

PREDICTING STUDENT PERFORMANCE BASED ON CLUSTERING AND CLASSIFICATION

Snehal Bhogan¹, Kedar Sawant², Purva Naik³, Rubana Shaikh⁴,
Odelia Diukar⁵, Saylee Dessai⁶

^{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 are 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

Roll No	Name	Semester1	Semester2	Semester3	Semester4	Semester5	Semester6	Semester7	Semester8
1	Amonkar P...	53.624	49.53	48.882	56.353	49.118	41.412	51.941	54.941
2	Ahnaiva Sh...	52.412	60.24	69.824	66.176	58.888	60.941	61.706	63.294
3	Banaukar ...	48.766	40.71	51.941	52.824	50.766	50.235	54.059	51.529
4	Braganza J...	53.706	50.24	53.471	55.412	48.294	46.941	46.059	49.755
5	Chodankar ...	45.118	47.76	42.529	55.50	50.882	50.118	47.471	48.706
6	Chopadea...	42.19	41.22	43.712	48.217	45.588	43.412	36.471	38.471
7	Coelho Viol...	54.059	54.59	58.882	64.471	63.706	59.412	58.647	62.235
8	Costa Jis...	48.706	48.59	57.118	51.766	57.118	50.235	53.235	56.706
9	D'souza Ro...	57.824	62.24	72.766	55.294	61.588	64.706	65.176	60
10	Devyani Jos...	66.412	67.237	65	60.118	67.706	70.706	62.176	66.941
11	Dipak Vishnu	49.824	45.06	52.294	46.824	40.647	45.647	48.647	51.176
12	Fernandes ...	60.647	55.41	69.706	62.294	59.824	43.353	62.941	60.588
13	Fernandes ...	63.706	48.84	64.176	71.294	64.059	62.941	64.882	65.235
14	Figueredo ...	55.706	53.18	58.647	63.412	60.176	52.706	59	50.706
15	Figueira ...	44.059	44.24	45.471	51.529	42.412	43.647	43.235	44.118
16	Giri Chinn...	56.176	59.18	69.941	65.766	60.412	58.824	60.766	65.059
17	Sardesh...	54.882	55.66	64.176	67.647	67.824	63.412	66.412	66.588
18	Joshi Aishw...	58.176	48.59	49.706	58.941	57.471	58.118	58.766	59.766

Name	Cluster	Name	Cluster	Name	Cluster
Ahnaiva ...	Medium	D'souza Ronald	High	Amonkar Poonar	Low
Banaukar ...	Medium	Fernandes Kanar	High	Braganza Jeremy	Low
Coelho V...	Medium	Joshi Anshata	High	Chodankar J Abhis...	Low
Costa ...	Medium	Pathan Sarah	High	Chopadekar Nikhil	Low
Devyani ...	Medium	Prabhu Divyana	High	Dipak Vishnu	Low
Fernand ...	Medium	Prabhu Divyana	High	Georgina Rahul	Low
Figuered ...	Medium	Rochea Edrich	High	Kulkarni Anshuddha	Low
Figueira ...	Medium	Saward Saurya	High	Nairdes Eshw	Low
Giri Chi ...	Medium	Shreri Desai Nare...	High	Naik Nilesh	Low
Sardesh ...	Medium	Tandil Suresh	High	Naik Rohit	Low
Joshi Ais...	Medium	Vedang pal	High	Pam Gautamiang	Low
Kanakar ...	Medium	Mohi Naik	High	Parbat Kumar	Low
Karkar G...	Medium	Kiya Naik	High	Charl Palash	Low
Machhel...	Medium	Conel Braganza	High	Vaaz Seeler	Low
Mehwan ...	Medium	Shantanu Kamat	High	Vijay Singh	Low
Navelkar...	Medium	Tantana Sakar	High	Nisha kamat	Low
Padgorn ...	Medium	Richa pedrekar	High		
Prabhu A...	Medium				
Rivankar...	Medium				
Sardesh...	Medium				

Fig1:- Implementation of K-means Clustering Algorithm

Roll No	Name	Semester1	Semester2	Semester3	Semester4	Semester5	Semester6	Semester7
1	Perera Alan	69.706	60	67.766	60	68.647	66.471	68
2	Vadji Anusuya	79	79	81	80	82.882	80.982	80
3	Bhangle Abhijeet	65	59.235	63.766	55.766	58.529	60.766	65
4	Braganza Analia	61	58.647	63.647	58.824	66.284	60.824	67
5	Chandekar Char...	48.647	40.176	43.647	42.706	45.529	45.529	50
6	Chatt Shrut	69.941	56.294	65.647	53.529	59.882	60.588	60
7	Chodankar Rudr...	55.471	43.941	44.235	43.059	51.941	46.706	57
8	D'souza Valen	51	61.05	58.26	51	49.529	51.059	60
9	Dangul Nishad	52	44.059	51.529	46	49.706	45.059	58.7
10	Desai Saylee	60.471	67.353	60.353	58.235	66.529	66	68.8
11	Dhavaskar Isha	72.529	62.529	76.353	72	73.706	78.235	70
12	Dhekne Shruti	67.706	60	66.847	56.941	69.235	70	66.9
13	Dias Marino	69.353	62.412	63.882	61	46.647	51.766	66.9
14	Dias Rachel	63.235	48.766	43.882	55.882	61	59.882	55
15	Diukar Anuj	50	60	58.257	49.751	60	57.647	58
16	Diukar Odelia	61	51.118	58.941	57.882	56.529	61.035	66

