Analyzing the Student Performance using Classification Techniques to find the better Suited Classifier

Sherine Dominick
Research Scholar
Jamal Mohamed College (Autonomous)
Tiruchirappalli-20

T. Abdul Razak, PhD
Associate Professor
Jamal Mohamed College (Autonomous)
Tiruchirappalli-20

ABSTRACT
Educational institutions are producing talented and intelligent students and workers, but when we consider quality and equity of the student's progress in his career, it is still a challenge or a question to be answered. These institutions focus on quality in education. Every year a huge number of students graduate from colleges and universities, with respect to the data collected from the feedback of students, classification a data mining technique is applied to it. It is a step to analyze the factors affecting the academic performance of students in order to evaluate the current student performance and take efficient steps in the prediction of the most likely occurring relationships between the various aspects of learning and to enhance the quality of education in future and help the educational planners to plan accordingly.

Keywords
Data Mining, Classification, Student Performance, Prediction.

1. INTRODUCTION
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.

Whether educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of meaningful hierarchy, which is often needed to be determined by properties in the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data.[1]

This research is a step to analyze the factors affecting the academic performance of students using the data mining technique classification in order to evaluate the current performance and take efficient steps to enhance the quality of education.

1.1 Motivation
This research draws its inspiration from the varying learning patterns among the students. This will help the educational institutions to identify the students who are at risk and to take necessary steps to reduce failing ratio at right time to improve the quality of education; this research is based on the collective data summarized by V.Ramesh, P.Thenmozhi, Dr.K. Ramar [2].

1.2 Research Objectives
It is an attempt to find if the external factors or hereditary factors have an impact on the academic or scholastic performance of a student using data mining techniques.

2. RELATED WORK
The idea of using data mining in higher education has been put forward by many researchers and authors who have explored and discussed the performance of several students.

V.Ramesh, et al.[2] have attempted to find suitable prediction techniques using data mining tool WEKA to enhance the quality of the higher educational system.

Guan Li [3] has compared the accuracy of data mining methods to classifying students in order to predict student's class grade.

J.F. Superby [4] conducted a study to investigate to determine the factors to be taken into account we will use a model adapted from that of Philippe Parmentier (1994). In other words the idea is to determine if it is possible to predict a decision variable using the explanatory variables which we retained in the model.

Bray[5] in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India. was relatively higher than in Malaysia, Singapore, Japan, China and Srilanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

Ramaswami[6] in his study on CHAID based performance prediction model, observed that the CHAID prediction model was useful to analyse the interrelation between variables that are used to predict the outcome on the performance at higher secondary school education. V.O.Oladokun[7] in his study on predicting student's academic performance using artificial neural network, observed that Multilayer Perception Topology was best to predict the performance of more than 70% of prospective students.

3. CLASSIFICATION
Classification is the process of learning a model that describes different classes of data. The classes are predetermined. Classification consists of assigning a class label to a set of unclassified cases.

3.1 Supervised Classification
The set of possible classes is known in advance.

3.2 Unsupervised Classification

Set of possible classes is not known. After classification we can try to assign a name to that class. Unsupervised classification is called clustering.

3.2.1 Building the Classifier or Models

This step is the learning step or the learning phase.

In this step the classification algorithms build the classifier. The classifier is built from the training set made up of database tuples and their associated class labels.

Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.[8]

Method

Through extensive search of the literature and discussion with experts on student performance, a number of socio- economic, environmental, academic, and other related factors that are considered to have influence on the performance of a university student was identified. These factors were carefully studied and harmonized into a manageable number suitable for computer coding within the context of the familiar algorithms. These influencing factors were categorized as input variables. The output on the other hand represents some possible levels of performance of a candidate in terms of the present college grading system.[2]

3.2.2 Using Apriori Algorithm

The basic algorithm for finding the association rules was first proposed in 1993. In 1994, an improved algorithm the Apriori algorithm [10] was proposed. The algorithm may be considered to consist of two parts. In the first part, those itemsets are called frequent itemsets. In the second part the association rules that meet the minimum confidence requirement are found from the frequent itemsets[11]. The second part is relatively straight forward, so much of the focus of the research in this field has been to improve the first part.[8]

$$
\begin{aligned}
&\text{Apriori}(T, \epsilon) \\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
&\quad k \leftarrow 2 \\
&\qquad \textbf{while } L_{k-1} \neq \emptyset \\
&\qquad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1}\} \\
&\qquad\quad \textbf{for } \text{transactions } t \in T \\
&\qquad\qquad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
&\qquad\qquad \textbf{for } \text{candidates } c \in C_t \\
&\qquad\qquad\quad count[c] \leftarrow count[c] + 1 \\
&\qquad\quad L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\} \\
&\qquad\quad k \leftarrow k + 1 \\
&\qquad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

Fig 1: Apriori algorithm

The data set chosen for this implementation is prepared by collecting data from the students directly. The data collected from the students are organized and put together in a form ready for processing.

3.2.3 Observations of Apriori Algorithm on the Dataset

The following rules can be framed based on the result of Apriori execution:

1. Atleast 50% of ug students recieve scholarship (because total no of students is 491)

2. Most of BC students recieve scholarship (see that no scholarship is 242)

3. Only very few of bc studens dont recive scholarship

4. PROPOSED WORK

The data mining technique classification is applied on the data collected in order to group the like classes together.

5. IMPLEMENTATIONS

The research uses WEKA and is based on the dataset collected by the students. The dataset to be used in WEKA is in an .arff or .csv file. The dataset, consists of 491 instances with 30 attributes, out of which 12 attributes are considered for this manipulation. The attributes denote the various factors that affect the students performance. The dataset is preprocessed, K-Means clustering technique is applied, and its effect is studied.

