Data Mining in Education for Students Academic Performance: A Systematic Review

Er. Anurag Jindal, Er. Williamjeet Singh

Research Scholar, Computer Engineering, Punjabi University, Patiala, India
Assistant Professor, Computer Engineering, Punjabi University, Patiala, India

ABSTRACT
Currently there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community. Higher education, throughout the world is delivered through universities, colleges affiliated to various universities and some other recognized academic institutes. The main objective of higher education institutes is to provide quality education to its students. Indian education sector has a lot of data that can produce valuable information which can be used to increase the quality of education. Good prediction of student’s success in higher learning institution is one way to reach the higher level of quality in higher education system. In this paper we analyzed the potential use of data mining in education section and survey the most relevant work in this area. Data Mining can be used for dropout students, student’s academic performance, teacher’s performance and student’s complaints. As we know large amount of data is stored in educational database, so in order to get required data and to find the hidden relationship, different data mining techniques are developed & used. Various algorithms and data mining techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees (CART and CHIAD), Genetic algorithms, Nearest Neighbor method etc. are used for knowledge discovery from databases and helps in prediction of students academic performance. In future work we can apply different data mining techniques on an expanded data set with more distinct attributes to get more accurate results.

KEYWORDS
Educational Data Mining (EDM); Student’s performance; Data mining; Prediction; Data mining techniques.
1. INTRODUCTION

Every year, educational institutes admit students under various courses from different locations, educational background and with varying merit scores in entrance examinations. Moreover schools and junior colleges may be affiliated to different boards, each board having different subjects in their curricula and also different level of depths in their subjects. Analyzing the past performance of admitted students would provide a better perspective of the probable academic performance of students in the future. This can very well be achieved using the concepts of data mining. Student’s academic performance hinges on diverse factors like personal, socio-economic, psychological and other environmental variables. Prediction models that include all these variables are necessitated for the effective prediction of the performance of the students. The prediction of student performance with high accuracy is beneficial to identify the students with low academic achievements initially. The identified students can be individually assisted by the educators so that their performance is better in future.

Education is an essential element for the progress of country. Mining in educational environment is called Educational Data Mining. Educational data mining is concerned with developing new methods to discover knowledge from educational database. In order to analyze student trends & behavior towards education an attempt to study the present behavioral pattern of student in a cross section is a must. Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students and the settings which they learn in [7]. Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. As we know large amount of data is stored in educational database so in order to get required data & to find the hidden relationship, different data mining techniques are developed & used. There are varieties of popular data mining task within the educational data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. We can use the data mining in educational system as: predicting drop-out student, relationship between the student university entrance examination results and their success, predicting student’s academic performance, discovery of strongly related subjects in the undergraduate syllabi, knowledge discovery on academic achievement, classification of student’s performance in computer programming course according to learning style, investing the similarity & difference between schools.

This paper is organized in four sections: Section 1 covers the basic Introduction about education whereas Section 2 discusses the Review of Literature. Section 3 includes Methods, Educational Data Mining and various data mining techniques. The conclusion is drawn in the section 4.

2. LITERATURE SURVEY

A number of reviews pertaining to not only the diverse factors like personal, socio-economic, psychological and other environmental variables that influence the performance of students but also the models that have been used for the performance prediction are available in the literature and a few specific studies are listed below for reference.

The author in [1] analyses the influencing factors (courses) that contributes to the prediction of student’s academic performance. It was determined whether a first year student will graduate higher or lower than a second class upper. 2427 complete records of bachelor of computer science students admitted from 2000 to 2006 were collected. The author in [13] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal. The author in [17] applied five classification algorithms namely Decision Trees, Perceptron-based Learning, Bayesian Nets, Instance-Based Learning and Rule-learning to predict the performance of computer science students from distance learning stream of Hellenic Open University, Greece. The author in [15] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. The author in [25] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The author in [24] presented a case study on educational data mining to identify up to what extent the enrollment data can be used to predict student success. The author in [10] compares the accuracy of Decision Tree and Bayesian Network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes: Can Tho University (CTU), a large national university in Viet Nam and the Asian Institute of Technology (AIT), a small international postgraduate institute in Thailand that draws students from 86 different countries. Table1 shows a summary of the main objectives of the corresponding papers and also shows the source of collection of data of students.
Table 1: Summary of the Pertinent Literature

