Review on Data Mining Techniques Used For Educational System

Agrawal Bhawana D¹, Gurav Bharti B²

¹M. Tech Scholar, ²Assistant Professor, Department of Computer Science & Engineering, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India

Abstract— Indian Government showing great interest in failure and dropout of children at different educational level from last decade. Great deal of research has been done on identifying the factors that affect the low performance of student at different educational levels. Data mining methods represent a valid approach for the extraction of precious information from existing Database. Educational Data Mining (EDM) deals with developing methods for discovering knowledge from data that come from educational domain. This paper is based on survey which proposes to apply data mining techniques such as association rule mining, classification techniques.

Keywords— Educational Data Mining (EDM), Data Mining, Dropout, Prediction, School Failure, Classification.

I. INTRODUCTION

From last few years many countries facing problem of school failure and drop outs, so Governments of many countries showing their interest in determination of its main contributing factors. Large amount of information that current computer can store in Databases is a “Gold Mine” of valuable information about students. Determination of its main contributing factors from this large amount of data is known as the “the one hundred factors problem” and a great deal of research has been done on identifying the factors that affect the low performance of students (school failure and dropout) at different educational levels (primary, secondary and higher) as described by Araque et al.,2009[1].

The solution to this problem is called Educational Data Mining (EDM). EDM concerned with developing methods that extract knowledge from data come from the educational context [2]. This new area of research focuses on the development of methods to better understand students and the settings in which they learn. This study proposes to predict student failure by using Data Mining. Data Mining is a process of extracting useful knowledge and information from data stored in databases and data warehouses. Data Mining is an integral part of KDD (Knowledge Discovery in Database) [3].

KDD can be used to learn the model for the learning process or student modeling. In fact, we want to detect the factors that most influence student failure in young students by using classification techniques and experiments taken to improve the accuracy for predicting which student might fail or dropout by first using all available attributes, then selecting best attribute and Finally rebalancing data using cost sensitive classification. Also we use different Data Mining techniques because data have high dimensionality (there are many factors that can influence) and often highly unbalanced (the majority of students pass and too few fail).

The final objective is to detect the students as early as possible who show these factors so that we can provide some type of assistance for trying to avoid or reduce school failure.

The paper is organized as follows section II describes Educational Data Mining process. The section III describes Algorithms for EDM. The section IV describes related work and finally section V describes conclusion.

II. EDUCATIONAL DATA MINING PROCESS

Data mining techniques can discover useful information that can be used in formative evaluation to assist educators establish a pedagogical basis for decisions when designing or modifying an environment or teaching approach. The application of data mining in educational systems is an iterative cycle of hypothesis formation, testing, and refinement (see Fig. 1) [3]. As we can see in Fig. 1, educators and academics responsible are in charge of designing, planning, building and maintaining the educational systems. Students use and interact with them. Starting from all the available information about courses, students, usage and interaction, different data mining techniques can be applied in order to discover useful knowledge that helps to improve the e-learning process. The discovered knowledge can be used not only by providers (educators) but also by own users (students).
Simply stated, data mining 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. The sequences of steps identified in extracting knowledge from data are shown in Figure 2.

The method proposed for predicting the academic failure of students belongs to the process of Knowledge Discovery and Data Mining (see Fig. 3)[4].

The main stages of the method are:

1) **Data gathering**: This stage consists in gathering all available information on students. To do this, the set of factors that can affect the students’ performance must be identified and collected from the different sources of data available. Finally, all the information should be integrated into a dataset as shown in fig 4.

2) **Pre-processing**: At this stage the dataset is prepared to apply the data mining techniques. To do this, traditional pre-processing methods such as data cleaning, transformation of variables, and data partitioning have to be applied. Other techniques such as the selection of attributes and the re-balancing of data have also been applied (Fig. 5) in order to solve the problems of high dimensionality and imbalanced data that are typically presented in these datasets.

3) **Data mining**: At this stage, DM algorithms are applied to predict student failure like a classification problem. To do this task, we propose to use classification algorithms based on rules and decision trees. Decision tree algorithms are like J48, AD Tree, C4.5, Random Tree etc. These are “white box” techniques that generate easily interpretable models. In addition, a cost sensitive classification approach is also used in order to solve the imbalanced data problem. Finally, different algorithms have been executed, evaluated and compared in order to determine which one obtains the best results.
4) **Interpretation:** At this stage, the obtained models are analyzed to detect student failure. To achieve this, the factors that appear (in the rules and decision trees) and how they are related are considered and interpreted.

III. **ALGORITHMS FOR EDUCATIONAL DATA MINING**

As number of schools and students is increasing day by day, we can use data mining techniques and algorithms for improving education standard so that reducing numbers of dropouts and maximizing education system efficiency. Following algorithms used in education mining:

**A) AD Tree:** An AD tree that is alternating decision tree contains two nodes namely decision node and prediction node. Predicate condition is specified by decision node and Prediction nodes contain a single number. Alternating Decision Tree has prediction node as both root and leave. Alternating Decision Tree classifies the instance by following all paths for which all decision nodes are true and summing any prediction nodes that are visited.

