International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7493-7498

An Effective Cure Clustering Algorithm in Education Data Mining Techniques to Valuate Student's Performance

[1] Manjula . V. and [2] A. N. Nandakumar

[1], Research Scholar, School of Engineering and Technology, Jain University, Bangalore, India.

[2], Professor, Computer Science and Engineering, GSSSIT for Women, Mysore, India.

Abstract

Current issues identified in Educational data mining(EDM) especially engineering student performance occurs mainly due to the large volume of data in educational databases. Other issues identified in EDM mainly incorporate identification or prediction of weak engineering student's performance. Various researches have been done so far for predicting the performance of weak students to make improvements in their performance. But in most research works, only few attributes like grades, results, assignments, gender, internal marks are considered in order to predict the students' performance. Though the teacher's maintain the performance of their students it is not correct in all cases. So a better mining algorithm has to be implemented to successfully identify the behavior of weak students. In most of the works, the attributes identified are irrelevant and are neglected as missing attributes leading to inconsistent results. The next drawback identified with the existing clustering algorithm, is the similarity achieved between the cluster partitions. Also with the increased dataset, the hierarchical clustering algorithm breaks down due to non-linear time complexity. In order to reduce this drawback an efficient Adaptive Clustering Using Representatives (CURE) Algorithm will be proposed in this work. Further this clustering improvement will also provide improvements in classification results.

Keywords: Dataset, Data Reduction, Random Sampling, Partitioning, CURE Clustering.

INTRODUCTION

EDM is a process of learning analytics which is described as a set of methods that applies data mining and machine learning techniques such as prediction, classification, and discovery of latent structural regularities .In education the ability to predict a student's final performance has gained increased emphasis. In practical, the students' performance prediction is used by the instructors to monitor students' progress and to recognize at-risk students for providing timely interventions. In educational field, data mining techniques are used to enhance the understanding of learning process to focus on identifying, extracting and evaluating variables corresponding to learning process of students. Hence a student performance prediction model has to be built which is both practical and understandable for users and is a challenging task fraught with confounding factors to collect and measure. Most of the current prediction models are difficult for teachers to interpret. The prediction issue occurs due to absence of sufficient existing methods during students' performance prediction. Additionally no detailed investigations were done with the factors affecting achievement of the students [5]. With the aid of prediction models, weak students are identified and their scores are monitored since an examination result plays a vital role in student's especially engineering student's life [6]. Clustering is an interesting data mining technique as it is responsible for accurate and efficient classification of the data [7]. Traditionally to predict the students' performance questionnaire are utilized to spot 'at –risk' students. But currently to predict students' performance mostly a neural network classification method is utilized [8]. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide proper advising [9].

LITERATURE REVIEW

Kiuand ChingChieh[1] has proposed a J48 decision tree supervised classification algorithm that can predict students' performance by providing a prediction model. The work does not include any unsupervised data mining techniques and different EDM techniques. Kaur*et al*. [2] designed a Multi-Layer Perceptron classification algorithm which helps institutions to identify slow learning students and thereby provide a decision to give special aid to them. Further this research is not suitable for other applications such as medicine, sport, share market etc.Márquez-Veraet[3] used genetic programming methods to predict students' performance . But the model performed with only small database and was performed only for school no solutions for primary, secondary and higher education. Data mining knowledge management algorithm used by Natek*et al*. [4] The dataset size is not too large and also not done with international student dataset.The Decision tree data mining technique proposed by Lin et al in [5] it improves the limitations of creativity learning systems and enhance outcomes. But then no additional personal traits are included.The Multi-Label classification algorithm in[6] overcomes manual qualitative analysis and computational analysis of large scale user generated textual content. Jishan *et al*. [7] proposed a Optimal Equal width Binning with Synthetic Minority Over-Sampling method that improves the accuracy when SMOTE oversampling and optimal equal width binning are utilized together also the level of misclassification error is minimized.Goga *et al*.[8] proposed

