Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 8, July 2021: 2994-3010

Regularized Canonical Variate Emphasis Jenks Breaks Boost Clustering For Student Performance Prediction With Big Data

Mrs. R. Pushpavalli[a], Dr. C.Immaculate Mary[b]

[a] Assistant Professor of Department of Computer Science, Sri Sarada College for Women(Autonomous), Salem – 636 016.
[b] Head of the Department & Associate Professor Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem-636 016.

ABSTRACT

Educational Data Mining is an emerging area due to the extensive growth of educational data. Recently, the amount of data stored in the educational database is increased rapidly. The stored database comprises hidden information about student performance and behavior. Predicting student's performance is more challenging due to a huge amount of data in educational databases. Many works carried out their research on data mining techniques to evaluate student performance. But the accurate prediction with minimum consumption is a challenging issue. In order to improve the prediction accuracy, a Regularized Canonical Variate Emphasis Jenks Breaks Boost Data Clustering (RCVEJBBDC) technique is introduced for student grade prediction with lesser time consumption. The RCVEJBBDC technique performs two processes namely Attribute Selection, and clustering for behavior analysis. In the RCVEJBBDC technique, the Regularized Canonical Variate Attribute Selection Process is carried out to select the relevant attributes from the input database using the radial basis kernel function. After attribute selection, Emphasis Jenks Breaks Boost Data Clustering is used to cluster the data based on student behavior analysis. Emphasis Boost Data Clustering combines the weak learner result to form the strong cluster output. Jenks Breaks Cluster (JBC) is considered as the weak learners for clustering the student data. Jenks Breaks is a data clustering technique to group the data values into different clusters. The Emphasis Boost technique combines the weak learners and provides strong clustering results by minimizing the quadratic error. This in turn helps to improve the student grade prediction accuracy and to minimize the time consumption based on their behavior. Experimental evaluation is carried out for factors such as prediction accuracy, false-positive rate, prediction time, and space complexity with respect to a number of student data. The empirical results demonstrate that the RCVEJBBDC technique provides higher prediction accuracy and lesser prediction time as well as space complexity than the conventional clustering techniques.

Keywords: Student Grade Prediction, big data, Regularized Canonical Variate Attribute Selection, radial basis kernel function, Emphasis Jenks Breaks Boost Data Clustering

1. INTRODUCTION

The performance of students in the academic field has exposed the consideration by researchers for improving their weaknesses. In recent days, digitization is used by the institution in teaching-learning and other academic processes creating a large volume of digital data. This data is helpful for teachers, and administrators for decision-making effectively and to examine the performance of students. In addition, educational institutions include a huge amount of academic databases comprised of student details. With the increasing growth in a large data warehouse, the necessity for analyzing the student data and extracting valuable information is a challenging task. Many researchers carried out data mining techniques such as clustering, data classification methods to evaluate student performance.

Clustering-Based Ensemble Meta-based Tree (CB-EMT) model was introduced in [1] for predicting the student performance based on the relevant features. But the model failed to provide accurate results while increasing the number of records in the dataset. A hybrid cluster-based LDA (CLDA) and ANN were developed in [2] for student performance prediction. But the time consumption of the performance prediction was not minimized.

An artificial neural network was introduced in [3] for performance predictions depend on student scores. But the model failed to predict the performance of students based on instant browsing data i.e. number of clicks. Besides, the accuracy score of the prediction was not improved. The student performance prediction was carried out in [4] with minimal available attributes. But it failed to compare student marks prediction accuracy with the results of this experiment.

A novel prediction algorithm classification and clustering techniques were developed in [5] to estimate the student's performance. But the technique failed to support large varieties of features of the student dataset. An ensemble meta-based tree model (EMT) was introduced [6] for predicting the student performance with the selected features. But the designed tree model failed to conduct the experiments using other clustering techniques.

A Learning Management system was developed in [7] for student behavior prediction using Big Data. However, the system was not efficient for accurate student behavior patterns and their grades prediction. A Firefly Grey Wolf-Assisted K-Nearest Neighbor Classifiers were developed in [8] for identifying the performance of college students using important features. But the time consumption was not minimized.

A deep artificial neural network was introduced in [9] to find the student's academic performance using big data. Though the model increases the accuracy, the analysis of prediction time was not performed. Automated Machine Learning algorithms were developed in [10] to improve the accuracy of predicting student performance. But it failed to use the ensemble models in predicting the student career success using academic data.

1.1. Contribution of the work

The major contribution of the proposed RCVEJBBDC technique is summarized as given below,

- To improve the student performance prediction, a novel RCVEJBBDC technique is introduced with two main processes namely attribute selection and clustering.

- Firstly, this RCVEJBBDC technique intends to develop the Regularized Canonical Variate Attribute Selection. The attribute selection process is carried out using the radial basis kernel function to find similar and dissimilar features. This process helps to minimize the performance prediction time and space complexity.

