IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN:2278-8727, Volume 19, Issue 3, Ver. V (May-June 2017), PP 49-52
www.iosrjournals.org

PREDICTING STUDENT PERFORMANCE BASED ON CLUSTERING AND CLASSIFICATION

Snehal Bhogan[1], Kedar Sawant[2], Purva Naik[3], Rubana Shaikh[4], Odelia Diukar[5], Saylee Dessai[6]

[1, 2](Asst. Prof., Computer Engineering Department, AITD, Goa University, India)
[3, 4, 5, 6](B.E. Student, Computer Engineering Department, AITD, Goa University, India)

Abstract : *In today's world the education field is growing, developing widely and becoming one of the most crucial industries. The data available in the educational field can be studied using educational data mining so that the unseen knowledge can be obtained from it. In this paper, various data mining approaches like Clustering, classification and regression our used to predict the students' performance in examination in advance, so that necessary measures can be taken to improvise on their performance to score better marks. A hybrid approach of Enhanced K-strange points clustering algorithm and Naïve Bayes classification algorithm is presented implemented and compared it with existing hybrid approach which is K-means clustering algorithm and Decision tree. Finally, to predict student performance, multiple linear regression is used. The results obtained after the implementation may be useful for instructor as well as students. This work will help in taking appropriate decision to improve student's performance.*

Keywords - *classification, clustering, data mining, student prediction, regression*

1. INTRODUCTION

Students tend to drop out or have a significant decrease in its academic performance. By predicting student performance, instructors can help to improve student performance in the examination and significantly reduce drop out ratio from college, which will enhance the performance of college.

In paper [1], K-means clustering algorithm and Decision tree has been used to predict student performance. But K-means clustering algorithm has a limitation that in case clusters or centroid does not converge that it can go into infinite iteration hence in this work, Enhanced K-strange point clustering algorithm is used since iteration depends upon number of clusters. The disadvantage of Decision Tree classification algorithm is that it is not considering all the attributes of the dataset which is essential to predict student performance hence in this paper, Naïve Bayes classification algorithm is proposed as this algorithm considers all the attributes while computing the result.

2. LITERATURE SURVEY

In paper[1],K-means clustering algorithm is used to form the clusters. The algorithm was applied on the student training data set, then three clusters were formed namely "High", "Medium" and "Low", according to their new grade. The new grade is calculated from the previous semester grade that means external assessment and internal assessment. Then Decision tree was applied to make correct decisions about the student's performance, which can use by the instructor to take the necessary steps.

In paper [2], the student performance is prediction is carried out using K-means clustering algorithms and decision trees, the results and analysis was done in WEKA tool. K-means algorithm was applied on the same dataset using WEKA tool. The decision tree algorithm was used to do the prediction which was displayed in tree-like structure. 143 students were classified as passed and 30 as failed which was true as per the original dataset.

K-means clustering algorithm is used on the student's data and then students have been clustered based on their class performance, sessionals and attendance in class [3]. Centroids are calculated from the educational data set taking K-clusters. This study helps in identifying students who are short of attendance and have shown poor performance in sessionals.

Paper [4] provides an enhancement to K Strange points clustering algorithm by correcting the location of the third strange point by trying to place it almost maximally and equally spaced both from Kmin and Kmax.This results in more accurate clusters.

3. CLUSTERING

3.1. K-Means Clustering Algorithm

In Clustering technique the data points of the dataset is partitioned into homogeneous clusters. The clustering problem can be solved using a simple unsupervised learning algorithm called K-Means. It can be used when there is unlabeled data. Using K-means a given data set can be gathered into number of clusters. Each of the cluster has its own centroid. These centroids do not have a fixed position in the cluster. Centroids of different cluster should be as far as possible to obtain better result.

Basically, K-Means clustering algorithm partitions *n* data points of the dataset into *k* clusters wherein each data points belongs to cluster with the minimum mean value. These mean value can be calculated using Euclidean distance formula. The algorithm works iteratively to partition the data point of the data set.

3.2. Enhanced K-Strange Point Clustering Algorithm

The Enhanced K-Strange points clustering algorithm is about discovering strange points that are hugely disconnected from each other if not exactly equidistant.

