Educational Data Mining techniques and their applications

Conference Paper · October 2015
DOI: 10.1109/ICGCIoT.2015.7380675

4 authors, including:

Shubha Puthran
Narsee Monjee Institute of Management Studies
6 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Educational Data Mining Techniques and their Applications

John Jacob
Computer Engineering Department
MPSTME, NMIMS
Mumbai, India
johnjacob.nmims@gmail.com

Kavya Jha
Computer Engineering Department
MPSTME, NMIMS
Mumbai, India
kavyajha.nmims@gmail.com

Paarth Kotak
Computer Engineering Department
MPSTME, NMIMS
Mumbai, India
paarthkotak.nmims@gmail.com

Shubha Puthran
Assistant Professor
Computer Engineering Department
MPSTME, NMIMS
Mumbai, India
shubha.puthran@nmims.edu

***Abstract*— Educational Data Mining (EDM) is a learning science, and an emerging discipline, concerned with analyzing and studying data from academic databases. Through the exploration of these large datasets, using various data mining methods, one can identify unique patterns which will help study, predict and improve a student's academic performance. This paper elaborates a study on various Educational Data Mining techniques and how they could be used for the benefit of all the stakeholders in the educational system. Correlation is used to see if a variation in one variable results in a variation in the other. Decision trees give possible outcomes and are used to predict students' performance in this study. Regression analysis is used in the construction of a model involving a dependent variable and multiple independent variables; if the model is satisfactory, then the value of dependent variable is determined using the values of the independent variables. Clustering finds groups of objects so that objects that are in a cluster are more like each other than to objects in another cluster, helping in arranging items under consideration; clustering would help in analyzing the job profiles that would be suited for each student.**

***Keywords*— Educational Data Mining, Cluster Analysis, Classification, Regression Model, ID3, J48, C4.5, K-Means.**

I. INTRODUCTION

Data Mining is a domain of computer science and the analysis step of the "Knowledge Discovery in Databases" process, or KDD. It is the process of recognizing patterns in huge data sets.

Educational Data Mining (EDM) is a sub-domain of Data Mining which deals with data from academic databases which is used to develop various techniques and to recognize patterns that are unique [1] [2]. The obtained knowledge can then be used to offer suggestions to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance, to decrease failure rates, to understand students' behavior in a better way, to assist instructors, to improve teaching, and to construct regression models and decision trees to predict student performance in terms of their grades or percentage [3]. EDM methods may also be used to categorize the students who require support, to analyze students' learning and cluster them according to their strengths and weaknesses for placement related activities.

A. Abbreviations and Acronyms

TABLE I. ABBREVIATIONS AND DESCRIPTION

Abbreviation	Description
EDM	Educational Data Mining
NFC	Near Field Communication
OTP	One Time Password
PPMC	Pearson Product Moment Correlation
SMS	Short Message Service
GPA	Grade Point Average
CGPA	Cumulative Grade Point Average
RFID	Radio-Frequency Identification

II. RESULTS AND ANALYIS

A. Study on whether Attendance is Statistically Significant for the Performance of Students

[11] [12] explain the various ways in which performance of students can be studied, one of the ways being Correlation. Pearson Product Moment Correlation (PPMC) is used to test for a relationship between two numerical variables. It represents the linear relationship between the two sets of data. Pearson Correlation are represented by the Greek letter rho (ρ) for a given population and the letter "r" for a sample under

consideration. The following formula is used to calculate PPMC:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Right at the inception, a scatter-plot should be made and studied which would unveil the possibility of a relationship between the two variables. The value of the Pearson correlation coefficient, R, ranges from values +1 to -1. A value closer to +1 would indicate a stronger relationship. We can determine the category of correlation by seeing what effect one variable's increment in value has on the other. The expected categories would be:

- Positive correlation: the value of the variable naturally increases (R>0)
- Negative correlation: the value of the variable naturally decreases (R<0)
- No correlation: the value of the variable stays constant (R=0)

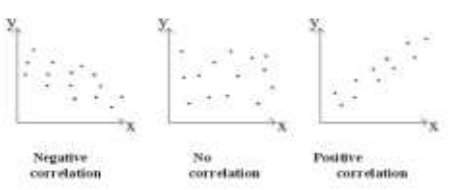

Fig. 1. Scatter-Plot for the three categories of Pearson Correlation

TABLE II. R-VALUE AND ITS MEANING

R-Value	Meaning
0.0-0.3	A weak uphill (positive) linear relationship between variables.
0.3-0.5	A Moderate uphill (positive) relationship between variables.
0.5-0.7	A Moderate - Strong uphill (positive) relationship between variables.
0.7-1.0	Extremely Strong uphill (positive) relationship.

