Data Mining Approaches on Detection of Students’ Academic Failure and Dropout: A Brief Survey

Devikala.D M.phil¹ and Kamalraj.N MCA, M.phil²

¹Department of Computer science, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore
²Head of the Department (B.Sc .CT), Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore

Abstract - Educational data mining (EDM) is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data. The discipline focuses on analyzing educational data to develop models for improving learning experiences and improving institutional effectiveness. Exams failure among school students has long fed a large number of debates, many education experts seeking to comprehend and explicate it, and many statisticians have tried to predict it. Understanding, predicting and preventing the academic failure are complex and continuous processes anchored in past and present information collected from scholastic situations and students’ surveys, but also on scientific research based on data mining technologies. Data mining is taken as a process of transform knowledge into some human understandable format like rule, formula, theorem, etc. This article provides a Review of the available literature on Educational Data mining, Classification method and different feature selection techniques that we should apply on Student dataset. The knowledge is hidden among the educational data set and it is extractable through data mining techniques.

Keywords: Data Mining, Education data mining, Classification techniques, association rule mining, Outlier Detection and Clustering techniques.

I. INTRODUCTION

An educational system has large number of educational data. This data may be students’ data, teachers’ data, alumni data, resource data etc. Educational data mining is used to find out the patterns in this data for decision-making.

There are two types of education system:

1) Traditional Education system: In this system there is direct contact between the students and the teacher. Students’ record including the information such as attendance, grades may be kept manually or digitally.

Students’ performance is the measure of this information.

2) Web based learning system: It is also known as e-learning. It is becoming more popular as the students can learn from any place without any time constraint. In a web based system, various data about the students are automatically collected through logs.

Educational data mining can answer number of questions from the patterns obtained from student data such as

1) Who are the students at risk?

2) What are the chances of placement of student?

3) Who are the students likely to drop the course?

4) What is the quality of student participation?

5) Which courses the institute should offer to attract more students?

Results of educational data mining can be used by different members of education system. Students can use them to identify the activities, resources and learning tasks to improve their learning. Teachers can use them to get more objective feedback, to identify students at risk and guide them to help them succeed, to identify the most commonly made mistakes and to organize the contents of site in efficient way. On the other hand, administrators can use them to decide which courses to offer, which alumni are likely to donate more to the institution etc.

Data mining is the process of analyzing data from different perspectives and summarizing it into important information so as to identify hidden
patterns from a large data set. Educational Data Mining (EDM) is an emerging discipline, concerned with data from academic field to develop various methods and to identify unique patterns which will help to explore student’s academic performance. EDM can be considered as learning science, as well as a feature of data mining. Assessing students learning process is a very complex issue. Education Data mining helps in predicting students’ performance in order to recommend improvements in academics. The past several decades have witnessed a rapid growth in the use of data and knowledge mining as a tool by which academic institutions extract useful unknown information in the student result repositories in order to improve students’ learning processes.

There are increasing research interests in using data mining in education. These new emerging fields, called educational data mining, concerned with developing methods that extract knowledge from data come from the educational context. The data can be collected from e-learning systems which have a large amount of information used by most institutes.

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students’ performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students’ academic performance and trim down failure rate, to better understand students’ behavior, to assist instructors, to improve teaching and many other benefits.

Educational data mining uses many techniques such as decision tree, rule induction, neural networks, k-nearest neighbor, naïve Bayesian and many others. By using these techniques, many kinds of knowledge can be discovered such as association rules, classifications and clustering. The main objective of this study is prediction of student’s performance 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.

