPA. There is 75% probability or confidence that student will get the grade poor and so on.

Classification:

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. A Rule-based classification extracts a set of rules that show relationships between attributes of the data set and the class label. It used a set of IF-THEN rules for classification.

If Attendance=excellent and Assignment=good and Sessional marks=excellent and GPA=good and Final_grade=excellent, then excellent.

If Attendance=excellent and Assignment=good and Sessional marks=good and GPA=good and Final_grade=good, then good.

If Attendance=average and Assignment=good and Sessional marks=good and GPA=good and Final_grade=good, then average.

If Attendance=poor and Assignment=poor and Sessional marks=average and GPA=poor and Final_grade=poor, then poor.

Association rules are characteristic rules (it describes current situation), but classification rules are prediction rules for describing future situation.

Clustering:

Clustering is a division of data into groups of similar objects. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others[8].

The K-means algorithm, probably the best one of the clustering algorithms proposed, is based on a very simple idea: Given a set of initial clusters, assign each point to one of them, and then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence [9].

The objective of this k-means test is to choose the best cluster center to be the centroid. The k-means algorithm requires the change of nominal attributes in to numerical. The clustering method produced a model with five clusters.

Mean of cluster:

Attribute	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Attendance	76.089	70.678	68.434	63.879	61.567
Assignment	73.098	70.678	66.987	60.667	57.234
Sessional	71.976	67.887	65.985	60.565	58.569
GPA	14.718	13.788	12.878	10.987	9.765
Final_grade	87.056	84.786	80.887	73.765	66.987

V. CONCLUSION

In this paper, we discussed the various data mining techniques which can support education system for decision making. It showed how useful data mining can be used in higher education specially to improve engineering students' performance. We applied data mining techniques to discover knowledge, association rules, classification rules to predict the students' performance, as well as we clustered the students in to groups using k-means clustering algorithm.

REFERENCES

- [1] M. Sukany et al, Data Mining: performance improvement in education sector using classification & clustering algorithm, ICCCE, April 2012.
- [2] Sajadin Sembiring et al, Prediction of student academic performance by an application of data mining techniques, IPEDR, vol 6, IACSIT press, Indonesia.
- [3] Qasem A., Eman nagi, Using data mining techniques to build a classification model for predicting employees performance, IJACSA, vol 3 no. 2, 2012.
- [4] Sonali et al, Data Mining in Education: Data classification and Decision tree approach, International journal of e-education, e-business, e-management and e-learning, vol 2 no. 2, April 2012.
- [5] Songkroh, ko, Evaluation of student performance with data mining: An application of ID3 and CART algorithm, Fourth international conference on S/W knowledge, Information Management and Application.
- [6] Chandra, Nandhini, Knowledge mining from student data, European journal of scientific research, vol 47 no. 1, 2010.
- [7] Moudary et al., Improving student performance using data clustering and neural networks in foreign language based higher education, The Research Bulletin of Jordan ACM, vol II (III).
- [8] Pavel Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, Inc.
- [9] Glenn Fung, A Comprehensive Overview of Basic Clustering Algorithms, June 22, 2001
- [10] Edin, Mirza, Data Mining approach for predicting student performance, journal of economics and business, vol x, issue 1, may 2012
- [11] Deen, Ahmad, classification based on association-rule mining techniques: a general survey and empirical comparative evaluation, Ubiquitous Computing and Communication Journal vol 5 no 3
- [12] A. Chan, et al, Online Course Refinement through Association Rule Mining, Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008