Machine learning algorithms analyze all background attributes but Experiments are not given for different levels of education such as primary, secondary etc. Guarín *et al*. [10] used Naïve Bayes and decision tree classification to predict loss of academic status due to low academic performance. When the size of data increases ,classifier improvement deteriorates. The decision tree data mining algorithm proposed in [11] provide information to improve prediction but lack deepen generalize ability of results

PROPOSED METHODOLOGY

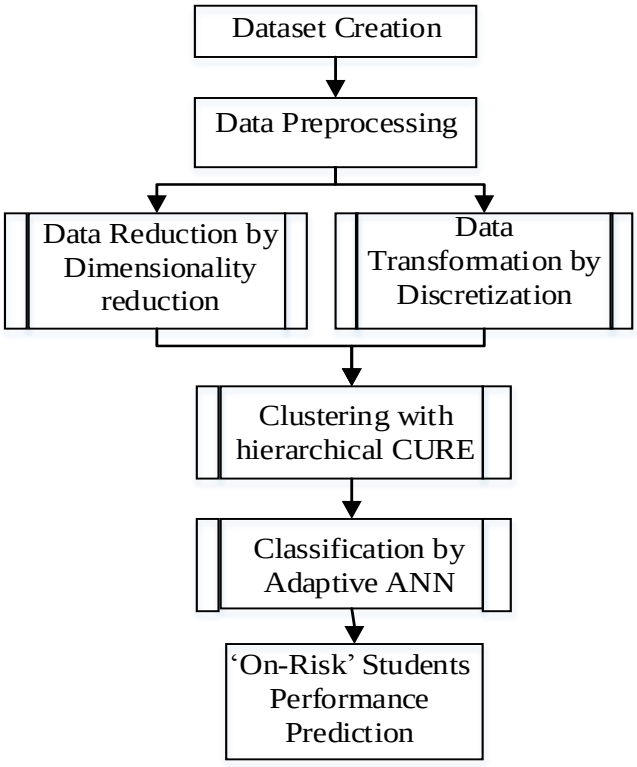


Figure 1: Proposed CURE – AANN Process Flow

Data set created:

Initially, a dataset is created by involving 27 different attributes which are not considered in other EDM such as Hours spent in study after college, Student family status ,Friends circle, Mother education , Class test grade, Seminar performance, Assignments , Attendance, Lab work, Grade in high school, Grade in senior secondary, Living location, Fathers qualification, Father occupation, Mother occupation, Result in first semester, Family size, Previous semester marks, Branch, Tenth board, Gap in study, Backlogs, Backlogs number , gender , social media, average family income, sponsor.

Data preprocessing:

Initially in EDM a data preprocessing technique will be performed in the available dataset by different stages such as data cleaning, data reduction and data transformation. Since this proposed research work aims on identification of reason for students (weak) performance by different attribute. Here ,only the weak students list will be processed for further processing. Thus the best students' performance are neglected or reduced by preprocessing and thus making data transformation in the dataset.

CURE Clustering:

This dataset then performs hierarchical clustering with aid of CURE algorithm that performs random sampling and partitioning. Apart from other clustering algorithms which undergo pre clustering CURE initially performs random sampling by which the input dataset is clustered correctly and also preserves cluster information. Then CURE partitions the random samples and partially clusters to speed up clustering. Thus the clustering process will be speed up when compared to traditional clustering algorithms and also reduces storage complexity. Since the clustering process utilizes only linear space the memory or storage complexity identified with existing clustering algorithms will be reduced.

The algorithm cannot be directly applied to large databases because of the high runtime complexity. Enhancements address this requirement.

- **Random sampling**: random sampling supports large data sets. Generally the random sample fits in main memory. The random sampling involves a trade-off between accuracy and efficiency.

  Churn off bound function

  For a cluster u,if the sample size s satisfies,

  Where, f belongs to 0<f<1 and δ belons to 0< δ <1

Pre-processing approaches can have significant drawbacks. Random sampling can throw out possibly useful data, while random sampling increases the size of the dataset and hence the training time. Random-Balance sustains the size of the training set and because it is a process which is repeated several times, the problem of removing important samples is reduced.