<table>
<thead>
<tr>
<th>Author</th>
<th>Objective</th>
<th>Data Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q. A. Al-Radaideh, E.M. Al-Shawafka and M.I. Al-Najjar (2006)</td>
<td>To enhance the quality of higher education system by evaluating student’s data and studying the attributes those affect the student performance in courses.</td>
<td>Records of students of C++ course in Yarmouk University, Jordan.</td>
</tr>
<tr>
<td>Z.N. Khan (2005)</td>
<td>To Analyze the Scholastic Achievement of Higher Secondary Students in Science Stream</td>
<td>Records of students (200 boys and 200 girls) selected from Senior Secondary School of Aligarh Muslim University, India.</td>
</tr>
<tr>
<td>Nguyen Thai Nghe, Paul Janecek and Peter Haddawy (2007)</td>
<td>To predict the academic performance of undergraduate and postgraduate students of two very different institutes.</td>
<td>Records of the students of Can Tho University (CTU), Viet Nam and Asian Institute of Technology (AIT), Thailand.</td>
</tr>
</tbody>
</table>

Various algorithms and data mining techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees (CART and CHIAD), Genetic algorithms, Nearest Neighbor method etc. are used for knowledge discovery from databases and helps in prediction of students academic performance. The objective of prediction is to estimate the unknown value of a variable that describes the student. In education, the values normally predicted are performance, knowledge, score or mark. This value can be numerical/continuous value (regression task) or categorical/discrete value (classification task). Regression analysis finds the relationship between a dependent variable and one or more independent variables [18]. Classification is a procedure in which individual items are placed into groups based on quantitative information regarding one or more characteristics inherent in the items and based on a training set of previously labeled items [14]. Prediction of a student's performance is one of the oldest and most popular applications of DM in education and different techniques and models have been applied (neural networks, Bayesian networks, rule-based systems, and regression and correlation analysis). Table 2 illustrates a sample of different data mining techniques which are implemented on the dataset of students.
Table 2: Summary of different data mining tools used on different datasets of students

<table>
<thead>
<tr>
<th>Author</th>
<th>Data Mining Techniques</th>
<th>WORK DONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. Cortez and A. Silva (2008)</td>
<td>Decision Tree, Random Forest, Neural Network and Support Vector Machine</td>
<td>Decision Tree and Neural Network algorithms have more predictive accuracy for two-class dataset and four-class dataset.</td>
</tr>
<tr>
<td>Z.N. Khan (2005)</td>
<td>Cluster Sampling</td>
<td>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.</td>
</tr>
<tr>
<td>Z. J. Kovacic (2010)</td>
<td>CHAID and CART</td>
<td>The algorithms were applied on student enrolment data to get two decision trees classifying successful and unsuccessful students.</td>
</tr>
<tr>
<td>Nguyen Thai Nghe, Paul Janecek and Peter Haddawy (2007)</td>
<td>Decision Trees and Bayesian Networks</td>
<td>Decision Tree was consistently 3-12 % more accurate than Bayesian Network.</td>
</tr>
</tbody>
</table>

Different Data Mining Tools are available like WEKA, TANAGRA etc. in which data mining techniques are easily implemented and help to predict the student's academic performance. WEKA [22] formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. The WEKA workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [11]. It is freely available software. It is portable & platform independent because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. WEKA has several standard data mining tasks, data preprocessing, clustering, classification, association, visualization, and feature selection. The WEKA GUI chooser launches the WEKA’s graphical environment which has six buttons: Simple CLI, Explorer, Experimenter, Knowledge Flow, ARFF Viewer & Log.

TANAGRA [19] is free data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. TANAGRA is an “open source project” as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license.
The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it) and allowing to analyze either real or synthetic data. Table 3 indicates accuracy of various data mining techniques which have been implemented on sample size of students.

Table 3: Summary of accuracy of student's performance using enrolled students

<table>
<thead>
<tr>
<th>Author</th>
<th>SAMPLE SIZE</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. Cortez and A. Silva (2008)</td>
<td>Students of two secondary schools</td>
<td>93%</td>
</tr>
<tr>
<td>Students of two secondary schools</td>
<td>91%</td>
<td>72%</td>
</tr>
<tr>
<td>S. Kotsiantis, C. Pierrakeas, P. Pintelas (2004)</td>
<td>365</td>
<td>74%</td>
</tr>
<tr>
<td>O. A. Al-Radaideh, E.M. Al-Shawakfa and M.I. Al-Najjar (2006)</td>
<td>Students of C++ course in Yarmouk University</td>
<td>38.33%</td>
</tr>
<tr>
<td>Z.N. Khan (2005)</td>
<td>400</td>
<td>-</td>
</tr>
<tr>
<td>Z. J. Kovacic (2010)</td>
<td>Student enrolment data of information system</td>
<td>59.4%</td>
</tr>
<tr>
<td>Nguyen Thai Nghe, Paul Janecek and Peter Haddawy (2007)</td>
<td>20,492 (CTU) 936 (AIT)</td>
<td>86% (CTU) 74% (AIT)</td>
</tr>
</tbody>
</table>

