IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

Review on Student Academic Performance Prediction using Data Mining Techniques

Prateek Sakaray[1], Snehal Kankariya[2], Chandini Lulla[3], Yash Agarwal[4], Pankaja Alappanavar[5]

Student, Department of Information Technology, Sinhgad Academy of Engineering, Pune, India[1,2,3,4,]

Assistant Professor, Department of Information Technology, Sinhgad Academy of Engineering, Pune, India[5]

Abstract: Educational Data Mining is an emerging discipline, concerned with student's performance prediction. In this paper student's performance is evaluated by selecting some attributes which generates rules for forming the classification of the instances in the dataset. Data from various sources of college can provide valuable knowledge to predict student's result at institutional level as well as it can provide insights to each individual performance. The process involves various steps, firstly pre-processing has to be carried out on the set of data secondly apply the classification rules on the data that has been processed in the previous step after which we test the results on the different categorical data input. Emerging institutions are using data mining approach to predict their results and also enhance the level of education in the society. The most commonly used algorithm is machine learning technique to predict performance is called Naïve Bayes and Neural Networks. This work presents a methodology on how student's information can provide insights into education at institutional level.

Keywords: Education, dataset, machine learning, Naïve Bayes, Neural Network.

I. INTRODUCTION

Student academic performance has been a concern for different institutions over a period of years. The quality of education has changed due to upcoming technology and up gradation of existing systems due to which there has been a drastic change in the field of academics. Educational data mining is a rising discipline and can be considered as learning science, which helps to develop rules to define the types of data from various educational databases. Predicting student's academic performance is more challenging because of the large data in educational institutes. The study that has been done in various institutes is still insufficient to identify the exact measures that are required for the institutions to yield better results and provide quality of education[2]. Lack of investigations on the factors affecting student's performance, achievement on periodic basis is also the major cause of the low results of the colleges and institutions.

The methods that were previously used for data collection related to students learning experiences mainly focusing on surveys, interviews, focus groups, classroom activities. These methods usually are tedious and time consuming. The data gathered from different sources can be mainly in two forms 1) Structured 2) Unstructured. Educational data mining has basically focused on analysing structured data which has some fields that are categorized according the rules set. The main aim of this study is to analyse different problems faced by students and categorize it to define the rules for prediction system.

Many datasets collected from different sources are highly susceptible to missing and inconsistent data due to which there will be low-quality mining results. We use different techniques for data pre-processing and data cleaning and merge two different datasets by using data integration. There are different reasons for inaccurate data as there may be human or computer errors occurring while data entry.

II. LITERATURE SURVEY

A. Naïve Bayes

The study conducted by Xin Chen is one of the initial studies that proposed students informal conversation on social media into their educational experiences to perform the analysis. This study included various factors which affects the student's performance like too much stress, depression, lack of sleep. The data was collected from twitter hashtags through an exploratory process [1].

The main focus of this study was to determine tweets based on the categories developed in content analysis stage. Naïve Bayes classifier was found to be more effective on the dataset compared to other multi-label classifier. Text processing was carried out by detecting special symbols to indicate different messages with different meaning. For example, # is used to indicate hashtag, @ is used to indicate a user account. Many users sometimes repeats the words so that they can highlight it or emphasize the word, for example "angryyyy", "Monnndayyy".

The technique used to implement multi-label classifier is to transform the multi-label classification problem into multiple single-label classification problems [1]. One simple transformation is called one-versus-all or binary relevance [1]. The following are the basic procedures of the multi-label Naïve Bayes classifier. Evaluation measures for classification models includes accuracy, precision,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

recall. In multi-label classification each documents gets assigned multiple labels.

Evaluation measures for multi-label classifier are most commonly accuracy, precision, recall and the harmonic average between precision and recall. One-versus all binary classification can be created in the matrix in Table 1.

TABLE I . CONTINGENCY TABLE PER CATEGORY

	True c	True not c
Predicted c	true positive (tp)	false positive (fp)
Predicted not c	false negative (fn)	true negative (tn)

B. Data Mining Models

Dataset used in this study by Mohammad Amir Hossain is of Portuguese student on two courses (mathematics and Portuguese) on which the decision tree was built. This dataset was integrated into two datasets related to mathematics (395 examples) and the Portuguese language (649 records) [2]. The data mining rule which was applied over the dataset was the decision tree which is commonly known as classifier.Decision tree is a branching structure that represents a set of rules. Classification is a form of analysis that extracts models describing important data classes [2]. Data classification is a two-step process, consisting of a learning step(where classification model is constructed) and a classification step (where the model is used to predict class labels for given data). This approach comprised of two algorithms 1) Decision Tree and 2) Random Forest.