ADAPTIVE CURE CLUSTERING ALGORITHM

I. Pseudo code for the Random Balance ensemble method.
RANDOM BALANCE

Require: Set A of Samples

Ensure: New set N of examples with Random Balance

1: $totalsize \leftarrow |A|$

2: $A_{N\leftarrow} \{(x_i, y_i) \in S \mid y_i = -1\}$

3: $A_p \leftarrow \{(x_i, y_i) \in S \mid y_i = +1\}$

4: $Majoritysize \leftarrow |A_N|$

5: $Minoritysize \leftarrow |A_p|$

International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7493-7498

6: new MajoritySize $\leftarrow$Random integer between 2 and totalSize-2
//Resulting classes will have at least 2 instances
7: newMinoritySize $\leftarrow$totalSize-newMajoritysize
8: if newMajoritysize<Majority size then
9: $A \leftarrow S_p$
10: Take a random sample of size newMajoritySize from A_N, add the sample to N
11: Create newMinoritySize-minoritySize from A_P, add these samples to N
12: else
13: N $\leftarrow$ A_N
14: Take a random sample of size newMinoritySize from A_P, add the sample to N.
15: Create newMajoritySize - majority Size artificial samples from A_N, add these samples to N.
16: end if
17: return N

II. Partitioning: The basic idea is to partition the sample space into *p* partitions. Each partition contains *n/p* elements. The first pass partially clusters each partition until the final number of clusters reduces to *n/pq* for some constant q ≥ 1. A second clustering pass on *n/q* partially clusters partitions. For the second pass only the representative points are stored since the merge procedure only requires representative points of previous clusters before computing the representative points for the merged cluster. Partitioning the input reduces the execution times.

III. Clustering:

- For every cluster m (each input point), in m.mean and m.rep store the mean of the points in the cluster and a set of r representative points of the cluster (initially r = 1 since each cluster has one data point). Also m.closest stores the cluster closest to m.
- All the input points are inserted into a k-d tree K
- Treat each input point as separate cluster, compute m.closest for each m and then insert each cluster into the heap H. (clusters are arranged in increasing order of distances between m and m.closest).
- While size(H) > *c*
- Remove the top element of H(say u) and merge it with its closest cluster m.closest(say n) and compute the new representative points for the merged cluster M.
- Remove m and n from K and H.
- For all the clusters i in H, update i.closest and relocate i
- insert M into H
- repeat