This algorithm first finds the minimum of the dataset. This point is referred to as Kmin. It then finds the maximum distance from Kmin which is referred to as Kmax point. Next, the algorithm computes maximally separated third point from Kmax and Kmin from the dataset. This computation of separation is calculated using Euclidean distance formula.

4. CLASSIFICATION

4.1. Decision Tree

Decision tree classifies examples into simple representation wherein leaf of decision tree represents the class label. Decision tree requires gaining information or entropy for making the decision.Decision tree classifies a dataset into smaller subsets of the tree and labels the leaves of the tree. Branch or arc of the decision tree represents the attribute that is required in order to extract the label of the leaf.

Initially, entropy is determined in order to produce a decision tree. Decision tree generates a root node, internal nodes having two or more branches and a leaf node based on entropy that is computed at the start. Decision tree begins with root node and ends with leaf nodes.

4.2. Naïve Bayes Algorithm

Naïve Bayes classifiers consider all the parameters of the dataset to produce the result. It represents supervised learning method as well as a probabilistic model.

The independent effect of an attribute value on a given class with the values of the other attributes which is assumed by Naïve Bayes classifier is called class conditional independence. Naïve bayes classification is based on computing probability in order to determine class for the given sample.

5. PREDICTION

5.1 Multiple linear Regression

Multiple linear regression is the extension of simple linear regression. It is a regression model that contains more than one repressor variable. Multiple linear regression can predict only one value at a time having one or more independent variables. Independent variables are the variables that are used to predict the dependent variable. Dependent variable is the variable that is been predicted.

Initially, least square method is used to calculate the coefficient of the independent variables. After substituting the value obtained from least square method, Multiple linear regression equation is formed based on which predicted value is been calculated.

6. PROPOSED WORK

6.1 Existing hybrid approach

K-means clustering algorithm and Decision tree has been used to predict student performance. But K-means clustering algorithm has a limitation that in case clusters or centroid does not converge that it can go into infinite iteration. And in case of decision tree which has a disadvantage of not considering all the attributes of the dataset to predict student performance which is essential.

6.2 Proposed hybrid approach

Enhanced K-strange point clustering algorithm and Naïve Bayes classification algorithm has been proposed to predict student performance as it overcomes the disadvantage of K-means clustering algorithm and Decision Tree. To enhance the existing approach, multiple linear regression is used that predict the student's percentage for the last semester.

Dataset which is been used is a student database of batch 2012-2016 and batch 2013-2017 from college named Agnel Institute of technology and design.

Batch 2012-2016 student database is the training dataset on which K-means clustering algorithm and Enhanced K-strange point clustering algorithm is applied. Batch 2013-2017 student database is the testing dataset on which Decision Tree, Naïve Bayes classification and multiple linear regression is applied.

