ISSN 2320 - 2602

Volume 3, No.5, May 2014

International Journal of Advances in Computer Science and Technology

Available Online at http://warse.org/pdfs/2014/ijacst03352014.pdf

Survey on Students' Academic Failure and Dropout using Data Mining Techniques

P.Senthil Vadivu[1], D.Bharathi[2]

[1]Head and Associate Professor, Department of computer Applications, Hindustan college of Arts and Science, India, sowju_sashi@rediffmail.com

[2]Research Scholar, Hindustan college of Arts and Science, India, bharathi196@gmail.com

ABSTRACT

Educational data mining is in the habit of learn the data available in the field of education and show up the hidden knowledge from it. Classification methods like decision trees, machine learning, rule mining, etc can be applied on the educational data for forecasting the student's behavior, performance of them in examination etc. This prediction will well helpful for tutors to classify the weak students and help them to score improved marks. The classification approach is applied on student's internal assessment data to predict their performance in the final exam. The result of the classification categorized the number of students who are to be expected to fail or pass. The outcome result is given to the tutor and steps were taken to improve the performance of the students who were predicted as fail in the examination. After the statement of the results in the final examination the marks acquired by the students are provide into the system and the results were investigated. The proportional analysis results states that the prediction has helped the weaker students to improve and brought out betterment in the result. The algorithm is also analyzed by duplicating the same data and the result of the duplication brings no much change in predicting the student's outcome. The goal of this survey is presented the several data mining techniques in determining of student failure. 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.

Keywords: Educational data mining, Classification techniques, Attribute selection techniques and Clustering techniques.

1. INTRODUCTION

An educational system has large number of educational data. The educational data is possibly 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 containing the information such as attendance, grades or marks which is get from the examination that 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 attractive more admired as the students can gain knowledge from any place without any time restriction. In a web based system, several data about the students are automatically together with the help of logs.

Educational data mining can response number of queries from the prototypes attained 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 intention feedback, to identify students at risk and guide them to help them succeed, to identify the most frequently made mistakes and to categorize the contents of site in efficient way. In addition to, administrators can use them to make a decision which courses to offer, which alumni are likely to contribute more to the institution etc.

Data mining is the course 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 a promising discipline, concerned with data from academic field to enlarge different methods and to discover unique patterns which will facilitate to explore student's academic performance. EDM can be well thought-out as learning science, as well as an important feature of data mining. Assessing students' knowledge process is a very difficult problem. Education Data mining is used to predicting students' performance with the aim of recommend improvements in academics. The past several decades have witnessed a speedy growth in the use of data and knowledge mining as a tool by which academic institutions take out useful unknown information in the student result repositories with the purpose of improves students' learning processes.

There are growing research interests in using data mining in education field. These new promising fields, called educational data mining, troubled with developing approaches that knowledge extracted from data come from the educational context. The data can be gathered form historical and prepared data is located in in the databases of educational institutes. The student data can be personal or academic. In addition it can be accumulated from e-learning systems which have a large amount of data used by most institutes.

The main intention of higher education institutes is to offer worth education to its students and to get better quality of managerial decisions. One way to attain highest level of quality in higher education system is by determining knowledge from educational data to learn the main features that may affect the students' performance. The discovered knowledge can be used to recommend a helpful and constructive suggestions to the academic planners in higher education institutes to improve their decision making process, to enhance students' academic performance and reduce failure rate, to better understand students' behavior, to support instructors, to get better teaching and many other benefits.

Educational data mining uses many methods such as decision tree, rule induction, neural networks, k-nearest neighbor, naïve Bayesian etc. Many categories of knowledge can be revealed such as association rules, classifications and clustering by using these techniques. 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.

2. DATAMINING TECHNIQUES

1. Attribute selection methods

Data reduction techniques can be classified into two types: Wrapper and filter method. Wrapper model method is used for classification itself to measure the importance of features set, hence the feature selected depends on the classifier model used. In general, Wrapper methods result in better performance than filter methods. However, wrapper methods are too costly for large dimensional database in terms of complexity of computation and complexity of time because each feature set considered must be analyzed with the classifier algorithm used. The filter approach actually comes first in the actual classification process. The filter method is free from the learning algorithm, computationally simple fast and scalable.