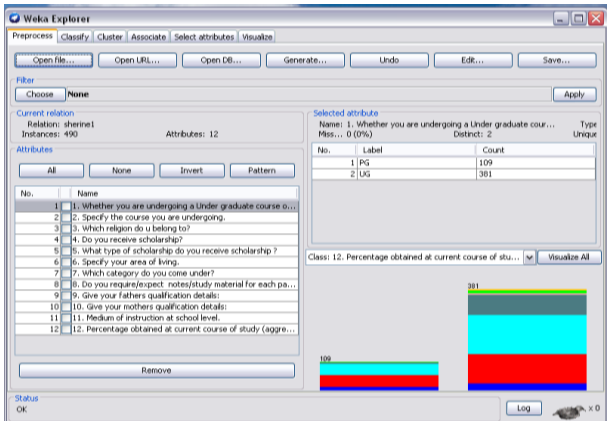

Fig 2: The imported dataset with its attributes.

This figure Fig.2 describes the elements of the dataset along with their attributes.

Fig 3: Visualization of the data set.

Fig.3 shows the visualization of the entire data set with respect to its attributes.

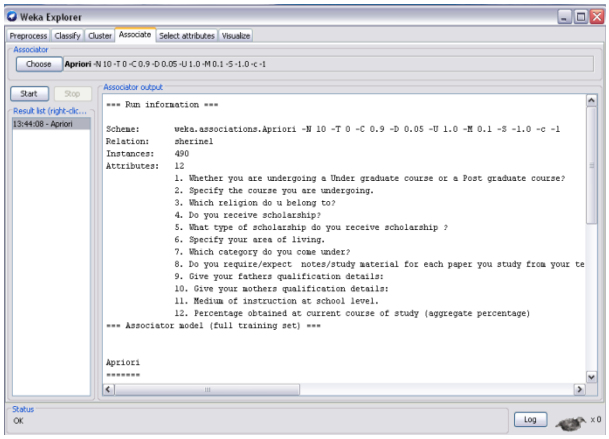

Fig 4: Association among the tuples.

In Fig.4 The associate tab is selected and the Apriori algorithm is chosen.

The total number of instances taken is 490, the chosen attributes are 12 out of the collected 30 attributes. The minimum Support is 0.35 and the confidence is 0.9. The number of cycles in which the execution is performed is 13.

Table 1: Using zeroR classifier.

Evaluation summary	**Observation**
Correctly Classified Instances	189
Incorrectly Classified Instances	301
Percentage of Correctly Classified Instances	38.5714%
Percentage of Incorrectly Classified Instances	61.4286%
Kappa statistic	0
Mean absolute error	0.2047
Root mean squared error	0.3194
Relative absolute error	100%
Root relative squared error	100%
Total Number of Instances	490

The Table. 1 shows how the training set elements are being classified using zeroR classifier. The correctly classified instances are 189 ie) 38.5714% and the incorrectly instances are 301 ie) 61.4286%.

Table 2: Using NavieBayes classifier.

Evaluation summary	**Observation**
Correctly Classified Instances	245
Incorrectly Classified Instances	245
Percentage of Correctly Classified Instances	50%
Percentage of Incorrectly Classified Instances	50%
Kappa statistic	0.2592
Mean absolute error	0.1801
Root mean squared error	0.3038
Relative absolute error	88.0223%
Root relative squared error	95.1101%
Total Number of Instances	490

The Table. 2 shows how the training set elements are being classified using NavieBayes classifier. The correctly classified instances are 245 ie) 50% and the incorrectly instances are 245 ie) 50%.

6. CONCLUDING REMARKS

There is significant variation in the classification process, as the numbers of incorrectly classified instances according to the different algorithms chosen as expected. Based on the results of the classification process, we find that the NavieBayes algorithm is more efficient when compared to the ZeroR classifer the incorrectly clustered instances are less in in NaviesBayes classifier which prove that any predictions made using the data will prove to be more precise and accurate. This would help to predict the impact of the factors on the student effectively.

7. REFERENCES

[1] International Educational Data Mining Society: http://www.educationaldatamining.org/

[2] V.Ramesh, et al., Study-of-influencing-factors-of-academic-performance-of-students-A-data-mining-Approach, International Journal of Scientific & Engineering Research, Volume 3, Issue 7, July-2012 ISSN 2229-5518.

[3] Guan Li, Liang Hongjun. Data warehouse and data mining. Microcomputer Applications. 1999, 15(9): 17-20.

[4] Adriaans P, Zantinge D. Data mining [M]. Addision_Wesley Longman, 1996.

[5] Rosemary Win and Paul W. Miller: The Effects of Individual and School Factors on University Students' Academic Performance

[6] Chen Rong, BP arithmetic and its structure optimization tactics. Journal of Autoimmunization.1997, 23(1), 43-49.

[7] Data Mining Classification & Prediction - Tutorialspoint: http://www.tutorialspoint.com/data_mining/dm_classification_prediction.html

[8] Jiawei Han, Micheline Kamber, and Jian Pei,"Data mining Concepts and Techniques", 2nd ed, Morgan Kaufmann, 2006.

[9] C. Romero, S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 40(6), 601-618, 2010.

[10] Shikha Maheshwari and Pooja Jain. The Research on Top Down Apriori Algorithm using Association Rule. IJARCSSE Volume 4, Issue 4, April 2014 http://www.ijarcsse.com/docs/papers/Volume_4/4_April2014/V4I4-0403.pdf

[11] Divya Jain and Sumanlata Gautam. Implementation of Apriori Algorithm in Health Care Sector: A Survey. International Journal of Computer Science and Communication Engineering Volume 2 issue 4 (November 2013 issue). http://static.ijcsce.org/wp-content/uploads/2013/12/IJCSCE110513.pdf