7. RESULT AND DISCUSSION

7.1 Result of Existing Hybrid Approach

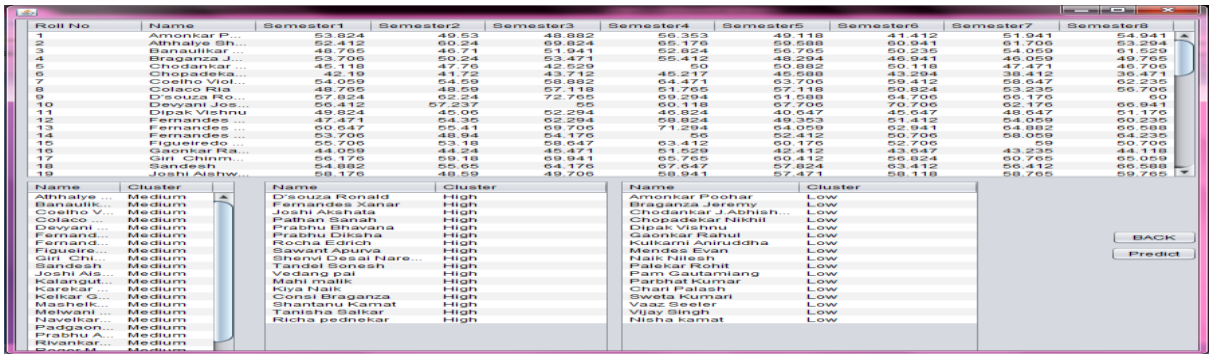

Fig1:- Implementation of K-means Clustering Algorithm

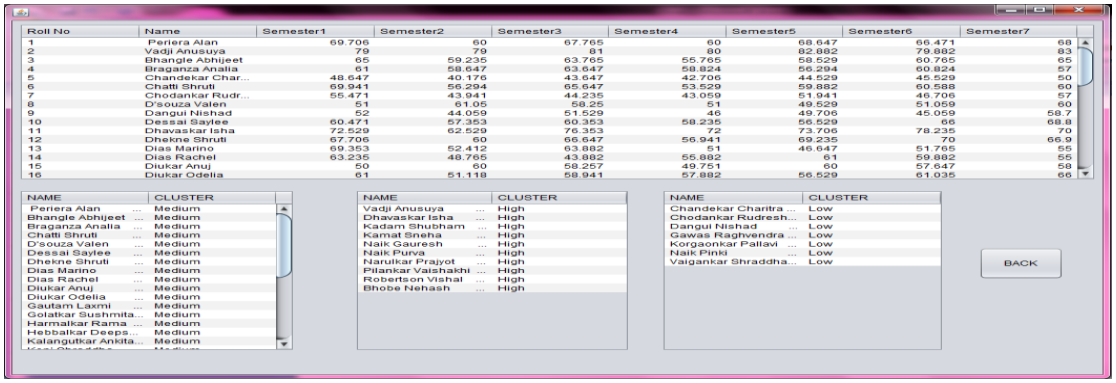

Fig2:- Implementation of Decision Tree

7.2 Result of Proposed Approach

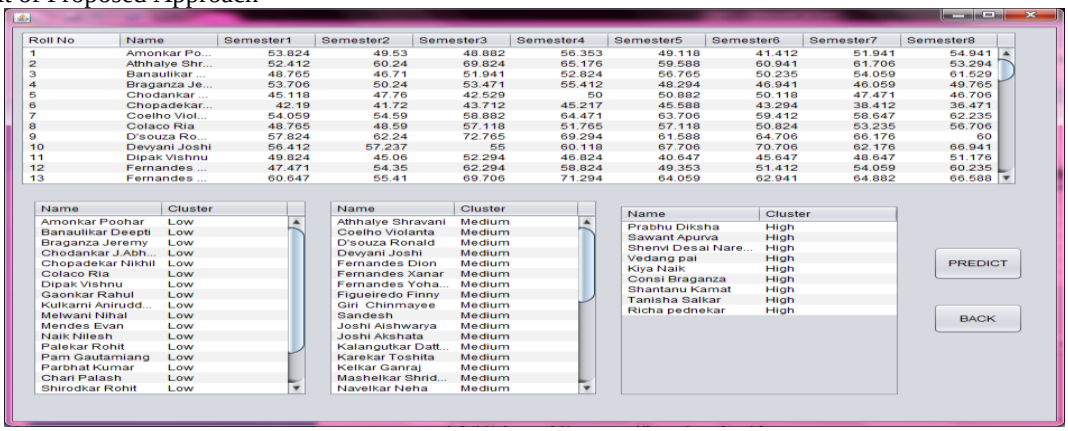

Fig3:- Implementation of Enhanced K-Strange Clustering Algorithm

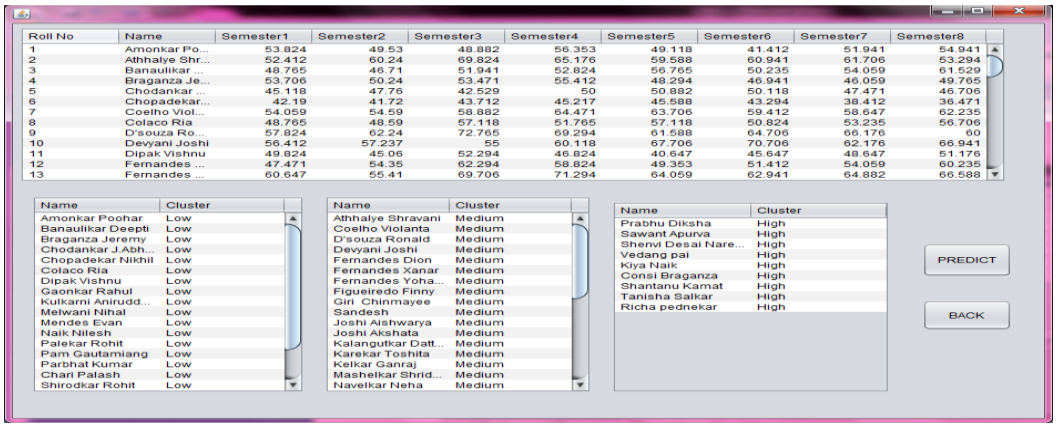

Fig4:- Implementation of Naïve Bayes Classification Algorithm

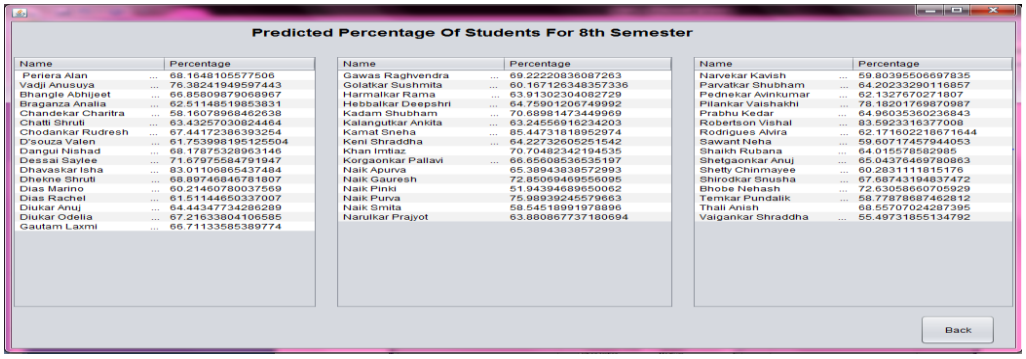

Fig5:- Implementation of Multiple Linear Regression

8. CONCLUSION

Existing approach proves to yield better result for the given data. While comparing the result of K-means clustering algorithm and Enhanced K-strange clustering point, there was observation that some of medium and low class tuple was assigned to high and medium class respectively in K-means Clustering algorithm which is not the case in Enhanced K-strange clustering point. Similarly, while comparing the result of Decision tree and Naive Bayes, there was an observation that class label was not accurate in decision tree as compare to naive bayes as decision tree did not consider all parameters of the attribute while computing the class label. From the result and analysis, ID3 Decision tree algorithm has a drawback of not defining new tuple's cluster if the tuple doesn't follow a particular range. It can be improved by using C4.5 Decision tree algorithm. Multiple Linear regression can help to predict only