In our experiment, after performing Pearson Correlation, i.e. > cor(CGPA,ATTENDANCE), we got 0.6035177 as the R-Value. Reading from Table II, 0.6 tells us that there is a moderate-strong relationship between the values of Attendance and CGPA, indicating that Attendance does, in fact, impact the CGPA of a student as students with lower CGPA have lower attendance and vice versa.

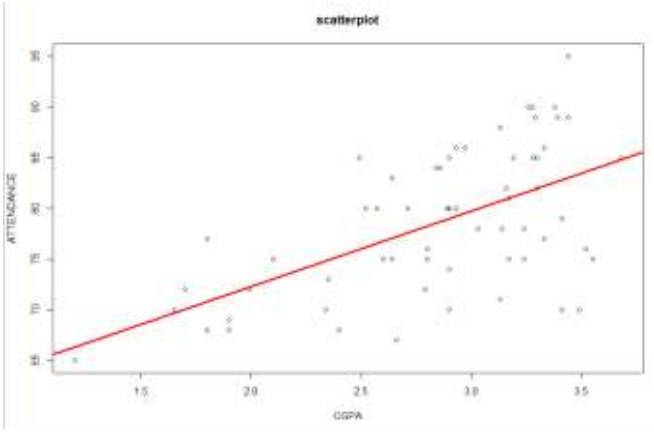

Fig. 2. Scatter-Plot for the Data Set under consideration

B. *Prediction of Student Performance*

For higher education institutions whose goal is to contribute to the improvement of quality of higher education, Human capital creation is assessed continuously. This makes prediction of students' success crucial for higher education institutions, to verify that the quality of teaching process is sufficiently meeting students' needs When the task of converting raw educational data into knowledge is carried out, the gratification of all stakeholders is ensured: students, professors or teachers, supporting management and the community and society of which we are a part of.

Prediction focuses on the estimation of the value of the variable describing the student, which is not known. This estimated value can be a numerical/categorical value. In our research, Classification using decision trees and Prediction using Multiple Regression has been used. M. Nasiri [13] justifies how GPA can be predicted using Classification Algorithms.- We aim to predict the Final Result (PASS or FAIL) of students in our university using decision trees and Predict Semester End GPA using Multiple Regression to help all stakeholders as mentioned above like Students, Professors, Administration, Supporting administration.

a. *ID3 (J48 Algorithm) to Predict Students Academic Failure or Dropouts*

A decision is a tree structure that follows a sequential path and, has nodes to help make a logical decision. The branches of a decision tree are used to illustrate the entire process and to show the result. The branches lead to the leaf nodes; leaf nodes are the final decision nodes at the end with no predicates.

The decision tree algorithm boils down to dividing leaves that are similar and dissimilar. The algorithm should reach a point where no further division is possible.

TABLE III. PARAMETERS FOR ID3 ALGORITHM

1.	CGPA
2.	Attendance
3.	Class Test Scores
4.	Backlogs
5.	Year Drops
6.	10th Grade Marks
7.	12th Grade Marks
8.	Board of Study
9.	Result (PASS/FAIL)

The algorithmic steps for decision tree algorithm are mentioned as follows.
Step 1: Let the data set be S.
Step 2: Make the continuous values discrete.
Step 3: Incorporate all of S in single tree node.
Step 4: If S is homogenous, then terminate.
Step 5: Divide the non-homogenous node by picking an independent attribute that best splits the node. Split that node based on the values of this chosen attribute.
Step 6: Terminate when the set becomes similar, otherwise continue splitting.

Once data was collected and cleaned, classification task is carried out using J48. S. Singhal [14] presented the steps of how to use WEKA tool for such technologies. It provides the facility to classify the data through various algorithms. The accuracy achieved is around 95%. The kappa statistic measures the correctness of prediction with the true class where a value of 1.0 signifies complete correctness or a complete match.