II. DATA MINING TECHNIQUES

A. Classification techniques

k-Nearest Neighbour classifier (k-NN)

k-NN algorithm is one of the well-known classification methods. It is based on learning by comparing a given test tuple with training tuples that are similar to it. When a new instance is introduced, k-NN finds the k-nearest neighbours of this new instance and determines the label of the new instance by using these k instances (Hand, Mannila & Smyth, 2002). In this study, closeness is defined in terms of a distance metric d called Euclidean distance (Han & Kamber, 2006). Although our data set is mostly consisted of categorical variables, each category has a numerical counterpart; thus, we have used the Euclidean distance. To assign a particular class to the test sample, the most common class among the k nearest neighbours is used and the unclassified test sample is classified by a majority vote of its neighbours. A good value for k, the number of neighbours, was determined experimentally. Starting with k=1, we used the 10-fold cross validation technique to estimate the error rate of the classifier. This process was repeated for k=10 times and in each iteration by incrementing k to allow for one more neighbour. The value k was selected as ‘3’ that gave the minimum error rate. According to Hamäläinen and Vinni (2010), this method has several advantages: The accuracy rate of classification can be very satisfying in some cases, there are just two parameters to learn, and the classification is very robust to missing values. However, the selection of distance function d might be difficult particularly for educational data sets (Hamäläinen & Vinni, 2010).

Naive Bayes classifier (NB)

A simple probabilistic classifier called as Naive Bayes classifier was also used in student dropout classification. Naive Bayes algorithm as the simplest form of Bayesian network (Domingos & Pazzani, 1997) is one of the easiest algorithms to perform and has very satisfactory accuracy and sensitivity rates (Kotsiantis, Pierrakeas & Pintelas, 2003). The posterior probability of each class, Ci, is obtained by the Naive Bayes classifier using Bayes rule. The classifier makes the simplifying assumption that the attributes, A, are independent given the class, so the likelihood can be obtained by the product of the individual conditional probabilities of each attribute given the class (Flach & Lachiche, 2004). Thus, the posterior probability, \( P(C_i | A_1, A_2, ..., A_n) \), can be given by the following equation/assumption:
\[ P(C_i|A_1, \ldots, A_n) = \frac{P(C_i) P(A_1|C_i) \ldots P(A_n|C_i)}{P(A)} \]

This assumption is usually called the Naive Bayes assumption, and a Bayesian classifier using this assumption is called the Naive Bayesian classifier, often abbreviated to ‘Naive Bayes’. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class.

Neural Network classifier (NN)

The prediction of the student dropouts was also performed by feed-forward NN. It is another inductive learning method grounded on computational models of neurons and their networks as in humans' central nervous system (Mitchell, 1997). NN is a set of connected input/output units where each connection has a distinct weight associated with each other (Kotsiantis, Pierrakeas & Pintelas, 2003). During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class of the input samples (Han & Kamber, 2006).

In this study, the back propagation algorithm was performed for learning on a multilayer feedforward neural network. The input layer of the network consisted of nine variables of the students. The hidden layer included 50 neurons and the output layer had one neuron, which was determined by our experimental studies.

Decision tree methods

Decision trees are often used in classification and prediction. It is simple yet a powerful way of knowledge representation.

The decision tree classifier has two phases [1]:

a) Growth phase or Build phase.

b) Pruning phase.

The tree is built in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belonging to each of the partitions bearing the same class label. The tree may over fit the data. The pruning phase handles the problem of over fitting the data in the decision tree. The prune phase generalizes the tree by removing the noise and outliers. The accuracy of the classification increases in the pruning phase. Pruning phase accesses only the fully grown tree. The growth phase requires multiple passes over the training data. The time needed for pruning the decision tree is very less compared to build the decision tree.

(i). ID3 (Iterative Dichotomies 3)

ID3 algorithm introduced by J. R. Quinlan [2] is a greedy algorithm that selects the next attributes based on the information gain associated with the attributes. The attribute with the highest information gain or greater entropy reduction is chosen as the test attribute for the current node. The tree is constructed in 2 phases. The two phases are tree building and pruning. ID3 uses information gain measure to choose the splitting attribute. It only approves categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise preprocessing technique has to be used. To build decision tree, information gain is calculated for every single attribute and select the attribute with the greatest information gain to designate as a root node. Label the attribute as a root node and the possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances. Continuous attributes can be handled using the ID3 algorithm by discretizing or directly, by considering the values to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