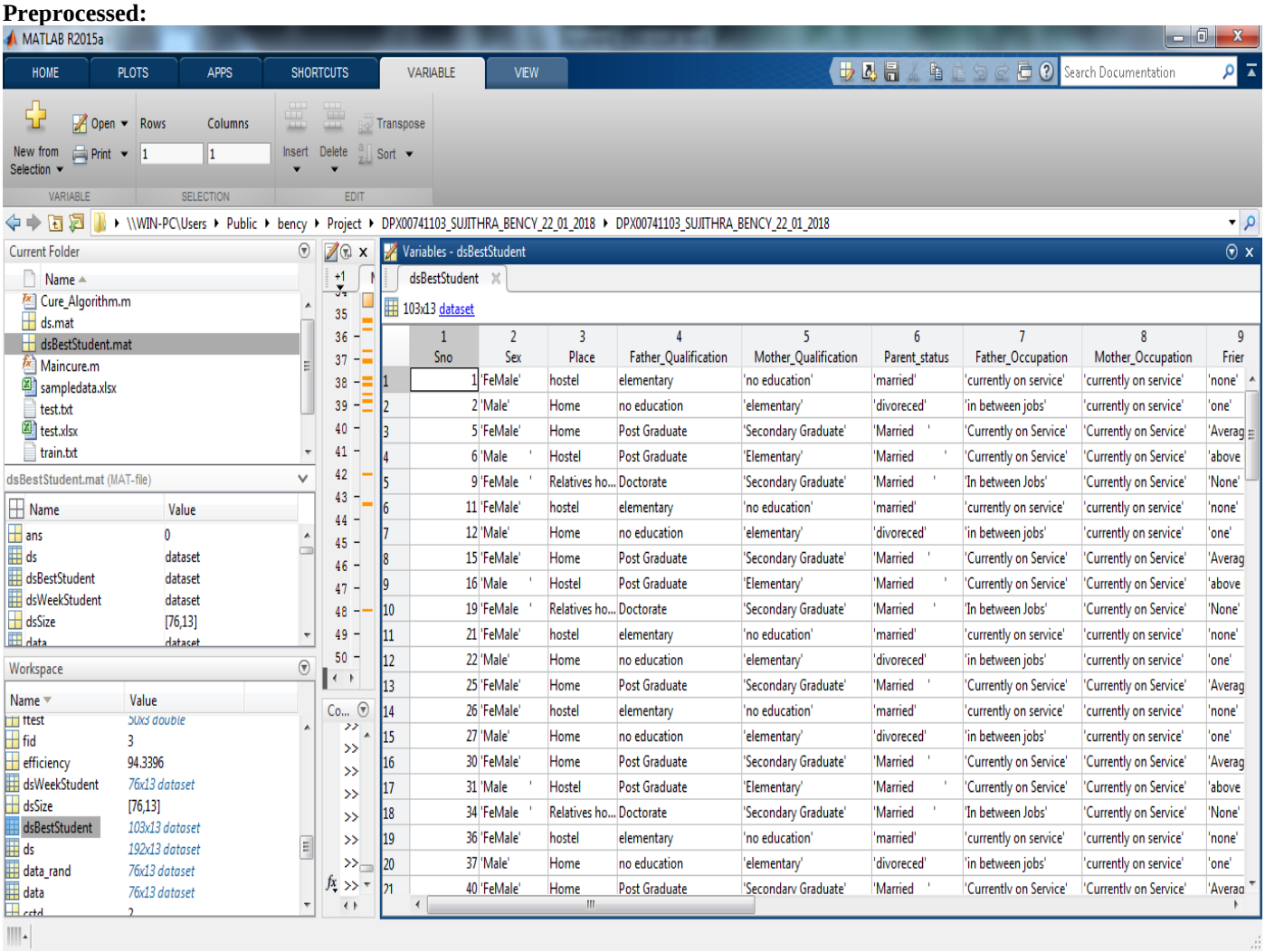
EXPERIMENTAL RESULT

Dataset:

