ISSN:0975-9646

Educational Data Mining and its role in Educational Field

P. Meena Kumari[1], SK.Abdul Nabi[2] and Puppal Priyanka[3]

[1]*Department of CSE,AVN Inst. of Engg. & Tech., Hyderabad.*
[2]*Professor & HOD, Department of CSE, AVN Inst. Of Engg. & Tech., Hyderabad.*
[3]*Department of CSE,AVN Inst. Of Engg. &Tech., Hyderabad.*

***Abstract*- Data mining is about explaining the past and predicting the future by means of data analysis. Educational Data Mining is a promising discipline which has an imperative impact on predicting students' academic performance. Thousands of students take admissions in Universities and colleges every year, at the time of admissions they collect the students' data. In the same way while the Teachers join in the institution they collect their personal and professional data. Understand the importance of data is essential from a business point of view. Data collected at the time of admission can be used for classifying and predicting students' behavior and performance as well as teachers' performance.**

Therefore, in this paper, we are examining the role of Data mining in an Educational Field. By using SDAR, we have identified possible grade values i.e., Excellent, Good, Average and Poor or Fail. We have used K-means clustering algorithm to find the best cluster center for attributes like attendance, Sessional marks and assignment marks etc. We have also discussed a Rule-Based Classification (RBC) method; it extracts a set of rules that shows relationships between attributes of the data set and the class label. In this paper we have also addressed the evaluation of Teachers' performance by using data mining techniques at University and College level.

Keywords- EDM, SDAR, K-Means Clustering, RBC, Student Feedback.

I. INTRODUCTION

The process of analyzing data from different perspectives and summarizing it into important information so as to identify hidden patterns from a large data set is called Data Mining. Educational Data Mining (EDM) is an application of Data Mining and statistics to information generated from educational settings like Universities and Colleges [1].

In fact, EDM refers to techniques and tools designed for automatically extracting meaning from large repositories of data generated by peoples learning activities in educational settings. Data Mining can also be defined as a multi-disciplinary field which combines artificial Intelligence statistics and database technology [3]. The relationship between data mining, data base and AI & machine learning is shown in figure 1.

Many businesses like Supermarkets, hotels, factories have stored large amounts of data over years of operations, and data mining can extract very valuable knowledge from this data. Then these business people can make greater profits by attracting more customers and by improving sales. This is possible with engineering and medical fields.

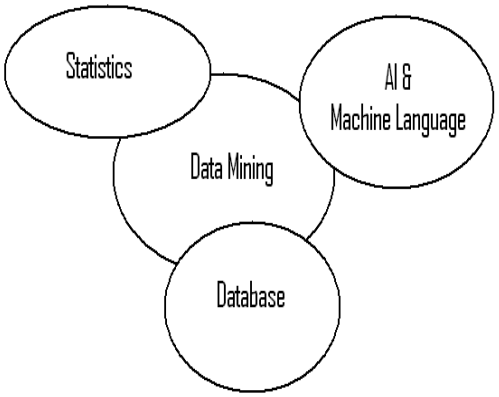

Figure 1. Relations ship among Datamining, AI with database

In this paper, association rule (data mining technique) is applied to evaluate student behavior. K means clustering is used to find out the best centroid for student data like attendance, Sessional marks, GPA, final grade etc. A Rule-Based Classification (RBC) is also used to extract relationships between attributes of the data set and the class label. And also we have discussed the teachers' performance evaluation using data mining techniques at University and College level. The objective of Teacher performance evaluation is as follows:

- Betterment of the Students
- To improve the Students' performance
- As a motivation to Teacher
- To guide the teacher for further improvement
- To improve self respect and ambition

Association rule (an important method of Data mining) is proposed by Agarwal in 1993. It is a popular and well researched method for discovering interesting relations between variables in a huge a database. Based on this we have proposed SDAR for evaluating the student performance and it was discussed in the further sections.

Clustering is the process of dividing the dataset into smaller groups such that each item in the group is similar or close to each other as possible. The two main groups of clustering algorithms are Hierarchical and Partitive.

Classification is a Data Mining task used to predict a value of categorical variable by building a model based on one or more numeric or categorical variable (or attributes)[5].