The classification rules extracted from the tree are:

- If Backlogs > 3, Result: FAIL (7.0)
- If Backlogs >=3 AND YearDrop > 0 AND CT <= 8, Result: FAIL (4.0)
- If Backlogs <=3 AND CT <=9, Result: FAIL(6.0/2.0)

Tree has another branch which represents following rules:

- If Backlogs <= 2: PASS (39.0)
- If Backlogs <=3 AND CT >9, Result: PASS (3.0)

Thus, based on these Classification rules, we can identify the students who are academically weak and need remedial classes or any other help in order to keep them from failing or dropping a year.

b. Multiple Regression to Predict Students GPA

[15] Regression is a numerical evaluation process that predicts the students' performance based on the already acquired data set. The student performance is given a numerical value Y and is predicted using the known values ranging from X_1 to X_k. In this case, X_1, X_2… are the parameters available to us in the educational database from previous semester performance of students. Y is the CGPA for which we are creating a prediction model. The Regression Model is built with the available data to predict CGPA from other parameters in the educational database. The data of Computer Engineering Students of the University was collected and a Multiple Regression Analysis was performed.

TABLE IV. PARAMETERS FOR REGRESSION MODEL

1.	CGPA (0.0 - 4.0)
2.	Lab Grade (LG) (1-10)
3.	Class Test Score (CT) (1-15)
4.	Assignment Score (A_Score) (1-10)
5.	Presentation Score (PPT) (1-10)
6.	Attendance (Att) (in %)
7.	Semester 6 Score (Sem 6) (0-100)
8.	Semester 7 Score (Sem 7) (0-100)
9.	Average of Semester Scores (Avg) (0-100)

TABLE V. SUMMARY OF REGRESSION MODEL

Coefficients							
Intercept	Sem 6	LG	PPT	CT	A_Score	Att	Avg
-0.93	-0.04	0.01	-0.07	0.13	0.07	0.01	0.04
Call: lm(formula = Sem 7 ~ Sem 6 + LG + PPT + CT + A_Score + Att + Avg)							
Residuals							
Min	1Q	Median	3Q	Max			
0.54	0.09	0.04	0.12	0.35			
Residual standard error: 0.21 on 49 degrees of freedom Multiple R-squared: 0.86, Adjusted R-squared: 0.83 F-statistic: 41.4 on 7 and 49 DF, p-value: < 2.2e-16							

R2 ranges from 0 to 1, the closeness of R2 to 1 indicates the goodness of the prediction.

According to the Model that we have constructed, the R2 value was found to be 0.85, which is very close to 1, thus verifying its accuracy in predicting GPA's of students. The model was put to test with the marks of the next semester as input. The predictions were found to be correct 80% of the time when compared with actual results of the next semester with an error of +/- 0.2 CGPA as acceptable error. 20% of the cases, the difference in predicted CGPA and Actual CGPA attained was greater than 0.2 CGPA.

The Prediction Models help students predict their GPAs and have a better understanding and idea of what they need to do and how much more effort they need to put in to achieve their GPA goals.

C. *Grouping students For Placements (Clustering based on Role Suitability)*

Clustering is the grouping of data based on their similarities and dissimilarities. Objects within one cluster need to be similar while the objects from two different clusters are expected to be different from each other. While analyzing the clusters, the data set is divided into categories based on the some attribute similarity and then a unique label is assigned to that cluster. The main benefit of using clustering is that it is dynamic and it adapts to changes quickly. Also, it easily aggrandizes the attribute on which the clusters were made.

Our objective was to cluster students according to their strengths, weaknesses, interests and traits, etc. Then, the clusters of students obtained would be used by the Placement team to gain insight into grouping these students into their prospective job profiles based on the attributes selected. For example, students who are proficient in coding would fall in a cluster of students who have higher marks in subjects related to Programming Languages; these students could be recommended to take up a Software Engineering or a Developer profile. We have used K-Means Clustering technique to recognize clusters of similar students on the basis of their strengths [16].

TABLE VI. CLUSTERING SUMMARY

K-means clustering with 4 clusters of sizes 17, 12, 13, 15							
Cluster means:							
	Comm Skill	Soft Engg.	QA Testing	Programming	Data Structure	DBMS	Data Mining
1	78.92	70.53	71.71	62.24	60.76	55.82	58.53
2	78.92	63.46	77.85	61.15	67.38	65.69	75.85
3	85.92	78.50	76.67	71.92	69.42	72.00	71.17
4	77.47	71.93	72.13	80.60	80.60	72.60	69.33
Clustering vector:							
[1] 4 3 2 2 4 2 4 2 1 4 3 1 1 2 2 3 4 4 4 3 4 4 4 4 4 3 3 1 2 2 4 2 1 3 3 1 3 3 [39] 1 1 3 1 2 2 4 3 4 2 1 2 1 1 1 1 1 1 1							
Within cluster sum of squares by cluster:							
[1] 4944.35 3368.08 4167.08 3888.53 (between_SS / total_SS = 48.6 %)							

