Student Performance Prediction by Pattern Mining Technique

Priyanka A. Patil¹, R. V. Mane²

¹Student, Department of Technology, Shivaji University, Kolhapur, India,
²Assistant Professor, Department of Technology, Shivaji University, Kolhapur, India,
¹priyanka.apatil1990@gmail.com, ²rvm_tech@unishivaji.ac.in

Abstract

Student’s academic performance is critical for educational institutions. Strategic programs can be planned for improving or maintaining student’s performance during their period of studies in the institutions. This system proposes the application of data mining techniques to predict student’s performance. Real world data of students such as engineering students can be used in sequential pattern mining algorithms, which finds most frequent patterns. Classification algorithm can be used to generate rules to predict the performance based on frequent patterns. Students who are at risk of failure can be provided guidance.

1. Introduction

The main goal of educational institutions is to develop the best quality of the education process. Examination plays a vital role in any student’s life. The marks obtained by the student in the examination decide his future. Although the educational level has improved in the last decades, recent years have shown a growing interest and concern in many countries about problem of student failure and the determination of its main contributing factors. Therefore it becomes essential to predict whether the student will pass or fail in the examination [1]. The research can be done on identifying the factors that affect the low performance of students using the large amount of information that current computers can store in databases.

Real-world data of student information (e.g. student marks, family background, social and academic related features) can be collected by using college reports and surveys. If properly analysed, these data could be used by the academic organizations to understand the behaviour of students. Identifying and finding useful information hidden in large databases is a difficult task.

Data mining is the process of analysing data from different perspectives and summarizing the results as useful information. A solution to achieve this goal is to use the knowledge discovery in databases techniques or data mining in education, called educational data mining, EDM [2]. The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research. Data mining has been recognized as promising area for database research. This area can be defined as efficiently discovering interesting rules from large database. This new area of research focuses on the development of methods to better understand students and the settings in which they learn. In fact, there are good examples of how to apply EDM techniques to create models that predict student performance specifically. This proposed system will show promising results with respect to those sociological, economic or educational characteristics that may be more relevant in the prediction of low academic performance.

Educational data mining researchers uses the different technical methods to accomplish the goal of their research such as prediction, clustering, classification, relationship mining [3]. Rule learning is a popular method for discovering interesting relations between variables in large databases [4]. Finding all frequent item sets in a database is difficult task since it involves searching all possible item sets that is item combinations.

Proposed system uses two data mining techniques to study the performance of a student which are: sequential pattern mining algorithm methodology [5] and classification algorithm. Pattern mining algorithm finds relationships between occurrences of sequential patterns. Classification algorithms can be used to generate rules of frequent patterns obtained by pattern mining algorithm. Based on these rules it is possible to predict the student performance in the future.
The rest of the paper is organized as follows. Section 2 describes related work. Section 3 describes methodology. Section 4 presents algorithm for finding frequent patterns. Section 5 describes prediction of student performance.

2. Related Work

H. Bydovska and L. Popelinsky used student academics related data and social behavioural data of students to predict the performance [6]. They carried out different experiments using data mining techniques such as support vector machine (SMO), OneR rule, Naïve bayes and decision tree. Support vector machine technique gave most accurate result.

In Canada there is a failure rate of more than 30% after the first two years in the Faculty of Engineering [7]. Different data mining techniques such as clustering and classification approaches e.g. K-means and hierarchical clustering, and K-nearest neighbour and naïve Bayes classifiers can be used to predict the failure rate of students.

For avoiding extra, unofficial seats during engineering admissions, New Zealand Government has adopted policy of penalizing the institute [8]. So as a result, New Zealand Universities adopted entry policy i.e. student’s performance can be evaluated during final year of secondary school. Based on this, performance prediction of success of particular student in engineering can be done. Students who do not perform well can be restricted from admission of engineering. Mathematics with calculus, Mathematics with statistics, Physics etc. subject’s marks were considered for prediction and data mining algorithms were applied on it.

S.Anupama Kumar and Vijayalakshmi M.N concluded that Decision rule and One R rule algorithms can be used to predict the result of the fifth semester of student in higher education based on the marks obtained by the students in the previous four semesters [9]. Rule based algorithm can provide efficiency in predicting the student’s performance in higher education using the previous historical data.

Chengguan Xiang and Shihuan Xiong applied GSP algorithm in dynamic cost prediction of enterprise [10]. By mining sequential pattern from the historical data, cost variation of the enterprise in the future can be predicted. It will be useful for the administrators to make decision in advance and to take measures for controlling the cost at the lowest level.

3. Methodology

The method proposed in this paper for predicting student performance consists of Data Mining techniques (see Fig. 1). The main stages are:

3.1 Data Collection

There are large amounts of risk factors that can affect student’s performance in exams such as demographics, family background, educational background, psychological profile, academic progress, and extra-curriculum. These attributes can be collected by different surveys from college as given below.

a. Specific Survey: will obtain personal and family background related information of students.

b. Departmental Survey: will identify academics related factors of students.

Finally all data should be integrated into single dataset.

3.2 Preprocessing

Before applying DM algorithms it is necessary to carry out some pre-processing tasks such as cleaning, integration and variable transformation. It is very important task in this work because the quality and reliability of available information, directly affects
the results. Other techniques such as attribute selection can be applied to reduce the dimensionality of the dataset.