According to paper [1], the prediction results using CFS as attribute selection technique shows that Naïve Bayes, AODE and RBF Network performed best on the data sets with 95.29% accuracy, on the other hand AODE score best with CoE showing 95.29% accuracy. As per paper [13], it was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset. As mentioned in paper [17], it was noticed that the Naïve-Bayes algorithm yielded high predictive accuracy (74%) for two-class (pass/fail) dataset. As given in paper [15], three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models with the predictive accuracy of 38.33% for four-class response variable. As provided in paper [25], 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. As mentioned in paper [24], the accuracy obtained with CHAID and CART was 59.4% and 60.5%. According to paper [10], the overall prediction accuracy from analysis was 86% (CTU) and 74% (AIT) for the 3-class prediction.
3. METHODS

3.1. Data Mining

Data mining also popularly known as Knowledge Discovery in Database refers to extracting or “mining” knowledge from large amounts of data. Data mining is an exploratory data analysis [12], a data analysis methodology [20], a process of obtaining knowledge [6], a process and methodology for applying tools and techniques [3], an art of extracting information from data [21], a process of discovering patterns in data [23] and a task of discovering meaningful data from big data [16] with the aim of obtaining clear and useful results [5]. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The sequence of steps used in educational knowledge discovery and data mining process as shown in figure 1.

The sequence of steps identified in extracting knowledge from data is given below:

- Establishing the mining goals: using domain knowledge to select data relevant to the research goal.
- Selection of data: identifying the characteristics of variables on which mining can be performed.
- Data pre-processing: removing noisy, erroneous and incomplete data.
- Data transformation: transforming the data into a new format in order to mine additional information.

![Diagram showing the educational knowledge discovery and data mining process](image)

**Figure1: Educational knowledge discovery and data mining process**

- Data warehousing: the process of envisioning, planning, building, using, managing, maintaining and enhancing databases.
- Data mining: discovering correlations among variables after performing data mining and finding interesting, meaningful and valuable knowledge based on the research topic.
- Evaluating the mining results: elaborating and evaluating the results after knowledge is obtained.

3.2. Educational Data Mining

Applying data mining (DM) in education is an emerging interdisciplinary research field also known as educational data mining (EDM). It is concerned with developing methods for exploring the unique types of data that come from educational environments. Its goal is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and to gain insights into and explain educational phenomena. Educational information systems can store a huge amount of potential data from multiple sources coming in different formats and at different granularity levels. Each particular educational problem has a specific objective with special characteristics that require a different treatment of the mining problem. Educational data mining (EDM) is concerned with developing, researching and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist. EDM has emerged as a research area in recent years aimed at analyzing the unique kinds of data that arise in educational settings to resolve educational research issues [2]. In fact, EDM can be defined as the application of data mining (DM) techniques to this specific type of dataset that come from educational environments to address important educational questions.
3.3. Data Mining Techniques

Various techniques and methods in data mining need brief mention to have better understanding [4].

3.3.1. Classification

Classification is the most commonly applied data mining technique which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

3.3.2. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification.

3.3.3. Prediction

Regression technique can be adapted for prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g. logistic regression, decision trees or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

3.3.4. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

3.3.5. Neural Networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

3.3.6. Decision Trees

A decision tree is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

The three widely used decision tree learning algorithms are: ID3, C4.5 and CART.

3.3.6.1. ID3

ID3 is a simple decision tree learning algorithm [8]. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given set, we introduce a metric - information gain. To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

3.3.6.2. C4.5

This algorithm is a successor to ID3 [8]. C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.
3.3.6.3. CART

CART [9] stands for Classification and Regression Trees. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

4. CONCLUSION

This paper analyzed the potential use of data mining in education section and surveys the most relevant work in this area. Data Mining can be used for dropout students, student’s academic performance, teacher’s performance and student’s complaints. One of the biggest challenges that higher education faces today is that how to improve the quality of education, how to improve the learning experience of students as well as their interest in education and which career options should be opted by students according to their ability.

The following will be the benefits of data mining in an education sector:

- Predicting the student’s performance on the basis of student’s database.
- Identifying student’s pattern trends.
- Identifying those students which needed special attention to reduce failing ration and taking appropriate action at right time.
- Helping the students to improve their performances.

Our future work include applying data mining techniques on an expanded data set with more distinct attributes to get more accurate results. The future work can be done using data mining techniques such as neural nets, genetic algorithms, k-mean and other data mining model.

REFERENCES


[7] International Educational Data Mining Societywww.educationaldatamining.org