Using filter method, feature selection is done once and then can be provided as input to different classifiers. A range of feature ranking and feature selection techniques have been proposed such as Correlation-based Feature Selection (CFS), Principal Component Analysis (PCA), Gain Ratio (GR) attribute evaluation, Chi-square Feature Evaluation, Fast Correlation-based Feature selection (FCBF), Information gain, Euclidean distance, i-test, Markov blanket filter. Some of these filter methods do not perform feature selection but only feature ranking hence they are combined with search method when one needs to find out the appropriate number of attributes. Such filters are often used with forward selection, which considers only additions to the feature subset, backward elimination, bi-directional search, best-first search, genetic search and other methods

Correlation-based Feature Selection (CFS)

CFS evaluates and ranks feature subsets rather than individual features. It prefers the set of attributes that are highly correlated with the class but with low intercorrelation [9]. With CFS several heuristic searching approaches such as hill climbing and best first are often functional to search the feature subsets space in reasonable time. CFS first computes the feature-class matrix and feature to feature correlations from the training data after that searches the feature subset space using a best first.

$$M_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Where Ms is the correlation between the summed feature subset S, k is the number of subset feature, $\overline{r_{cf}}$ is the average of the correlation between the subsets feature and the class variable, and $\overline{r_{ff}}$ is the average inter-correlation between subset features [10].

Information Gain (IG)

The IG evaluates attributes by measuring their information gain with respect to the class. It discretizes numeric attributes first using MDL based discretization technique [9]. Let C be set consisting of c data samples with m distinct classes. The training dataset ci contains sample of class I. Expected information needed to classify a given sample is calculated by:

$$\text{I (c1, c2, ... ,cm)} = - \sum_{i=1}^{m} \frac{c_i}{c}\, log_2\,(\frac{c_i}{c})$$

Where $\frac{c_i}{c}$ is the probability that an arbitrary sample belongs to class Ci. Let feature *F* has v distinct values *{ f1, f2, …, fv } which* can split the training set into v subsets *{C1,C2, …, Cv }* where Ci is the subset which has the value *fi* for feature *F*. Let Cj contain Cij samples of class *i*. The entropy of the feature *F* is given by

$$\text{E(F)} = \sum_{j=1}^{m} \frac{c_{ij} + \ldots\ldots\ldots + c_{mj}}{c} \times I(c_{ij} + \ldots\ldots\ldots + c_{mj})$$

Information gain for F can be calculated as:

$$Gain\,(F) = I(\text{c}1,...,\text{c}m) - E(F)$$

Gain Ratio (GR)

The information gain is helpful to select attributes which having a large number of values. The gain ratio an extension of info gain, attempts to overcome this bias. Gain ratio applies normalization to info gain using a value defined as

$$\text{SplitInfo}_F\ (C) = \sum_{i=1}^{v} \left(\frac{|C_i|}{|C|}\right) log_2 \ \left(\frac{|C_i|}{|C|}\right)$$

The above value represents the information generated splitting the training data set C into v partitions corresponding to v outcomes of a test on the feature F [11].

The gain ratio is defined as

$$Gain\ Ratio(F) = Gain\ (F)/\ SplitInfo_F\ (C)$$

3. CLASSIFICATION TECHNIQUES

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 is one of the easiest algorithms to perform and has very satisfactory accuracy and sensitivity rates. The posterior probability of each class, *Ci*, is obtained by the Naive Bayes classifier using Bayes rule. The classifier formulates the simplifying statement that the attributes, *A*, are independent in the class, so the possibility can be obtained by the product of the individual conditional probabilities of each attribute given the class. Thus, the posterior probability, □ *i n* □ *P C A* ,..., *A* 1 , can be given by the following equation assumption:

$$P(C_i|A_1,. . . . .,A_n) = P(C_i)\ P(A_1|C_i) . . . .. P(A_n\,|\ C_i) / P(A)$$

This assumption is generally called the *Naive Bayes assumption*, and a Bayesian classifier using this assumption is called the *Naive Bayesian classifier*, and also it is frequently abbreviated to ‘Naive Bayes’. In actual fact, 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 carried out 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. NN is a set of connected input/output units where each connection has a distinct weight associated with each other. During the learning phase, the network learns by altering the weights so we can identify the correct class of the input samples.

In this study, the back propagation algorithm was performed for learning on a multilayer feed forward neural network. The input layer of the network consisted of nine variables of the students. The hidden layer contains 50 neurons and the output layer contains one neuron, which was found out by 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.