II. RELATED WORK

Data Mining is a powerful tool that enables educational institutions to better allocate resources and staff and proactively manage student outcomes [2]. The educational system can improve their policies, can enhance their strategies and their by can improve quality of their management system [8]. Different Data Mining techniques have the capability to provide effective, improving tools for student and teacher performance.

Data Mining could be used to improve business intelligence process, including the education system to enhance the overall efficiency by optimally utilizing the resources available.

The performance or success of students in the examination as well as their overall personality development could be exponentially accelerated by thoroughly utilizing Data Mining techniques to evaluate their admission academic performance and finally the Placement in an organization.

III. DATA COLLECTION

In our case study, we have collected student data of III year B.Tech. (CSE & ECE) of AVNIET Engineering college (Affiliated to JNTU, Hyderabad). In this process, we have used a questionnaire to collect the real data from the students that describes the relationship between learning behavior and their academic performance. And also we collected the Faculty (or teacher) feedback from the students.

The variables for judging the learning and academic behavior of students used in the questionnaire are Assignment, Attendance, Sessional marks, GPA (grade point average for general performance in the lab or extracurricular) and Final_ grade in last semester. We grouped all grades into five possible values they are Excellent, Good, Average, Poor and fail which is shown in the table 1.

TABLE .1

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

IV. IMPLEMENTATION (EDUCATIONAL DATA MINING)

A. Students Dataset

There is a work methodology which governs our work. The methodology starts from the problem definition, then data collection from Students Database. And next stage is preprocessing, but it doesn't require as it organizes data. Next Data Mining methods that are association, clustering and classification followed by evaluation of results which is shown in the figure 2.

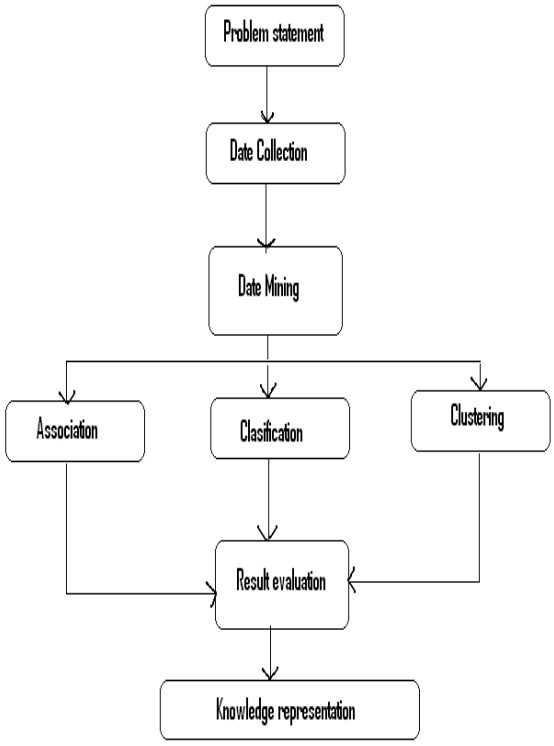

Figure 2. Working Methodology

1. Association

Association Data Mining aims at analyzing the data to identify an occurrence of events and uses the criteria of Support and confidence. In this paper, it is applied to evaluate student behavior. Association rules are nothing but mining for interesting relationships among items in a given data set. Here Student Data set Association Rule(SDAR) mining is used to identify possible grade values. i.e., Excellent, Good, Average, Poor or Fail.

[Attendance=poor, Assignment=poor, GPA=poor] → [Grade=poor]
(Support:0.196, Confidence:0.757)

[Attendance=poor, Sessional marks =poor, GPA=poor] → [Grade=poor]
(Support:0.166, Confidence:0.657)

[Assignment =poor, Final grade=poor, GPA=poor] → [Grade=poor]
(Support:0.176, Confidence:0.737)

[Attendance=poor, Sessional marks =poor, Final grade =poor] → [Grade=poor]
(Support:0.296, Confidence:0.747)

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

We can interpret the above association rule result as,

- 19%(support) students are poor in their class attendance, poor in assignment, having poor in GPA. There is 75% probability or confidence that student will get the grade poor and so on.

-16%(support) students are poor in their class attendance, poor in sessional marks, having poor in GPA. There is 6% probability or confidence that student will get the grade poor and so on.

2. Clustering

Clustering is a division of data into groups of similar objects. Clustering plays an outstanding role in data mining applications such as information retrieval and text mining, scientific data exploration, web analysis, spatial database applications, medical diagnostics, marketing and many more.

