IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 8 Issue 2, February 2021
ISSN (Online) 2348 – 7968 | Impact Factor (2020) – 6.72
www.ijiset.com

A Review of Education Data Mining to Predict Students' Performances

Dr. Himanshu Maniar
Assistant Professor - Shri C U Shah College Of Commerce, Management and Computer Education – Surendranagar, Gujarat, INDIA
collogehimanshu@gmail.com

Abstract.
In recent years, with the fastest growing of computerization, every field is changed drastically as far as the management of data is concerned. In today's era of competition, education field is one of the most important fields which people always analyze critically. From large scale universities to the small scale schools, every educational institution has started maintaining data in computerized form. Such data analytics are often used while seeking admissions, placements or for performance improvements. Various researchers have proposed solutions to improve students' performances in systematic way. Considering the large set of research activities in this domain, a new sub domain called Education Dta Mining is introduced. This paper discusses some of the recent proposals of education data mining in detail. The paper concludes with some future directions.

Keywords: Education Data Mining, Students' Performance, Classification, Decision Tree

1 Introduction

Data mining is a field of computer science which is based on retrieving information from the data in the form of hidden patterns. Such information is not directly stored in the database but could be generated by applying data mining algorithms. Data mining techniques help organizations to see insight from their data. Such techniques help to find out those hidden truth which are not directly stored. Other than of retrieval of information, Data mining techniques help in classification, prediction, data cleaning, data transformation etc. Data mining techniques are used in various fields like medical science, educational data analysis, business analysis, market analysis, web analysis, network traffic analysis, trend analysis etc. Data mining along with data warehousing helps an organization to view its data with various levels in terms of dimensions. This chapter introduces the basis of data mining and how it is useful with the education data mining [1-10]

Data mining techniques used for analysis of educational data is called educational data mining.

Educational institutes maintain a large amount of data to keep records of students, faculties and courses. This data has student's personal and academic information, faculties' personal and academic information, syllabus, question papers, circulars etc. Educational data

mining is started being adopted by various universities and individual institutions for betterment of their students and faculties. These techniques are implemented with their application programs to be compatible with their databases.

Students' performance is one of the most important criteria for any institute. Students' performance can be predicted according to their previous academic performances. Further to the analysis, students' skills and interests can be related with their performances. Such analysis helps teachers to pay more attention on the weak students.

This paper discusses some of the recent education data mining related approaches for students performance predictions using classification techniques.

2 Classification with Data Mining

Classification is a process of predicting values of one or more categorical attributes given a set of values corresponding to rest of the categorical or numerical attributes. For example, a student's grade in current semester is predicted based on their previous semester performances. A bank analyzes loan holders' payment history to identify unsafe customers. A mall classifies their customers into various categories (high profile, middle profile and low profile or vegetarian or non-vegetarian) [2-6].

Classification is a process of finding values of one or more unknown categorical variables. The purpose of classification is to build a model based on existing data which helps in classification of new data. Classification model construction can be done using various algorithms. This section explains a few of the most widely used classification algorithms with examples. The results of association rule mining can also be useful to decide track of classification. The main goal here is to find the relationships among a set of variables to determine the class label for new data. Various classification techniques such as decision tree algorithms, naive bayes, neural network are proposed. The selection of a method depends upon various parameters such as required speed, accuracy, efficiency, effectiveness of classification. İt also depends on do we need a shorter classification time or longer classification time. Various algorithms may differ in terms of how they react to the limitations of the datasets such as incompleteness, errors etc. Due to limited scope of a paper, in this paper the selection criterias and these methods are not discussed in detail [7-10].

3 Recent Proposed Solutions

Three of the recent proposed solutions are discussed here along with their merits and limitations. These directives will help the researchers in doing research with these solutions.

3.1 Proposal – 1

Dr. Varun Kumar and Anupama Chadha have analyzed the requirements of using data mining techniques in education specially in higher education [11].

Methodology [11]:-

This proposal starts with explaining how data mining techniques work and how they can be used to analyze educational data with its classification and clustering techniques. Along with prediction, this proposal explains use of association rule mining in education data mining too.

Purpose[11].**:-**

1. To analyze subjects and to find out related subjects this reduces search space and improves efficiency.
2. To predict students' performances in a specific subject based on their performances in related subject derived from (1).
3. To detect cheating during online examinations.
4. To detect any abnormal or error in results.

Limitations:-

This proposal shows the possible ways data mining techniques can be used in education data mining. The purposes discussed in this proposal need to be implemented.

3.2 Proposal –2

Paul Baepler and Cynthia James Murdoch have analyzed requirements of using data mining techniques for academic analytics – a form of education data mining[12].

