International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Enhanced Student Bunglers Detection using Association Rules and Predicting Outliers

Devikala. D[1]
Research Scholar1,
Department of Computer Science,
Dr. SNS Rajalakshmi College of Arts & Science,
Coimbatore – 641 049, India.

Kamalraj. N[2]
Head of the Department,
Department of Computer Technology
Dr. SNS Rajalakshmi College of Arts & Science,
Coimbatore-641 049,India.

***Abstract* - Recently many countries show interest and concern about problem of failure students and the way to determine the main contributing factors that affects the student's performance. The great deal of research is undergoing for identifying the factors for the low performance of students using the large amount of information stored in databases.**

This paper proposes a novel classification approach with association rule mining and outlier detection. Data mining is applied after preprocessing the data and continues with association, classification and outlier detection. The main objective of the paper is to detect dropout and failure data as early as possible which shows the factors trying to reduce dropout and failure students. The outcomes are compared and best result is identified.

Keywords: Educational data mining (EDM), Classification, Association, Outlier detection.

I. INTRODUCTION

The innovation of information technology from various disciplines such as database technology, scientific data, machine learning, neural networks, information retrieval, statistics, etc leads to usage of large volumes of data storage in various formats like records, files, documents, images, sound, videos and many new data formats. The process of identifying meaningful patterns and relationships of a data within very large databases is data mining and it is also called as KDD-knowledge discovery in databases. The steps involved before carrying out data mining are data cleaning, data selection, and pre-processing and data transformation.

The great deal of research [1] has been done on identifying the factors that affect the low performance of students at different educational levels using the large amount of information that current computers can store in databases. Current area of research in educational data mining is based on the development of methods for the better understand about students and the settings in which they learn [2]. The works show promising results with respect to, economic, sociological, educational characteristics which are more relevant in prediction of low academic performance [8] with some complexity of time and process by using various classification based algorithms. This paper proposes apriori algorithm in association rule mining for classification which provides more efficient results that the existing system. It reduces the complexity of the system and the extreme data that is the data which is abnormal is detected by the outlier detection method. Density based outlier is used to detect the abnormal data. The result produced by the system is more accurate takes less time complexity and provides better performance.

II. LITERATURE REVIEW

Romero.c et al [2] studies about the educational data mining and the development of the studies by exploring the data. The paper deals with the introduction of the educational data mining with different types of user groups and types of educational environment of the user group which provides the data. The most common task by data mining technique to resolve the educational environment is listed out and finally some promising features are discussed.

N. V. Chawla et al [3] proposed a method of over-sampling the abnormal class and under-sampling the normal class can achieve better classifier performance by varying the loss ratios in class.

S. Kotsiantis et al [4] studies about the various methodologies that have been proposed for the betterment of failure students in the academics. The author proposes a local cost sensitive technique and concludes the framework which is more effective solution for the problem.

M.N.Quadril et al [6] studies about the work of data mining in predicting the drop out feature of students. He proposed decision tree technique for choosing the best prediction and analysis about the features of failure students. The author produces the lists that are predicted as likely to drop out of students from college that are handled by the management and teachers.

III. METHOD

This paper proposes a method for predicting the academic student failure belongs to the process of Knowledge discovery and Data mining. The stages of the method are:

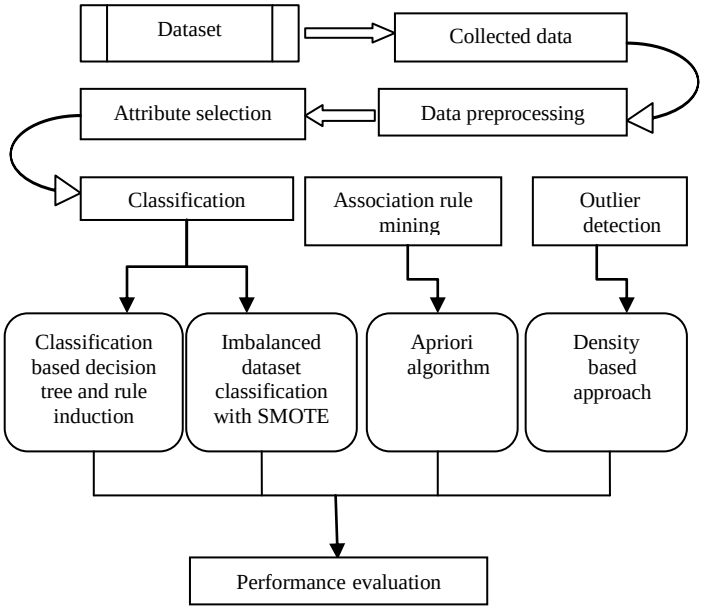

Figure: 1.Method Proposed for Prediction of student failure

1. Data Gathering: All the available information about the student is gathered from which the set of factors that affects the student are identified and is fed into the dataset.
2. Pre-processing: In this stage before applying data mining algorithm the gathered data must be cleaned, transformed to variables with required selection, integrated properly and solve the problems in the dataset.
3. Classification based decision tree and rule induction: In this stage, Data mining algorithm is applied to predict the student failure. The "white box" technique is preferred for generating easily interpretable models. Classification method followed based on decision tree which is organized in a hierarchical structure. Rule induction is used to avoid many problems associated with decision trees.
4. Imbalanced dataset classification with SMOTE: SMOTE (Synthetic Minority Oversampling Technique) is used to solve the problem of imbalanced data classification which occurs when the number of instances in one class is much smaller than the number of instances in another class. It is activated during the pre-processing of data for balancing the class distribution.
5. Association rule mining: Association rule mining is used to find the relation between variables in the large database. Various algorithms are used for association rule mining. Apriori algorithm is used in this system. It follows the breath-first search type to count the support of the itemsets and use candidate generation function for the process [9].
6. Outlier Detection: The data objects in the database that does not have general behavior that of normal data is called outliers. It is detected with outlier detection method and in this paper density based outlier detection method is used [10].
7. Interpretation: The obtained models are analyzed to detect the failure student in the database.

IV. PROPOSED ALGORITHM

4.1. 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. Association rule Mining is mainly intended to recognize strong rules from databases based on confidence and different measures support. The preliminaries required for performing data mining on any data are discussed below.

Let I= {I1, I2, I3….Im} be the set items and let *D*, be the task relevant data, a set of database transactions where each transaction $T \subseteq I$. 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*, only if $A \subseteq T$. Association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cup B = \emptyset$.The rule interest is of measures that are Support (s) and confidence (c). The result reflects the usefulness and certainty of the discovered rule. A support of 2% of the rule $A \Rightarrow B$ means that A and B exist together in 2% of all the transactions under analysis. The rule $A \Rightarrow B$ having confidence of 60% in the transaction set D means that 60% is the percentage of transactions in D containing A which also contains B. A set of items is also referred to as an item set. An item set which contains k items in it is also called as k-item set. The number transaction of the itemset is the occurrence frequency of the itemset. If I the relative support of an itemset get satisfied the minimum support of threshold, then it is a frequent item set. Association rule mining follows two-step process:

1) To *d*iscover all frequent itemsets: Each itemset will occur at least as frequently as a predetermined minimum support count.

2) Create strong association rules from frequent itemsets: The rules must always satisfy minimum support and confidence and the rules are called as strong rules.

3.2.2 Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agarwal and R. Srikant for mining frequent itemsets for Boolean association rules. The algorithm is based on prior knowledge of frequent itemset properties. The steps in generating frequent item set in Apriori algorithm are.

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

• Find frequent set Lk-1

• Join step: Ck is generated by joining Lk -1 with itself (Cartesian product Lk-1 x Lk-1)

• Prune step (apriori property): Any (k − 1) size itemset is not frequent and cannot be a subset of a frequent k size itemset, hence it should be removed.
• Frequent set Lk has been achieved.

Figure: 3 depicts association rules discovered from data of students with grade, with their support, confidence, and lift

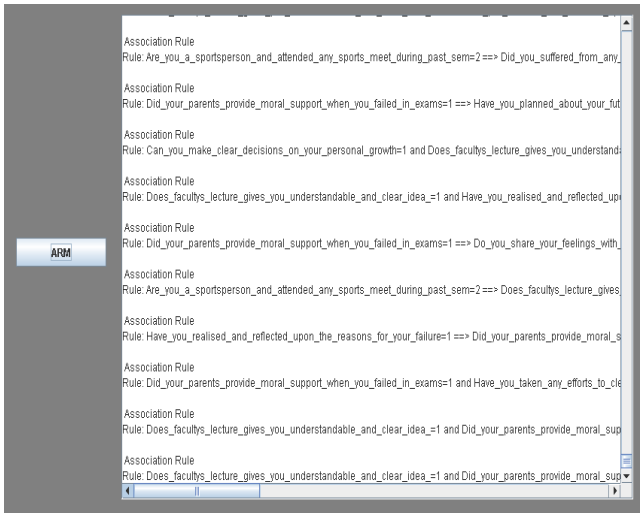

Figure: 2. Association Rule Mining

4.2. Outlier detection

A database may contain data objects that do not comply with the general behavior of the data and are called outliers. The analysis of these outliers may help in fraud detection and predicting abnormal values. The data stored in a database may reflect outliers/noise, exceptional cases, or incomplete data objects. The incomplete data objects may confuse the analysis process which cause over fitting of data to the knowledge model constructed. So, as a result, the accuracy of the discovered patterns can be poor. The abnormal values in the result sheet of the students are detected by an application of outlier analysis. This may be due to many factors like data entry operator negligence, software fault, or an extraordinary performance of the student in a particular subject.