1 if (precondition)
2      if (condition)
3          return score_one
4      else
5          return score_two
6      end if
7    else
8        return 0
9    end if

**B) J-48:** J-48 is used when we have a dataset with a list of predictors and a list of targets or we can say J-48 is used when we have list of independent variables and dependent variables. Then we can apply a decision tree like J48 on that dataset. It would allow you to predict the target variable of a new dataset record.

**C) JRiP:** JRiP algorithm is based on association rules with reduced error pruning (REP). Reduced error pruning (REP) is a very common and effective technique in decision tree algorithms. In Reduced error pruning for rules algorithms, the training data is divided into two parts namely a growing set and a pruning set. Firstly using some heuristic method an initial rule set is formed that over the growing set, then this large rule set is then repeatedly simplified by using one of a set of pruning operators. At each stage of simplification, the pruning operator is chosen such that it yields the greatest reduction of error on the pruning set. And finally simplification terminates by applying any pruning operator that would increase error on the pruning set.

**D) OneR:** OneR is short for "One Rule". This is very simple and accurate classification algorithm. OneR algorithm generates one rule for each predictor in the data, and then selects the rule as its "one rule" which has the smallest number of total error. We construct a frequency table for each predictor to create a rule for a predictor against the target. OneR produces rules only quite less accurate than state-of-the-art classification algorithms, but OneR produces rules that are simple for humans to interpret.

OneR Algorithm

For each predictor,

- For each value of that predictor, make a rule as follows:
  - Count how often each value of target (class) appears
  - Find the most frequent class
  - Make the rule assign that class to this value of the predictor

Calculate the total error of the rules of each predictor

Choose the predictor with the smallest total error.

IV. **RELATED WORK**

Although EDM is a recent research field there are many works in this area. Considerable amounts of EDM work are published at the peer-reviewed International Conference on Educational Data Mining, organized by the International Educational Data Mining Society in UK, USA etc.

EDM has emerged as an independent research area in recent years with establishment of annual international conference of EDM and journal of EDM.

C. Romero and S. Ventura [3] provides a survey of EDM from 1995 to 2005. It describes the need for analyzing the student data which can be used by students, educators and administrators.
It also describes that the application of data mining in Education system has specific requirements not present in other domain. They concluded that educational data mining is young area of research.

Khan [10] in 2005 conducted a study on 400 students containing 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

El-Halees [8] in 2008, gave a case study that used educational data mining to analyze students learning behavior. He used student’s data from database course and collected all available data including personal records and academic records of students, course records and data came from e-learning system. Then, he applied data mining techniques to discover many kinds of knowledge such as association rules and classification rules using decision tree.

Shannaq et al. and Rafael Y. [6] in 2010 use Classification as Data Mining technique to enhance the quality of the higher educational system. They applied the classification to predict the main attributes that may affect the students’ loyalty (number of enrolled students). The extracted classification rules are based on the decision tree; the extracted classification rules are studied and evaluated using different evaluation methods. It allows obtaining a deep understanding of student’s enrollment pattern in a University where the faculty and managerial decision makers can utilize any action to provide extra basic course skill classes and academic counseling. In addition the management system can improve their policy, enhance their strategies and thereby improve the quality of that management system.

Ayesha et al. Mustafa, Sattar and Khan [7] in 2010 used k-means clustering algorithm as a data mining technique to predict students learning activities including class quizzes, mid and final exam and assignments. This information will be given to the class teacher before the conduction of final exam. This study helps the teachers to reduce the dropout ratio by taking appropriate steps at right time and improve the performance of students. The information generated after the implementation of Data Mining techniques may be helpful for instructor as well as students.

Chandra and Nandhini [9] in 2010, uses the association rule mining analysis based on students failed courses to identify student’s failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity student’s performances. The extracted association rules shows some hidden patterns of students failed courses which could serve as a base for academic planners in making academic decision.

Bhardwaj and Pal [5] in 2011 conducted study on the student performance based by selecting 300 students from 5 different degree colleges. They applied the classification as data mining technique on 17 attribute, it was found that the factors like students grade in exam, living location, medium of teaching, mothers qualification, students other habit, family annual income and student's family status were correlated with the student academic performance. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.

Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto[11] in Feb 2013 uses 670 middle-school students from Zacatecas, México, and apply white-box classification methods, such as induction rules and decision trees. Then first they use experiments that attempt to improve their accuracy for predicting which students might fail or dropout by using all the available attributes; next, selecting the best attributes; and finally, rebalancing data and using cost sensitive classification. They face a complex problems such that data have high dimensionality (there are many factors that can influence) and often highly unbalanced (the majority of students pass and too few fail). The final objective of their study is to detect as early as possible the students who show these factors in order to provide some type of assistance for trying to avoid and/or reduce school failure.

This paper provides overall information about Educational data mining and Data mining techniques used such as classification techniques.

V. CONCLUSION

In this paper we studied newly emerging field of Educational Data Mining (EDM) and work done on this field up till now. As we have seen, predicting student failure at school can be a difficult task because it is a multifactor problem and also because the available data are normally imbalanced.

To resolve these problems, we can use KDD, different Data Mining algorithms and approaches for predicting student failure.
REFERENCES

[1] Araque F., Roldan C., Salguero A. Factors Influencin
[10] Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto(Fe