Methodology[12].**:-**

This proposal shows various tasks which can be done more efficiently using data mining techniques. Three main categories are identified: Academic Analytics, Data Mining in Higher Education and Course Management System Auditing.

Purpose[12].**:-**

1. Academic Analytics refers to predicting performances of students' with algorithms called signal algorithms – to identify performance issues of students and teachers.
2. Data Mining in Higher Education refers to the activities to analyze non academic record like to find out students with similar interests, students resources utilization analysis, misuse and cheating done by students etc.
3. Course Management System Auditing refers to the analysis of a course and its usage in real life. This type of analysis helps to determine whether course content revision is required or not.

Limitations:-

This proposal shows the possible ways data mining techniques can be used in education data mining. The purposes discussed in this proposal need to be implemented.

3.3 Proposal – 3

Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer and William F. Punch have proposed a solution to analyze resources utilization and performance of students[13].

Purpose[13]**:-**

1. To find out students groups who use the resources in similar way. Using this information, we can motivate a student to use resources in a better way.
2. To find out what kind of problems are solvable by students. Using this information, we can guide teachers to prepare assignments and practicals in an effective and efficient way.

Database[13]**:-**

This proposal is based on using Learning Online Network with Computer-Assisted Personalized Approach (LON-CAPA) based large database. The database is composed of two types of data.

1. Educational resources like examinations, term works, home works, assignments, quiz, material, web resources etc.
2. Resource utilization data like how many resources are actually utilized for how much duration by various students.

Michigan State University offers physics course in engineering in spring semester. This course is based on 12 home work assignments with total 184 unsolved problems. Students have to submit their work through online portal only. Students may opt out a subject if not able to complete assignments. In 2002 batch, 261 students' enrolled and later on 227 students completed the course whose database is used[13].

Students' performances are evaluated at three levels. At 1st level, every student's grade is assigned in a rank of 0 to 9. At 2nd level, every student's grade is assigned in {High,Middle,Low} and at 3rd level, every student's grade is assigned in {Passed,Failed} as shown in Table 2.1[13].

TABLE 1
SELECTING 9 CLASS LABELS REGARDING TO STUDENTS' GRADES

Class	Grade	Student #	Percentage
1	0.0	2	0.9%
2	0.5	0	0.0%
3	1.0	10	4.4%
4	1.5	28	12.4%
5	2.0	23	10.1%
6	2.5	43	18.9%
7	3.0	52	22.9%
8	3.5	41	18.0%
9	4.0	28	12.4%

TABLE 2
SELECTING 3 CLASS LABELS REGARDING TO STUDENTS' GRADES

Class	Grade	Student #	Percentage
High	Grade >= 3.5	69	30.40%
Middle	2.0 < Grade < 3.5	95	41.80%
Low	Grade <= 2.0	63	27.80%

TABLE 3
SELECTING 2 CLASS LABELS REGARDING TO STUDENTS' GRADES

Class	Grade	Student #	Percentage
Passed	Grade > 2.0	164	72.2%
Failed	Grade <= 2.0	63	27.80%

Table 2.1 – Students' Grade Structure [13]

Features [13]**:-**

From the above database, various features are extracted as below.

1. Student's success ratio as per the number of correct answers.
2. Student's success ratio trial base. Those students who could solve problems correctly vs. those students who could solve problems with multiple trials.
3. Total number of trials a student took to complete home work.
4. Time required by a student to complete a problem correctly.
5. Time required by a student to complete a problem correctly or incorrectly.
6. Amount of participation a student does to perform group activities. Analysis of whether a student works independently or in a group.
7. How much resources a student uses before starting solving a problem. Analysis of whether a student accesses more resources prior to start problem solving or a student access more resources if he fails to solve a problem.
8. How much resources a student uses between submitting solutions of two solutions. Whether a student shows hurry in submitting problems all together without referring resources properly or not.
9. How a student reacts to difficult problems. Whether a student tries to solve it up to the time limit or quit after a few failed trials.
10. How much time a student spent online to solve all problems.