In the above Cluster Graph, we observe four different clusters and these four clusters together include all the students of the class. Analysis of these clusters indicate the following:

- Students in cluster 1 are weak In DBMS and Data Mining and perhaps should not be considered for these roles. As their QA Scores are their strength, they should be considered for Testing Roles.
- Students in Cluster 2 are the best at Data Mining and QA and these students could be considered for MIS / Data Mining or Quality Assurance Profiles.
- Students in Cluster 3 are excellent in Communication Skills and have good understanding of the Software Development process. They could be thus, considered for Business Analyst or Project Manager Roles.
- Students in Cluster 4 are undoubtedly the best at coding with good understanding of Data Structures and Programming. They could be offered the role of Software Engineers or Developers.

III. CONCLUSION

Data Mining techniques like Regression and Decision Trees to predict academic performance are studied and executed and are found to effectively predict student performance and also, to predict academic failure. Clustering is successfully used to group the students into clusters according to their academic strengths and weaknesses. These methods will greatly help the university teachers to know what changes need to be made, provide remedial courses to weak students, identify weak students at risk of failure or year drops and to make learning a better experience for their students; and it would also help the students and the placement committee in getting to know which job profiles they could apply to on the basis of their skill set.

Along with technological advancements are costs and challenges associated with implementing Educational Data Mining applications. These include the costs to store logged data and the cost associated with hiring staff dedicated to managing data systems. Moreover, data systems may not always integrate seamlessly with one another and even with the support of statistical and visualization tools, creating one simplified version of the data can be difficult. Furthermore, choosing which data to mine and analyze can also be challenging, making the initial stages very time consuming and labor-intensive. From beginning to end, the EDM strategy and implementation require one to uphold privacy and ethics for all stakeholders involved.

ACKNOWLEDGMENT

For this project, we would like to thank our families for their continuous support and faith in us.

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] Umamaheswari. K, and S. Niraimathi "A Study on Student Data Analysis Using Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, ISSN: 2277 128X, August 2013.*

[3] E. Osmanbegović, and M. Suljić "Data Mining Approach for Predicting Student Performance", *Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2011.*

[4] S. L. Prabha, Dr.A.R.M. Shanavas "Educational Data Mining Applications", *Operations Research and Applications: An International Journal (ORAJ), Vol. 1, No. 1, August 2014.*

[5] R. R. Kabra "Performance Prediction of Engineering Students using Decision Trees", *International Journal of Computer Applications (0975 – 8887) Volume 36– No.11, December 2011.*

[6] M. Durairaj, C. Vijitha "Educational Data mining for Prediction of Student Performance Using Clustering Algorithms", *International Journal of Computer Science and Information Technologies, Vol. 5 (4).*

[7] W. Hämäläinen and M. Vinni. "Comparison of Machine Learning Methods for Intelligent Tutoring Systems", *8th International Conference on Intelligent Tutoring Systems, ITS 2006, volume 4053, pages 525–534, Springer, 2006.*

[8] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R.Zaıane "Clustering and Sequential Pattern Mining of Online Collaborative Learning Data", *IEEE Transactions on Knowledge and Data Engineering, 21(6):759–772, 2009.*

[9] O. R. Zaıane "Web Usage Mining for a better Web-based Learning Environment", *Conference on Advanced Technology for Education, Banff, Alberta, pages 60–64, 2001.*

[10] R. K. Arora, D. Badal "Subject Distribution Using Data Mining", *International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.*

[11] L. Ahmedi, E. Bytyçi, B. Rexha, V. Raça, "Applying data mining to compare predicted and real success of secondary school students", *Advances in Applied Information Science, Kosovo.*

[12] E. Ogor "Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques", *IEEE Computer Society, IEEE, 2007.*

[13] M. Nasiri, F. Vafaei, and B. Minaei "Predicting GPA and Academic Dismissal in LMS Using Educational Data Mining: A Case Mining", *6th National and 3rd International conference of e-Learning and e-Teaching (ICELET2012), IEEE, 2012.*

[14] S. Singhal, "A Study on WEKA Tool for Data Pre-processing, Classification and Clustering", *International Journal of Innovative Technology and Exploring Engineering* (IJITEE), *Vol. 2(6), May 2013.*

[15] R. S. Bichkar "Predicting Students Academic Performance Using Education Data Mining", *World Journal of Computer Application and Technology 2(2): 43-47, 2014.*

[16] Md. H. Shovon, and M. Haque "Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 7, ISSN: 2277 128X, July 2012.*