3.3 Data Mining

At this stage, data mining algorithms are applied to predict student performance. For doing this task, we propose the use of pattern mining algorithm and classification algorithm.

3.4 Interpretation

At this stage, the obtained result is analyzed to detect student performance.

4. GSP Algorithm for Finding Frequent Patterns

This section describes the proposed data mining techniques for obtaining student performance prediction. After pre-processing is done, data mining algorithms can be applied. Pattern mining algorithms that discover sequential patterns affecting student performance from student dataset can be applied. In recent years, many sequential pattern mining algorithms have been proposed.

GSP algorithm scales linearly with the number of data sequences and has very good scale-up properties with respect to the average data sequence size [5]. The support of a particular sequence is defined as the total number of data sequences that contain this sequence. GSP algorithm works with finding all sequential patterns with user-defined minimum support. GSP algorithm is much faster than AprioriAll algorithm.

The basic structure of Pattern mining (GSP) algorithm for finding sequential patterns will be as follows [5].

As shown in Table 1, the algorithm starts with finding support for each attribute in the student database. Attributes which exceed the threshold are considered as frequent items in the first pass. The frequent items found in previous passes are used for generating frequent candidates in each pass. Each pass contains 1-item more than previous passes. The algorithm terminates when there are no candidate sequence generated at the end of the pass.

At the end of the last pass, candidate sequence patterns can be used as the most frequent patterns that affect student performance. These frequent patterns can be given as input to classification algorithms.

Table 1. GSP Algorithm

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Let, $S = {S_1, S_2, ..., S_n}$ be a sequence of itemset of students. Let $X = {X_1, X_2, ..., X_m}$ be elements of each set. Let, $D$ be student database consisting of data sequences. Let, $F$ be the set of frequent items. Let, $C$ be the Candidate set.</td>
</tr>
<tr>
<td>2.</td>
<td>$F = \emptyset$;</td>
</tr>
<tr>
<td>3.</td>
<td>Evaluate Support of each element $X_i$ in sequence $S_i$ as $\text{Sup}_{X_i}(S_i)$.</td>
</tr>
<tr>
<td>4.</td>
<td>Let, $\text{Min}<em>\text{Sup}$ be user defined minimum support, where, $0 \leq \text{Min}</em>\text{Sup} \leq 1$.</td>
</tr>
<tr>
<td>5.</td>
<td>If $\text{Sup}<em>{X_i}(S_i) \geq \text{Min}</em>\text{Sup}$ then $F = S_i$ Else Remove sequence $S_i$ Repeat for each $i$</td>
</tr>
<tr>
<td>6.</td>
<td>Let, $S_j$ and $S_j$ be frequent sequence</td>
</tr>
<tr>
<td></td>
<td>a. Delete last item of $S_i$ and Delete 1st item of $S_j$</td>
</tr>
<tr>
<td></td>
<td>b. If same sequence</td>
</tr>
<tr>
<td></td>
<td>If last item in $S_i$ is separate element $F = S_i$ union last item of $S_j$ Else $F = {S_i$ - last item of $S_i} \cup {\text{Last item of } S_i + \text{Last item of } S_j}$</td>
</tr>
<tr>
<td></td>
<td>Repeat for each $i$ and $j$</td>
</tr>
<tr>
<td>7.</td>
<td>$C = F_1 \cup F_2 \cup ... \cup F_n$</td>
</tr>
<tr>
<td>8.</td>
<td>Evaluate support for $C$</td>
</tr>
<tr>
<td></td>
<td>If $\text{Sup}(F_i) \geq \text{Min}_\text{Sup}$ then $C = F_i$ Else Delete $F_i$ from $C$</td>
</tr>
<tr>
<td></td>
<td>Repeat for each $i$ until no candidate sequence $C$ generated.</td>
</tr>
</tbody>
</table>
5. Prediction of Student Performance

After pattern mining algorithms last stage, most frequent candidate sequences are found. One of the decision tree algorithms such as ID3, C4.5, CART, J48 etc. [1] can be used to classify the frequent candidate sequences into different classes such as high performance students, low performance students. A decision tree is tree-shaped structure that represents sets of decisions. It can be directly transformed into if-then rules for interpretation of the result [2].

1) Decision tree starts with root node on which actions has to be taken. Possible values of student attributes can be represented as arcs.

2) From this node, each node is split recursively according to threshold given to decision tree algorithm. Threshold can be given such that it generates different classes such as high performance, low performance students.

3) After the tree is built, each branch represents possible rules for performance prediction.

Decision tree can be interpreted easily as if-then rules. These rules can show factors and relationships that influence student academic performance. Students who are at the risk of failure can be given guidance by the teacher.

6. Conclusion

Accurately predicting student performance is useful in different ways in universities. For example, identifying bright students for scholarships and identifying weak students who are likely to fail is also important for their right guidance.

The student database may contain many records. It is impossible to find patterns from such large database. Proposed pattern mining algorithm can find patterns that most influence student performance by counting fewer candidates sets. Proposed system can be used for predicting performance of students with data mining techniques such as pattern mining algorithm and classification algorithm. Once students are found at the risk of failure, they would be provided guidance for improvement.

7. References