Outlier detection discovers data points that are significantly different than the rest of the data. In educational data mining outlier analysis can be used to detect students with learning problems. In this paper, we used outlier analysis to detect outliers in the student data. Outlier methods are used in this work which is 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. Intuition (density-based outlier detection): The density around an outlier object is significantly different from the density around its neighbors.

Method: The relative density of an object is used against its neighbors as an indicator of the degree of the object being the outliers.

- *k-distance* of an object o, distk(o): distance between o and its k-th NN
- *k-distance neighborhood* of o, Nk(o) = {o'| o' in D, dist(o, o') ≤ distk(o)}
- Nk(o) could be bigger than that of k since multiple objects may have identical distance to o.

Reachability distance from *o'* to *o*:

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

– where k is a user-specified parameter.

• Local reachability density of *o*:

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

• LOF (Local outlier factor) of an object o is the average of the ratio of local reachability of *o* and those of *o*'s k-nearest neighbors.

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$$

• The lower is the local reachability density of o, and the higher is the local reachability density of the kNN of o, and the higher is LOF.

• A local outlier is captured whose local density is relatively low comparing to that of local densities of its kNN.

Figure: 3. depicts the outlier detection using local outlier factor for students.

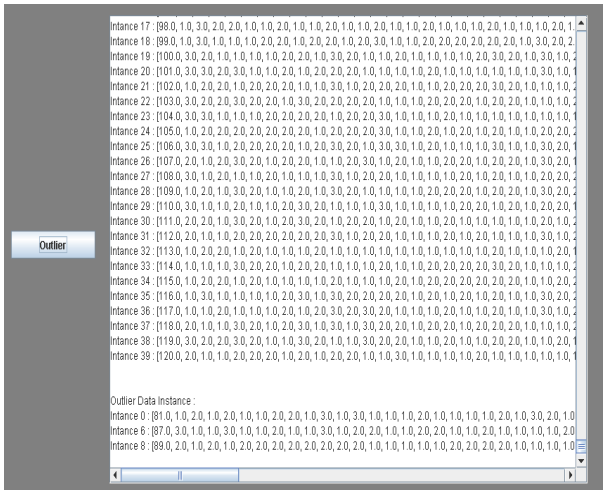

Figure: 3. Outlier Detection

V. PERFORMANCE EVALUATION

The performance of the existing classification and prediction system with proposed grammar based genetic programming approach to derive the pass/failure result are tested. Measure the performance results in terms of the true positive rate (TPR), False positive rate (FPR), False Negative Rate (FNR) and True negative Rate (TNR), accuracy, Time comparison.

We analyze and compare the performance offered by classification, classification with feature selection, imbalanced classification with SMOTE oversampling technique, and prediction using association rule mining, outlier detection approaches. The performance is evaluated by the parameters such as accuracy. Based on the comparison and the results from the experiment show the proposed approach works better than the existing system.

Accuracy

Accuracy is calculated from the below given formula as

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

- **TP (True positive)**

In a statistical hypothesis test, two types of incorrect conclusions can be drawn. The hypothesis can be inappropriately. A positive test results accurately reflects the test for the activity is analyzed. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP);

True positive rate (TPR) =TP/P

P= (TP+FN)

Where P is the positive. TP is the True Positive

- **TN (True negative)**

A result that appears negative when it should not. A true negative (TN) has occurred when both the prediction outcome and the actual value are n is the number of input data.

True negative rate (TNR) =TN/N

N= (TN+FN)

Where

N is the Negative value.

TN is the True Negative.

- **FP (False positive)**

A result that indicates that a given condition is present when it is not. However if the actual value is n then it is said to be a false positive (FP).

False positive rate (α) = FP / (FP + TN)

- **FN (False negative)**

False negative (FN) is when the prediction outcome is n while the actual value is p.