A. 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 greatest 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.

B. 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 contains both categorical and continuous attributes to build a decision tree. With the purpose of 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 supports missing attribute values. Gain Ratio is an attribute selection measure for C4.5 to build a decision tree. It eliminates the biasness of information gain when there are many outcome values of an attribute. In the beginning for each attribute we are determining the gain ratio. The root node will be the attribute whose gain ratio is high. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

C. *CART*

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

D. ADT (Alternating Decision Tree)

ADTrees were introduced by Yoav Freund and Llew Mason [3]. It simplifies decision trees and has connections to boosting. An alternating decision tree contains of two nodes. One is decision nodes and other is prediction nodes. First nodes specify a predicate state. Second nodes contain a single number. A prediction node is both root and leaves also in ADTrees.

4. 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. 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 types of methods to assign an object to a group, that is to say 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 [4], [5]. 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 initialization and is easily trapped in local optima [4], [5] and [6]. As well, the important limitation of the *k*-means algorithm is that it depends seriously on the initial choice of the cluster centers, which degrades its convergence reliability and efficiency.

The *k*-means algorithm is a non-parametric technique that intends to partition objects into *k* different clusters by minimizing the distances between objects and cluster centers. 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-organizing maps

The self-organizing map (SOM) or self-organizing feature map network (SOFM) was proposed by Kohonen. 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 [6], [7]. Particularly, SOM provides a topological structure imposed on the nodes in the network, and preserves neighborhood relations from the input space to the clusters. The learning algorithm of SOM is described as follows:

1. Initialize the map: this stage aims to initialize reference vectors, set up the parameters of the algorithm, such as the distance of neighborhoods and the learning rate;

2. Obtaining the winning node: select the best matching node that minimizes the distance between each input vectors by the Euclidean distance;

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

4. Iteration: iterate Steps 2 and 3 until the solution can be holed 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* [8]. BIRCH is an integrated hierarchical clustering method. It proposed the idea of clustering feature (CF) and clustering feature tree (CF tree), and these models used to attain 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. Dissimilar to *k*-means and SOM, the BIRCH clustering algorithm characterizes a desirable exploratory tool, for which the number of clusters does not need to be specified at the beginning. BIRCH performs the following steps:

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

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

3. Perform global clustering;

4. Perform cluster refining (optional).

5. CONCLUSION

This study introduced the data mining approach to modeling drop out feature and some implementation of this approach. The key to gaining a competitive advantage in the educational industry is found in recognizing that student databases, if properly managed, analyzed and exploited, are unique, valuable assets. Data mining uses predictive modeling, database segmentation, market basket analysis and combinations to more quickly answer questions with greater accuracy. Several strategies can be developed and implemented enabling the educational institutions to transform a wealth of information into a wealth of predictability, stability and profits. In this article, we are presenting the various data mining techniques related to the students' failure prediction. From this survey article we can improve the performance of the students' failure and dropout prediction.

REFERENCES

[1] U. Fayyad and R. Uthurusamy, “Data mining and knowledge discovery in databases,” Commun. ACM, vol. 39, pp. 24–27 , 1996.

[2] J. R. Quinlan, “Introduction of decision tree”, Journal of Machine learning”, pp. 81-106, 1986.

[3] Yoav Freund and Llew Mason, “The Alternating Decision Tree Algorithm”. Proceedings of the 16th International Conference on Machine Learning, pp. 124-133, 1999.

[4] Hosseini, S. M. S., Maleki, A. & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, 37(7), 5259-5264. http://dx.doi.org/10.1016/j.eswa.2009.12.070

[5] Yang, F., Sun, T. & Zhang, C. (2009). An efficient hybrid data clustering method based on Kharmonic means and particle swarm optimization. *Expert Systems with Applications*, 36(6), 9847-9852. http://dx.doi.org/10.1016/j.eswa.2009.02.003

[6] Mingoti, S. A. & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, Kmeans and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, 174(3), 1742-1759.

[7] Budayan, C., Dikmen, I. & Birgonul, M. T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. *Expert Systems with Applications*, 36(9), 11772-11781.

[8] Markov, Z. & Larose, D. T. (2007). *Data mining the web: Uncovering patterns in web content, structure, and usage.* New York: John Wiley & Sons.

[9] I.H.Witten, E.Frank, M.A. Hall “Data Mining Practical Machine Leanrning Tools & Techniques” Third edition, Pub. – Morgan kouffman.

[10] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer Science, University of Waikato.http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.

[11] j.Han ,M Kamber, Data mining : Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers(2001).