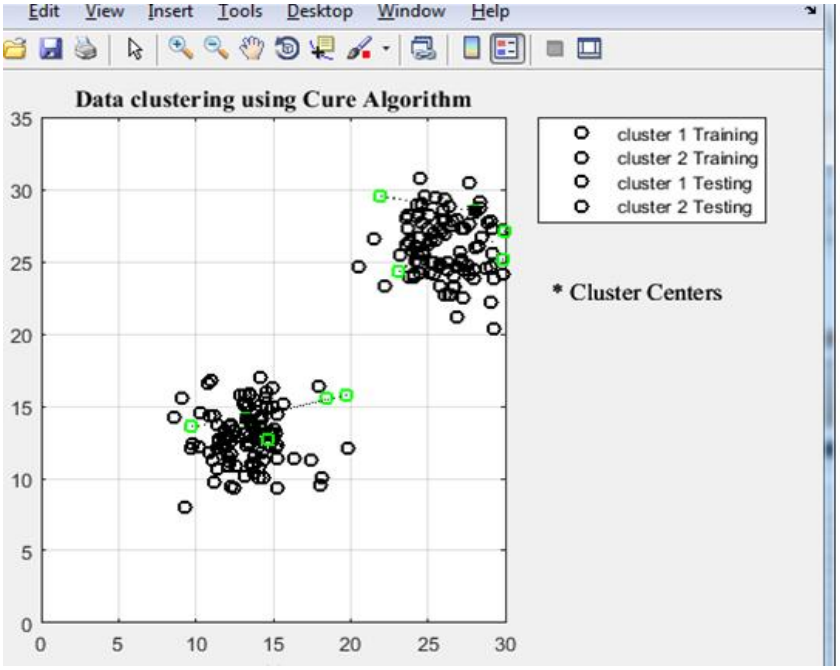
	A–F	G	H	I	J	K	L
1	Mention your Roll No :		1	2	3	4	
2	What about your Gender:		female	male	male	female	female
3	What is your Living location?		hostel	Day scholar from own home	Day scholar from relatives home	hostel	Day scholar fro
4	What was your Grade in High school?		excellent	Very Good	Good	Pass	excellent
5	What was your Grade in senior secondary?		excellent	Very Good	Good	Pass	excellent
6	In which board you were studied in Tenth Board?		state board	State board	central board	State board	entral board
7	Mention about your Student family size:		with both parents	with both parents	big family	Medium family	with both paren
8	What is your parent status?		married	divoreced	Separated	widowed	Married
9	What is your Father Qualification?		elementary	no education	Secondary Graduate	Secondary Graduate	Post Graduate
10	What is your Mother Qualification?		no education	elementary	elementary	Secondary Graduate	Secondary Grad
11	What is your Father Occupation Status?		currently on service	in between jobs	Retired	N/A	Currently on Ser
12	What is your Mother Occupation Status?		currently on service	currently on service	Currently on Service	Currently on Service	Currently on Ser
13	How many number of Friends you have?		none	one	average	medium	Average
14	How much Hours you spent with your friends per week?		none	very limited	medium	medium	Very limited
15	Do you have an interest to studying in the branch belonging to you?		strongly agree	somewhat agree	Strongly disagree	somewhat disagree	strongly agree
16	What about your Class test performance?		excellent	good	Pass	Pass	Excellent
17	Do you have an interest to take Seminars?		strongly agree	somewhat agree	strongly disagree	somewhat agree	strongly agree
18	Could you have the intention to write the Assignments regularly?		somewhat agree	strongly agree	somewhat disagree	somewhat agree	strongly agree
19	Do you have an attention to maintain the regular attendance?		strongly agree	strongly agree	Somewhat agree	somewhat agree	strongly agree
20	Are you having an interest to do your Lab works perfectly?		somewhat agree	strongly disagree	Somewhat agree	somewhat agree	strongly agree
21	What was your Result in first semester?		excellent	good	Pass	Pass	Excellent
22	How do you could perform in the previous semester?		excellent	Very Good	Good	Good	Excellent
23	Are you having an knowledge to spent Hours in study after college timings		somewhat agree	strongly agree	Strongly disagree	somewhat agree	strongly agree
24	Whether are you having any Gap in study?		no	no	yes	no	no
25	Whether are u having Backlogs, mention the backlogs no		no backlogs	one	Three	Two	No backlogs
26	Do you have an interest to access the Social media?		strongly disagree	somewhat disagree	strongly agree	somewhat agree	Somewhat agre