KNIME Analytics platform tool was used to perform the required work on the data provided. Konstanz Information Miner (KNIME) is a modular, open source platform for data integration, processing, analysis and exploration [2]. KNIME describes the work flow. Work-flows mainly consists of nodes that process data and the transportation of data between different nodes. Nodes read the data from the database and these data clusters are stored internally in table based format consisting of columns with a certain datatype. These tables are then transferred to different nodes which first preprocesses it like handling missing values, filter columns or rows and then builds the predictive model with the help of machine learning algorithms.

The main goal of this [2] is to find the alcohol consumption by secondary school student. There are two cases in which alcohol consumption can be determined 1) alcohol taking in weekdays (Dalc) 2) alcohol taking in weekends (Walc). Only one of the two instances can be predicted at one time so an overall study was conducted by using the Equation 1.

$$\text{Alc} = \frac{\text{Walc X } 2 + \text{Dalc X } 5}{7} \quad (1)$$

Weka tool was used to analyze the different statistical output of the student's data that was then tested for alcohol consumption. This paper [2] basically uses the Business Intelligence (BI) and Data Mining (DM) techniques to predict the alcohol consumption of the student's which affects the academics and results of the college.

C. Apriori Algorithm and Association Rule Mining

Attributes	Description	Values
Schooling Education	Medium in which schooling education is done	English/Non English
Previous programming knowledge	Previous Knowledge About programming	Yes/No
Father/Mother is educated	Whether either father or mother is educated or not	Yes/ No
Graduation%	Percentage of marks obtained in graduation.	Good, Avg, Poor
Attendance%	Attendance of the student.	Good, Avg, Poor
Assignment%	Assignment performance given during the semester.	Good, Avg, Poor
Unit Test Performance%	Percentage marks obtained by a student in Unit Test.	Good, Avg, Poor
University Result%	Percentage marks obtained by the student in university Examination	Good, Avg, Poor

This research on student's academic performance by Suchita Borkar involves the data of students who are pursuing Master of Computer Application (MCA) degree from Pune University. Neural network technique is used for selecting the attributes from a set of attributes [3].

Datasets of students can be categorized in the above Table 1. Methodology used in this approach is used to obtain the accuracy on all attributes and also comparison is done on accuracy with selected attributes. WEKA tool offers a collection of machine learning and data mining algorithms for data preprocessing, classification, regression etc. This tool uses ARFF file format as an external representation format [3]. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling [5].

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

The main two algorithms for which the data is to be predicted are 1) Association rule mining 2) Apriori. Apriori is used for mining frequent item-sets [3]. This algorithm uses prior knowledge of frequently occurring data in the dataset. The following lines state the steps in generating frequent item set in Apriori algorithm [3].

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
- 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 Lk has been achieved [3].

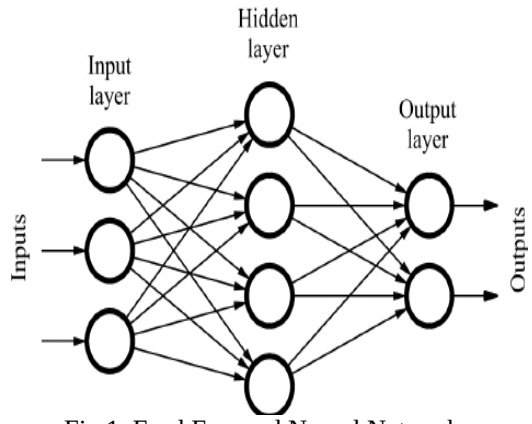

Fig 1. Feed Forward Neural Network.

Association rule mining used in [3] is a method to determine interest relations between attributes in large database.

Let I = {I1, I2, I3….Im} be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction T□I. 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□T. An association rule is an implication of the form A□B, where A □I, B□I, and A□B =□[3].
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 [3].

This approach uses a dataset of 60 students 8 attributes were considered and multi-layered perceptron was applied on the given data. Validation technique known as 10 cross validation was applied and each attribute was removed subsequently. The attributes taken in [3] are:
1. Schooling Education.
2. Previous programming knowledge.
3. Father/Mother is educated.
4. Graduation%.
5. Attendance%.
6. Assignment%.
7. Unit Test%.
8. University result%.