one semester percentage of the students at a time. One of the drawbacks of multiple linear regression is the rank deficiency problem for which tuple in the particular cluster should be more than number of variables.

9. REFERENCES

Journal Papers:

[1] Md. Hedayetul Islam Shovon and Mahfuza Haque, An Approach of improving Student's Academic performance by using k-means clustering algorithm and decision tree, (IJASC) International Journal of Advanced Computer Science and Applications Vol.3, No. 8, 2012.

[2] Thaddeus Matundura Ogwoka, Wilson Cheruiyot and George Okeyo, A model for predicting student's Academic Performance using a Hybrid of k-means and decision tree algorithm, International Journal of Computer Applications Technology and Research, Vol.4, Issue 9, 693-697, 2015, ISSN:2319-8656.

[3] M. Durairaj and C. Vijitha, Educational Data mining for Prediction of Student Performance Using Clustering Algorithm, International Journal of Computer Science and Information Technologies, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol5(4), 2014, 5987-5991, ISSN:0975-9646.

[4] Terence Johnson and Dr. Santosh Kumar Singh, "Enhanced K- Strange Points Clustering Algorithm", 2015 International Conference on Emerging Information Technology and Engineering Solutions.

Books:

[5] Jiawei Han and Micheline Kamber, Data Mining - Concepts and Techniques, Second Edition, Original ISBN: 978-1-55860-901-3, Indian Reprint ISBN: 978-81-3120535-8