Sheet1 Sheet2

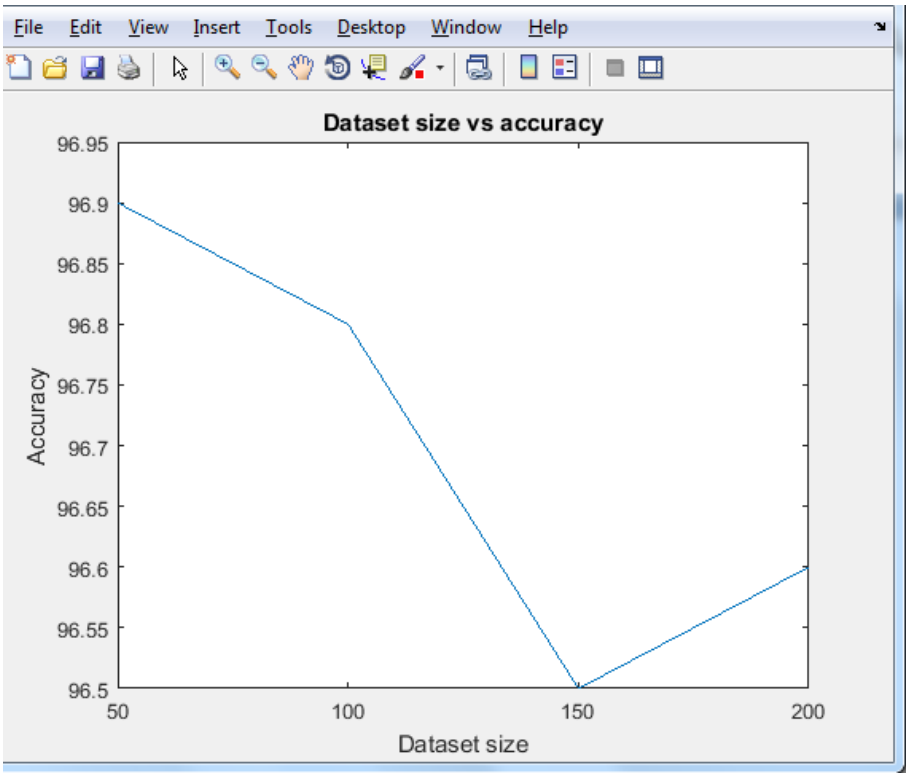
Preprocessed:



Clustering using CURE:



Clustering:



The Overall Accuracy is 96.6

Elapsed time is 11.929215 seconds.

CONCLUSION

In this paper CURE Clustering algorithm is used though the importance of analysing and predicting the reasons for the poor performance of the students has been addressed in other disciplines like psychology and sociology , its importance has been very less addressed in the area of educational data mining . The main objective of this research paper is to find out the reasons that contribute to the poor performance of students in educational institutions. Experiments and results have proved our claims regarding prediction and accuracy.

REFERENCES

[1] Kiu, Ching Chieh. "Supervised Educational Data Mining to Discover Students' Learning Process to Improve Students' Performance." In Redesigning Learning for Greater Social Impact, pp. 249-258. Springer, Singapore, 2017.

[2] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. "Classification and prediction based data mining algorithms to predict slow learners in education sector." Procedia Computer Science Vol.57, pp: 500-508, 2015.

[3] Márquez-Vera, Carlos, Alberto Cano, Cristóbal Romero, and Sebastián Ventura. "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data." Applied intelligence Vol.38, no. 3, pp: 315-330, 2013.

[4] Natek, Srečko, and Moti Zwilling. "Student data mining solution–knowledge management system related to higher education institutions." Expert systems with applications Vol.41, no. 14, pp: 6400-6407, 2014.

[5] Lin, Chun Fu, Yu-chu Yeh, Yu Hsin Hung, and Ray I. Chang. "Data mining for providing a personalized learning path in creativity: An application of decision trees." Computers & Education Vol.68, pp: 199-210, 2013.

[6] Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan. "Mining social media data for understanding students' learning experiences." IEEE Transactions on Learning Technologies Vol.7, no. 3, pp: 246-259, 2014.

[7] Jishan, Syed Tanveer, Raisul Islam Rashu, Naheena Haque, and Rashedur M. Rahman. "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique." Decision Analytics Vol.2, no. 1, pp: 1, 2015.

[8] Goga, Maria, Shade Kuyoro, and Nicolae Goga. "A recommender for improving the student academic performance." Procedia-Social and Behavioral Sciences Vol.180, pp: 1481-1488, 2015.

[9] Mrs. Manjula V and Dr.A.N.Nandakumar "Predicting Student's Learning Behavior prior to

University Admission" International Journal of Computer Applications (0975 – 8887) Volume 164 – No 5, April 2017 38.

[10] Guarín, Camilo Ernesto López, Elizabeth León Guzmán, and Fabio A. González. "A model to predict low academic performance at a specific enrollment using data mining." IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje Vol.10, no. 3, pp: 119-125, 2015.

[11] Asif, Raheela, Agathe Merceron, Syed Abbas Ali, and Najmi Ghani Haider. "Analyzing undergraduate students' performance using educational data mining." Computers & Education (2017).