The most important and relevant features are listed in Table 2.2[13]

FEATURE IMPORTANCE IN 3-CLASSES USING ENTROPY CRITERION

Feature	Importance %
Total_Correct _Answers	100.00
Total_Number_of_Tries	58.61
First_Got_Correct	27.70
Time_Spent_to_Solve	24.60
Total_Time_Spent	24.47
Communication	9.21

Table 2.2 – Important Features[13]

Classification:-

Decision Tree, k-nearest neighbor, Quadratic Bayesian classifier, Parzen window and multilayer perceptron methods of classification are used for analysis purpose. The main benefit of using multiple classifiers is to find out accuracy of the classification. CMC – Combination of Multiple Classifiers based testing is done. There two ways we can implement CMC: Offline and Online. In offline CMC, out of n classifiers the one with the highest accuracy is selected to classify new data. In online CMC, all n classifiers classify new data and then the class label with highest votes is considered as the class label of new data. Further to enhance the efficiency, proposal has been implemented using GA – Genetic Algorithms. GA can be implemented in two ways with data mining. Either directly in a classifier or after classification, to smooth out accuracy and to set parameters. The details are given in the paper. Table 1 shows the results of various classifiers with respect to various levels of class attribute. Table 2 shows the results of CMC without GA vs CMC with GA[13].

COMPARING THE ERROR RATE OF ALL CLASSIFIERS ON PHY183 DATASET IN THE CASES OF 2-CLASSES, 3-CLASSESS, AND 9-CLASSES, USING 10-FOLD CROSS VALIDATION, WITHOUT GA

		Performance %		
Classifier		2-Classes	3-Classes	9-Classes
Tree Classifier	C5.0	80.3	56.8	25.6
	CART	81.5	**59.9**	**33.1**
	QUEST	80.5	57.1	20.0
	CRUISE	81.0	54.9	22.9
Non-tree Classifier	Bayes	76.4	48.6	23.0
	1NN	76.8	50.5	29.0
	kNN	**82.3**	50.4	28.5
	Parzen	75.0	48.1	21.5
	MLP	79.5	50.9	-
	CMC	**86.8**	**70.9**	**51.0**

Table 1 – Results of Various Classifiers [13]

COMPARING THE CMC PERFORMANCE ON PHY183 DATASET USING GA AND WITHOUT GA IN THE CASES OF 2-CLASSES, 3-CLASSESS, AND 9-CLASSES, 95% CONFIDENCE INTERVAL.

	Performance %		
Classifier	2-Classes	3-Classes	9-Classes
CMC of 4 Classifiers without GA	**83.87 ± 1.73**	**61.86 ± 2.16**	**49.74 ± 1.86**
GA Optimized CMC, Mean individual	**94.09 ± 2.84**	**72.13 ± 0.39**	**62.25 ± 0.63**
Improvement	**10.22 ± 1.92**	**10.26 ± 1.84**	**12.51 ± 1.75**

Table 2 – Results of CMC with / without GA [13]

Limitations:-

This proposal shows mining of resource utilization by students and quality of problems given to the students. There is no information given by teachers is processed to analyze students performances.

2 Future Work

This research work is based on predicting student performance using classification where the predictions are classes in terms of grades. A more complex system could be designed for the prediction of performances in terms of marks. This research work has designed a database including student's assessment based information, skill based information and teacher's view.

This system could be extended to include student's view. Care must be taken to get genuine information from the students about what they think of their future performances. At present, predictions are done independently for every set of dependent subjects. Prediction of one subject will not affect prediction of other subject if both subjects are not dependent on each other. Further to the extension, some dependencies could be identified for finding final result of the entire course. Such concept could be extended to predict performances in competitive exams and placement related activities. More complex classification algorithms could be used with the support of Bigdata analytics to train classifier over a huge data.

References

1. Weiss, Sholom M., and Nitin Indurkhya. Predictive data mining: a practical guide. Morgan Kaufmann Publishers, 1998
2. Jiawei Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001
3. Simoff, Simeon J., and Graham J. Williams. Data Mining. Springer-Verlag Berlin/Heidelberg, 2006.
4. Lawrence, Kenneth D., et al. Data mining methods and applications. Auerbach Publications, 2008.
5. Cao, Longbing. Data mining and multi-Agent integration. Springer, 2009.
6. Abraham, Ajith. Data mining. Springer, 2009.
7. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
8. Romero, Cristobal. Handbook of educational data mining. CRC Press, 2011
9. Aggarwal, Charu C. Data mining: the textbook. Springer, 2016.
10. Roiger, Richard J. Data mining: a tutorial-Based primer. Taylor & Francis, CRC Press, 2017
11. Kumar, Varun, and Anupama Chadha. "An empirical study of the applications of data mining techniques in higher education." International Journal of Advanced Computer Science and Applications 2.3 (2011): 80-84.
12. Baepler, Paul, and Cynthia James Murdoch. "Academic analytics and data mining in higher education." International Journal for the Scholarship of Teaching and Learning 4.2 (2010):
13. Bidgoli, B. Minaei, et al. "Predicting student performance: An Application of data mining methods with the educational web-based system LON-CAPA." Proceedings of ASEE/IEEE frontiers in education conference. 2003.