False negative rate (β) =FN / (TP + FN)

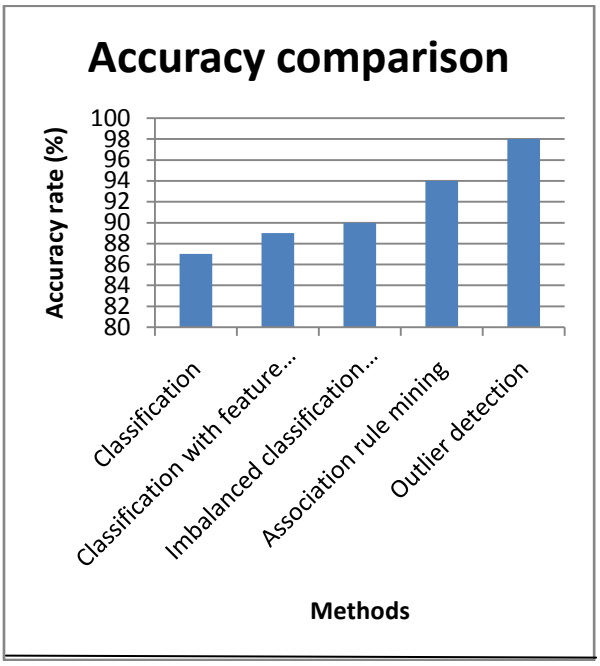

Figure: 4. Shows Accuracy Comparison graph

The graph shows the accuracy rate of existing system such that classification, classification with feature selection, imbalanced classification with SMOTE oversampling technique, and proposed system such as prediction using association rule mining, outlier detection approach using density based approach based on two parameters of accuracy and methods such as existing and proposed system. From the graph we can see that, accuracy of the system is reduced somewhat in existing system than the proposed system. From this graph we can say that the accuracy of the proposed system is increased which will be the best one.

VI.CONCLUSION AND FUTURE WORK

The aim of this system is to analyze the factor that affects the academic achievement of the students. It is useful in identifying weak students who are likely to perform poorly in their studies. Data mining and machine learning depend on classification which is the most essential and important task. An educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics. The various data mining techniques can be effectively implemented on educational data. From the results it is clear that classification techniques can be applied on educational data for predicting the student's outcome and to improve their performance for results. The efficiency of various decision tree algorithms is analyzed based on their accuracy and time to derive the tree. The predictions obtained from the system have helped the tutor to identify the weak students and improve their Performance. The classification accuracy and performance is high in the proposed system when compared to the existing system. The experimentation result gives the proposed system is more efficient than the existing system.

Finally, as the next step in our research can be carry out with more experiments using more data and also with different educational levels to test whether the same performance results are obtained with different DM approaches.

The future work continues as, to predict the student failure as soon as possible. To detect students risk in time before it is too late. To propose actions for helping students identified within the risk group. Then, to check the rate of the time to prevent the fail or dropout of that student previously detected.

REFERENCES

1. F. Araque, C. Roldán, and A. Salguero, "Factors influencing university dropout rates," *Comput. Educ.*, vol. 53, no. 3, pp. 563–574, 2009.
2. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.
3. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
4. S. Kotsiantis, "Educational data mining: A case study for predicting dropout—prone students," *Int. J. Know. Eng. Soft Data Paradigms*, vol. 1, no. 2, pp. 101–111, 2009.
5. Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto "Predicting School Failure and Dropout by Using Data Mining Techniques", IEEE journal of latin-american learning technologies, vol. 8, no. 1, february 2013.
6. M.N.Quadri, Dr.N.V.Kalyankar, " Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", Global Journal of Computer Science and Technology, Vol 10, No 2 (2010).
7. Devikala.D , Kamalraj.N, "Data Mining Approaches on Detection of Students' Academic Failure and Dropout: A Brief Survey", International Journal of Computer Trends and Technology (IJCTT) – volume 14 number 3 – Aug 2014
8. J. Más-Estellés, R. Alcover-Arándiga, A. Dapena-Janeiro, A. Valderruten-Vidal, R. Satorre-Cuerda, F. Llopis-Pascual, T. Rojo-Guillén, R. Mayo-Gual, M. Bermejo-Llopis, J. Gutiérrez- Serrano, J. García-Almiñana, E. Tovar-Caro, and E. Menasalvas-Ruiz, "Rendimiento académico de los estudios de informática en algunos centros españoles," in Proc. 15th Jornadas Enseñanza Univ. Inf., Barcelona, Rep. Conf., 2009, pp. 5–12.
9. http://en.wikipedia.org/wiki/Association_rule_learning#Apriori_algorithm
10. http://en.wikipedia.org/wiki/Local_outlier_factor