The K-means is one of the best clustering algorithms in data mining[6].K-Means Clustering (KMC) proposes to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. Exactly *k* different clusters have been produced by this method with greatest possible characteristic. Initially best number of clusters *k* leading to the greatest separation (distance) is not known and must be computed from the data. K-Means clustering's objective is to minimize the squared error function or total intra-cluster variance.

Choosing the best cluster center to be the centroid is the objective of this k-means test. The k-means algorithm requires the change of small attributes in to numerical. Five clusters have been produced by the clustering method which is shown in table 2.

TABLE 2. Mean of Cluster

Attribute	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5
Attendance	74.023	72.124	68.434	62.879	58.567
Assignment	73.098	70.678	65.248	60.157	52.487
Sessional marks	74.257	65.578	63.025	58.247	55.248
GPA	14.578	12.247	11.247	10.478	09.478
Final Grade	89.247	85.247	79.247	72.247	66.057

3. Classification

Machine learning Classification [7] is a typical data mining technique. Normally classification is used to classify each item in a set of data into one of predefined set of groups or classes**.** A Rule-Based Classification (RBC) extracts a set of rules that show relationships between attributes of the data set and the class label.

A set of IF-THEN rules are used for classification.

If Final_grade=excellent and GPA=good and Sessional marks=excellent and assignment=good and attendance=excellent, then excellent.

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

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

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

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

B. Teacher Dataset

Teachers' performance [4] is evaluated using data mining techniques in this paper. Initially a survey of the teachers' requirements and students' requirements is made. Then we interacted with the teachers and got some knowledge about their methods. We met different teachers that have been given some ideas about the finding of the teacher's performance.

1. Data Preparations

For this study, Feedback from the students' were collected from B.Tech., students of AVNIET (Affiliated to JNTU, Hyderabad) , for two semesters from two different branches for the purpose of investigating if there is any changes in their evaluation of lecturers over that period. The demographic data is not collected at the time of Survey but the college has that information and used for this paper. The profile of students is defined based on the academic and cultural backgrounds of students. The students' academic background is measured using the entry requirements to be fulfilled to get entry into the university/college.

2. Data Selection and Transformation

In this stage only the data required for data mining are selected. A few derived variables were selected. From the available database, some of the information for the variables is collected. The data collected from Feedback forms and database is shown in Table 3.

TABLE 3. Feedback forms data

Variable	Description	Possible value
Name	Lecturer Name	Text
PR	Presentation	{1,2,3,4,5}
CO	Communication	{1,2,3,4,5}
CA	Content arrangement	{1,2,3,4,5}
SD	Speed of delivery	{1,2,3,4,5}
CD	Content delivery	{1,2,3,4,5}
EP	Explanation power	{1,2,3,4,5}
DC	Doubts clearing	{1,2,3,4,5}
OCS	Overall completion of Syllabus	{1,2,3,4,5}
SA	Students attendance	{1-below,2-average,3-high}
RE	Result	{1-pass, 2-fail, 3-pramoted}
PT	Performance of Teacher	{1,2,3,4,5}

The survey uses the following scale for items,

5 = strongly agree
4 = agree
3 = disagree
2 = strongly disagree
1 = neither agree nor disagree

The domain values for some of the variables are as follows.

SA (Attendance of the Student- Minimum attendance required to appear for the end examination is 75%. But in some special cases low attendance students are allowed to appear for the end examinations on genuine reasons. Attendance is divided into three classes. Poor<65%, Average >=65% and <75%, Good>=75%

RE (Students Result in End examinations- It is split into three classes like "Pass", "Fail" or "Promoted". If student passes in all the papers then "pass" is awarded, if student fail in up to three theory and two lab exams then the student will be "promoted" to next academic year else awarded as "Fail".

3. Implementation of Mining Model

Weka is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. Weka stands for Waikato Environment for Knowledge Analysis[9]. From the above faculty data, teacher.arff file was created, and then this file was loaded into WEKA explorer. To visualize erroneous predictions, to estimate the accuracy of the resulting predictive model, and, or the model itself, the classify panel enables the user to apply classification and regression algorithms to the resulting dataset. Naive Bayes, ID3, CART and LAD are the algorithm used for classification.