By applying 10 cross fold validation on a dataset if all attributes of the dataset are considered then the accuracy of correctly classified data is 44.5%. Thus this accuracy is the measure of certainty and uncertainty in the dataset.

D. Multivariate Linear Regression

This study by Havan Agarwal mainly focuses on the past records of the student where his academic performance levels are considered. This relationship can be expressed accurately using Multivariate Linear Regression [4]. In [4] the algorithm uses past semester marks of student and marks scored by the senior batches to predict future of current batch. The dataset of 80 BE-IT students was gathered and was trained on only 60 of them, tested on a cross-validation set of 10 students. The error statistics in [4] were as follows:
Average error = 6
Accurate = 296
Erroneous = 124
Accuracy Rate = 70.48%
Comparative study was carried out between Bayesian classification and neural network in [4], results proved that neural network outperformed Bayesian classification. This was justified by the fact that neural network uses continuous values and for Bayesian classification it requires discrete values hence the results of neural network was much more accurate than the other one.

The training dataset was increased in increments of 10. The test data was of 10 students to predict a single subject. The accuracy results were as follows:

TABLE 3. NEURAL NETWORK ACCURACY

Training Dataset Size	Accuracy
40	50 %
50	50 %
60	60 %
70	70

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

III. CONCLUSION

Present studies carried out over student academic performance prediction shows that the results of the predictive system is totally based upon the past experiences or past performance. Our survey confirms that performance of neural network as compared to other algorithms mentioned in the paper like decision trees, Multi Linear regressionand Apriori yields much better results in terms of accuracy, consistency and scalability. The survey also proves that performance factor of neural network increases with the increase in dataset size as this algorithm takes continuous values.
This study can also be carried out with the algorithm C4.5 and C5.0 which is classification algorithm but uses continuous values rather than discrete values like neural networks. The work can be tested by using limited dataset as the complexity of classifier is higher than that of other algorithms. Machine learning can contribute a powerful predictive tool which can help the various institutions and as well as students to evaluate themselves at initial level of their academic phase.

Further, we are continuing this research by using other Machine Learning Algorithm such as ID3 (Iterative Dichotomiser 3) which works on only discrete values. We would mainly be focusing on the attribute selection and ordered weightage of each attribute for a dataset provided by Sinhgad College. This study will help the student to analyze his/her monthly performance and also generate overall college performance for the college authorities to take respective measures.

ACKNOWLEDGMENT

We avail this opportunity to express our deep sense of gratitude and whole hearted thanks to our guide Ms. Pankaja Alappanavar for her valuable guidance and encouragement.

REFERENCES

[1] Xin Chen, Mihaela Vorvoreanu and Krishna Madhavan, “Mining Social Media Data for Understanding Students’ Learning Experiences,”IEEE transactions on learning technologies.
[2] Fabio Pagnotta, Hossain Amran, “Using Data Mining to Predict Secondary School Student Alcohol Consumption”.
[3] Suchita Borkar and K.Rajeswari, “Attributes Selection for Predicting Students’ AcademicPerformance using Education Data Mining and Artificial Neural Network”International Journal of Computer Applications (0975 – 8887),Volume 86 – No 10, January 2014.
[4] Havan Agrawal andHarshil Mavani, “Student Performance Prediction using Machine Learning”,International Journal of Engineering Research & Technology (IJERT),ISSN: 2278-0181,Vol. 4 Issue 03, March-2015.

BIOGRAPHIES

Prateek Sakaray, Pursuing Bachelor of Engineering (B.E) in Information Technology from Sinhgad Academy of Engineering, Savitribai Phule Pune University (S.P.P.U)

Snehal Kankariya, Pursuing Bachelor of Engineering (B.E) in Information Technology from Sinhgad Academy of Engineering, Savitribai Phule Pune University (S.P.P.U)

Chandini Lulla, Pursuing Bachelor of Engineering (B.E) in Information Technology from Sinhgad Academy of Engineering, Savitribai Phule Pune University (S.P.P.U)

Yash Agarwal, Pursuing Bachelor of Engineering (B.E) in Information Technology from Sinhgad Academy of Engineering, Savitribai Phule Pune University (S.P.P.U)

Pankaja Alappanavar, ME Computer Engineering, Sinhgad Academy of Engineering Pune