(ii). C4.5

C4.5, the most popular algorithm, is a successor of ID3. C4.5 made a number of improvements to ID3. This algorithm is a successor to ID3 developed by Quinlan Ross [2]. It is also based on Hunt’s algorithm. 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. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.
CART

CART stands for Classification and Regression Trees introduced by Breiman [2]. It is also based on Hunt’s algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. If the target variable is nominal it generates classification tree and for continuous-valued numerical target variable it generates regression tree. Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

ADT (Alternating Decision Tree)

ADTrees were introduced by Yoav Freund and Llew Mason [3]. It generalizes decision trees and has connections to boosting. An alternating decision tree consists of two nodes. One is decision nodes and other is prediction nodes. First nodes specify a predicate condition. Second nodes contain a single number. ADTrees have prediction nodes as both root and leaves also.

B. Clustering techniques

Clustering analysis is a common unsupervised learning technique. Its aim is to group objects into different categories. That is, a collection of data objects that are similar to one another are grouped into the same cluster and the objects that are dissimilar are grouped into other clusters. It is an important technique in data mining to analyze high-dimensional data and large scale databases.

Clustering algorithms can be classified into hierarchical and non-hierarchical algorithms (Han & Kamber, 2006). The hierarchical procedure produces a tree-like structure, which is able to see the relationship among entities. The hierarchical clustering procedure can be agglomerative or divisive. On the other hand, nonhierarchical methods do not possess tree-like structures but assign some cluster seeds to central places, also called k-means clustering. There are three methods to assign an object to a group, namely the sequential threshold, parallel threshold and optimization partitioning procedures.

K-means

The k-means algorithm is one of the best known and simplest clustering algorithms. It was proposed over 50 years ago and still widely used (Hosseini et al., 2010; Jain, 2009; Yang et al., 2009). This is due to its ease of implementation, simplicity, and superior feasibility and efficiency in dealing with a large amount of data. However, it is sensitive to initialisation and is easily trapped in local optima (Hosseini et al., 2010; Kanungo et al., 2002; Mingoti & Lima, 2006; San et al., 2004; Yang et al., 2009). In addition, the main shortcoming of the k-means algorithm is that it depends heavily on the initial choice of the cluster centres, which reduces its convergence reliability and efficiency (Kao et al., 2008; Mingoti & Lima, 2006; Qiu, 2010; Yang et al., 2009).

The k-means algorithm is a non-parametric approach that aims to partition objects into k different clusters by minimising the distances between objects and cluster centers (Qiu, 2010). The k-means algorithm contains the following steps:

1. Select initial centers of the k clusters,
2. Assign each object to the group that is closest to the centroid,
3. Compute new cluster centers as the centroids of the clusters,
4. Repeat Steps 2 and 3 until the centroids no longer move.

Self-organising maps

The self-organising map (SOM) or self-organising feature map network (SOFM) was proposed by Kohonen (1982, 2001). SOM is an unsupervised neural network consisting of an input layer and the Kohonen layer. It is usually designed as a two-dimensional arrangement of neurons that maps an n-dimensional input to a two-dimensional map (Budayan et al., 2009; Mingoti & Lima, 2006). Particularly, SOM provides a topological structure imposed on the nodes in the network, and preserves neighborhood relations from the input space to the clusters (Kohonen, 1989; 2002). The learning algorithm of SOM is described as follows (David & Yong, 2007):

1. Initialise the map: this stage aims to initialise reference vectors, set up the parameters of the
2. Determine the winning node: select the best matching node that minimises the distance between each input vectors by the Euclidean distance;

3. Update reference vectors: updating reference vectors and its neighborhood nodes based on the learning criterion;

4. Iteration: iterate Steps 2 and 3 until the solution can be regarded as steady.

Two-step clustering: BIRCH

The BIRCH (balanced iterative reducing and clustering using hierarchies) algorithm contains two main steps and hence is known as a two-step clustering (Markov & Larose, 2007). BIRCH is an integrated hierarchical clustering method (Han & Kamber, 2006). It introduces the concepts of clustering feature (CF) and clustering feature tree (CF tree), and these structures help achieve good speed and scalability for very large datasets. A CF is a triple that stores the information about sub-clusters of objects; a CF tree is a height balanced tree used to store the clustering features (Han & Kamber, 2006). Unlike k-means and SOM, the BIRCH clustering algorithm represents a desirable exploratory tool, for which the number of clusters does not need to be specified at the beginning (Markov & Larose, 2007).