Under the "Test options", the 10-fold cross-validation is selected as our evaluation approach.

We can say it is necessary to get a reasonable idea of accuracy of the generated model because there is no separate evaluation data set.

To better understand the importance of the input variables, it is customary to analyse the impact of input variables on the performance of teacher's evaluation, in which the impact of input variable of the model on the output variable has been analysed. Tests were conducted using four tests for the assessment of input variables: Chi-square test, Gain Ratio test and Info Gain test. Different algorithms were provide different results, i.e. each of them accounts the relevance of variables in a different way, therefore the average value of all the algorithms is taken in the final result of variables ranking, instead of selecting one algorithm and evaluating it.

V. CONCLUSION

In this paper, we have discussed the various data mining techniques which can support education system for decision making. Initially, we discussed that how Data mining can be used to evaluate 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.

Teachers' regular assessment is suggested to maintain the quality in higher education. As a conclusion, we have reached our objective which is to evaluate and investigate teacher's performance by the four selected classification algorithms based on Weka.

On working on performance, many attributes have been tested, and some of them are found effective on the performance prediction. The content arrangement was the strongest attribute, and then the result plays an important role in the performance of teachers. The speed of delivery attribute did not show any clear effect while the overall completion of course and regularity attribute has shown some effect in some of the experiments for predicting the performance. Other attributes had a degree of effect on predicting the performance.

In this paper Students and Teachers performances are evaluated using different Data mining methods. This can be useful in the educational system like Universities and colleges; they can improve their standards and reputations by introducing the new courses or branches. As a result the quality of education can be improved.

ACKNOWLEDGMENT

We would like to thank everyone who has motivated and supported us for preparing this manuscript.

REFERENCES

[1] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques" Simon Fraser University Morgan Kaufmann Publishers, 2001.

[2] Richard A. Huebner, Norwich University, "A Survey on Educational Data Mining", Research in Higher Education Journal. (www.aabri.com)

[3] B.K. Bharadwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance", International Journal of Advance Computer Science and Applications, Vol. 2, No. 6, 2011.

[4] Ajay Kumar Pal1, Saurabh Pal, "Evaluation of Teacher's Performance: A Data Mining Approach", IJCSMC, Vol. 2, Issue. 12, December 2013,

[5] Aggarwal, C. Charu and Yu, S. Philip. "Data Mining Techniques for Associations, Clustering and Classification." in Zhong, Ning and Zhou, Lizhu (Eds.) methodologies for knowledge discovery and data mining, third pacific Asia Conference, PAKDD, Beijing, China, April 26-28, 1999 proceedings, Springer, New York

[6] Suchita Borkar, K. Rajeswari, "Predicting Students Academic Performance Using Education Data Mining", IJCSMC, Vol. 2, Issue. 7, July 2013

[7] M. Sukany et al,DataMining:performance improvement in education sector using classification & clustering algorithm,ICCCE ,April 2012

[8] Dr. Mohd Maqsood Ali, Jazan University, "Role Of Data Mining In Education Sector", IJCSMC, Vol. 2, Issue. 4, April 2013.

[9] An Introduction to the WEKA Data Mining System Zdravko Markov Central Connecticut State University

AUTHORS

Prof Shaik.Abdul Nabi[2] is the Head of the Dept. of CSE, AVN Inst.Of Engg.& Tech, Hyderabad, AP, India. He completed his B.E (Computer Science) from Osmania University, A.P. He received his M.Tech. from JNTU Hyderabad campus and now he was submitted his Ph.D. thesis in the area of Web Mining from AcharyaNagarjuna University, Guntur, AP, India. He is a certified professional by Microsoft. His expertise areas are Data warehousing and Data Mining, Data Structures & UNIX Networking Programming.

P. Meena Kumari[1] is pursuing her Post Graduate in Master of Technology with specialization of Computer Science & Engg. at AVN Inst. of Engg.& Tech, Hyderabad, AP, India. Her interested research area is Data warehousing & Data Mining , Database Security and Data Structures .

Puppala Priyanka[3] is pursuing her Post Graduate in Master of Technology with specialization of Computer Science & Engg. at AVN Inst. of Engg.& Tech, Hyderabad, AP, India. Her interested research area is Data warehousing & Data Mining, Network Security and Data Structures.