NAME	CLUSTER	NAME	CLUSTER	NAME	CLUSTER
Perera Alan	Medium	Vadji Anusuya	High	Chandekar Charitra	Low
Bhangle Abhijeet	Medium	Dhavaskar Isha	High	Chodankar Rudresh...	Low
Braganza Analia	Medium	Kadam Shubham	High	Gawas Radhendra	Low
Chatt Shrut	Medium	Kamat Sneha	High	Korgaonkar Pallavi	Low
D'souza Valen	Medium	Naik Gauresh	High	Naik Pinki	Low
Desai Saylee	Medium	Naik Purva	High	Vaigantkar Shradha...	Low
Dhekne Shruti	Medium	Nanekar Prayot	High		
Dias Marino	Medium	Plankar Vaishakhi	High		
Dias Rachel	Medium	Robertson Vishal	High		
Diukar Anuj	Medium	Bhobe Nehash	High		
Diukar Odelia	Medium				
Gautam Laxmi	Medium				
Colatkar Dushmita	Medium				
Harmalkar Rama	Medium				
Hebbalkar Deepa	Medium				
Kalanekar Anshita	Medium				

Fig2:- Implementation of Decision Tree

7.2 Result of Proposed Approach

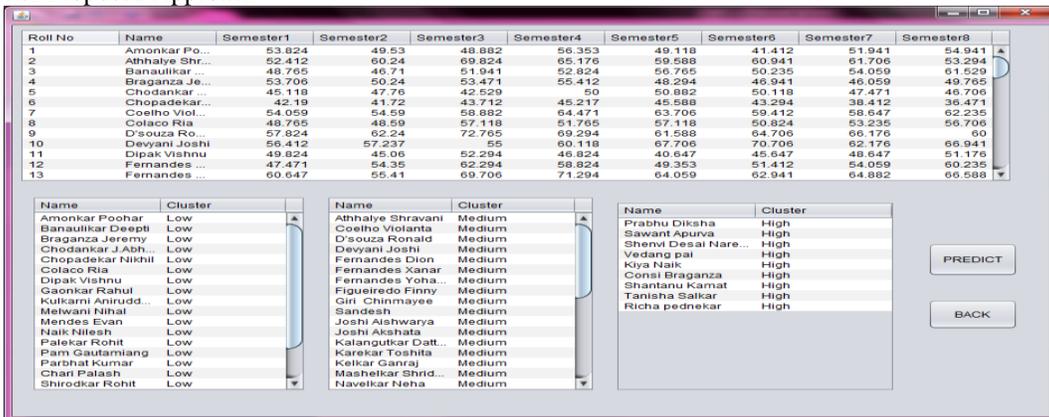


Fig3:- Implementation of Enhanced K-Strange Clustering Algorithm

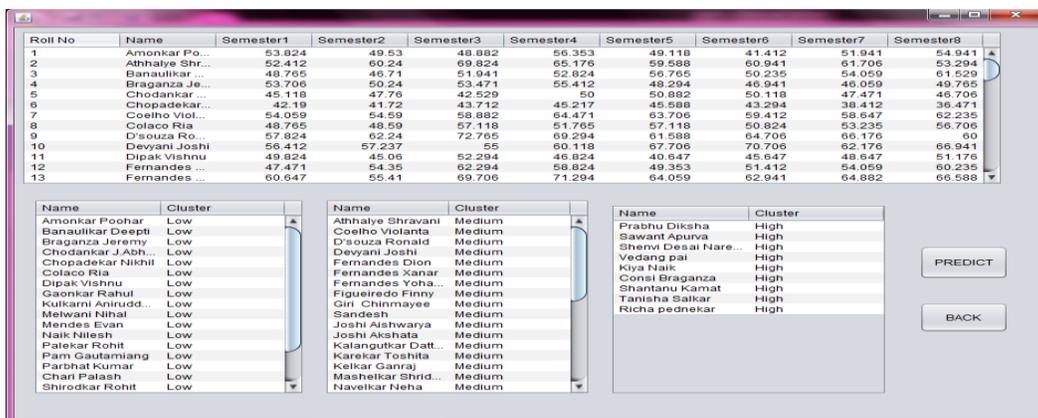


Fig4:- Implementation of Naïve Bayes Classification Algorithm

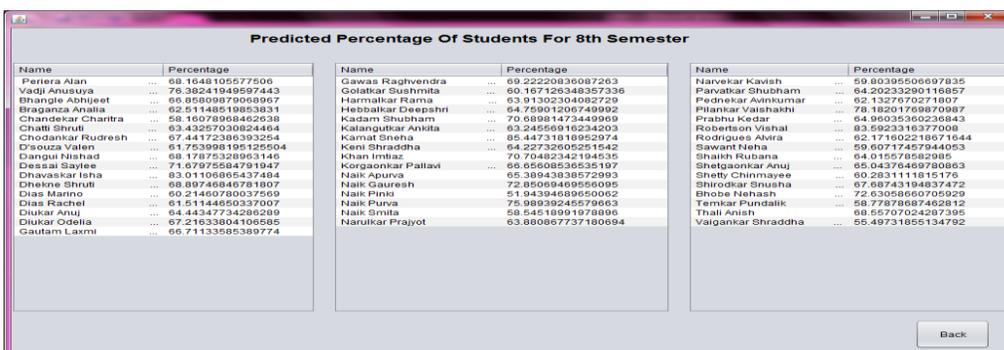


Fig5:- Implementation of Multiple Linear Regression

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