BIRCH performs the following steps:

1. Load data into memory by building a CF tree;

2. Condense the initial CF tree into a desirable range by building a smaller CF tree (optional);

3. Perform global clustering;

4. Perform cluster refining (optional).

C. Association rule mining

In education data mining, association rule learning is a conventional and well researched method for determining interesting relations between attributes in large databases [4]. Association rule Mining is mainly intended to recognize strong rules from databases using different measures support and confidence. The preliminaries necessary for performing data mining on any data are discussed below. Let I= {I1, I2, I3….Im} be a set items. Let D, the task relevant data, be a set of database transactions where each transaction Ti. Each transaction is an association with an identifier, called transaction identification (TID).

Let A be a set of items. A transaction T is said to contain A if and only if A Î T. An association rule is an implication of the form A⇒ B, where A Î I, B Î I, and A∪B =Æ. [4] Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule A⇒ B means that A and B exist together in 2% of all the transactions under analysis. The rule A⇒B having confidence of 60% in the transaction set D means that 60% is the percentage of transactions in D containing A that also contains B.

A set of items is referred to as an item set. An item set that contains k items is a k-item set. The occurrence frequency of an item set is the number of transactions that contain the item set. If the relative support of an item set I satisfies a prescribed minimum support threshold, then I is a frequent item set. The association rule mining can be viewed as a two-step process:

1) Find all frequent item sets: Each of these item sets will occur at least as frequently as a predetermined minimum support count.

2) Generate strong association rules from the frequent item sets: The rules must satisfy minimum support and confidence. These rules are called strong rules. [4]

Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agarwal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. The following lines state the steps in generating frequent item set in Apriori algorithm. [5]

Let Ck be a candidate item set of size k and Lk as a frequent item set of size k. The main steps of iteration are:

• Find frequent set Lk-1

• Join step: Ck is generated by joining Lk -1 with itself (Cartesian product Lk-1 × Lk-1)
• Prune step (apriori property): Any \((k - 1)\) size item set that is not frequent cannot be a subset of a frequent \(k\) size item set, hence should be removed

• Frequent set \(L_k\) has been achieved [4].

D. Outlier Detection

Outlier detection discovers data points that are significantly different than the rest of the data [6]. In educational data mining outlier analysis can be used to detect students with learning problems [7]. In this paper, we used outlier analysis to detect outliers in the student data. Two outlier methods are used which are Distance-based Approach and Density-Based Approach.

Distance-based Approach

It Identifies the number of outliers in the given data set based on the distance to their \(k\) nearest neighbors, and the result of applying this method is to flag the records either to be outlier or not, with true or false

Density-Based Approach

It Computes local densities of particular regions and declare instances in low density regions as potential outliers. The method used is Local Outlier Factor (LOF), the Basic idea of LOF is to compare the local density of a point with the densities of its neighbors, and the result of applying this method is to flag the records with a percentage of outlier. The larger score means larger possibility of being outlier [8].

III. CONCLUSION

Educational data mining (EDM) is an area full of exciting opportunities for researchers and practitioners. This field assists higher educational institutions with efficient and effective ways to improve institutional effectiveness and student learning. The results presented in this article are a part of a larger research which is to be used to make numerous correlations, analysis and to be presented to the higher education institution managers, to offer a better knowledge of students’ present scholastic situations, their opinions regarding the each component of the educational process, and to predict some important aspects of their future scholastic situation. The purpose is to contribute to optimal managerial decision taking, in preventing students’ exams failure, improving learning abilities and scholastic results.

Our studies continue with deeper mining of academic failure, to detect with pragmatic exactness what and how much the students know, to understand precise learning gaps, and also improve teaching methods and educational management processes.

REFERENCES