- Secondly, Emphasis Jenks Breaks Boost Data Clustering is performed to predict the student grade based on selected attributes**.** The Jenks Breaks technique groups the data into different clusters based on the mean and deviation. The weak learner results are combined. Then the Emphasis function is applied to measure the quadratic error of each weak learner. The ensemble technique finds the weak learner with minimum error. This helps to improve the prediction accuracy and minimize the false positive rate.

- Finally, extensive experimental assessments are carried out with various performance metrics to highlight the improvement of the proposed RCVEJBBDC technique over conventional clustering techniques.

1.2. Outline of paper

The rest of the paper is organized into five various sections. Section 2 introduces the related works. Next, a detailed description of our proposed RCVEJBBDC technique is presented with a neat diagram in section 3. Section 4 provides the experimental settings. In section 5, the performance of different proposed and existing methods is discussed with different metrics. Finally, section 6 provides a conclusion.

2. RELATED WORKS

A hybrid classifier approach called fuzzy and Bayesian was developed in [11] for the prediction of student performance with higher accuracy. But the time consumption of performance prediction was not minimized. A hybrid regression model was introduced in [12] that optimize the prediction accuracy of student academic performance. But the complexity analysis was not performed.

An Adaptive Sparse Self-Attention Network (AS-SAN) was introduced in [13] for fine-grained student learning performance prediction. But the higher accuracy was not obtained in the performance prediction. Many data mining techniques were introduced in [14] to predict student performance with higher accuracy. But the other data mining techniques such as clustering were not applied for student performance prediction.

A cluster-based distributed architecture was designed in [15] for predicting the student's performance. But the distributed architecture failed to improve the performance of the clustering process. The ensemble approach was developed in [16] based on a heuristic system to improve the prediction models with higher accuracy. But the approach failed to perform the attribute selection for minimizing the dimensionality of the dataset.

Different types of Artificial Neural Networks (ANNs) were developed in [17] for improving the academic performance. But the system failed to consider a large number of student's data in the prediction process. Student performance prediction was performed in [18] using a clustering algorithm. But the accurate prediction was not performed with a minimum error rate.

A multi-layered neural network was introduced in [19] o predict the students' final degree. But the feature selection process was not carried out. The student course achievement was predicted in [20] using Moodle and homework submission data. The clustering and classification techniques were introduced to sort the students into the various class of course achievement. But, the approach was not considered to more information on students' activities for accurate prediction performance.

3. METHODOLOGY

Predicting students' performance is the most significant topic for learning circumstances such as schools, colleges since it helps to validate the academic results. The amount of data in the education field is getting increased gradually with the help of admission system, academic information system, and learning management system, e-learning, and so on. The collected data from the students are typically used for behavior prediction. But most of the data unused because of complexity and large volume data set. Therefore, to analyze such kind of a large amount of educational data is the great significance to predict student performance. Therefore, a novel RCVEJBBDC technique is introduced for student grade prediction based on their behavior with higher accuracy.

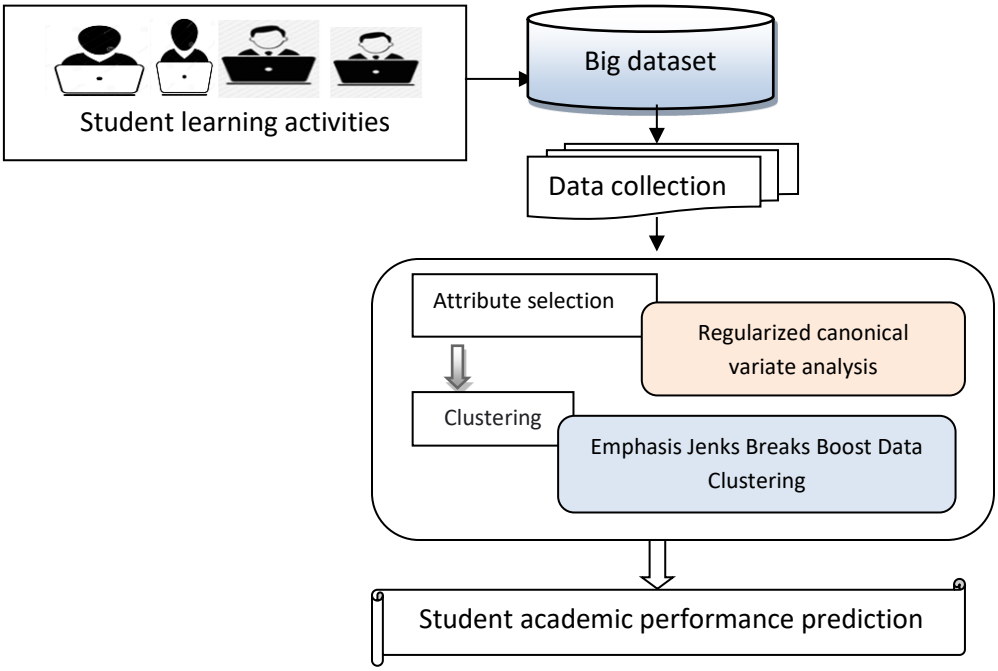

Figure 1 Architecture of the proposed RCVEJBBDC technique

Figure 1 given above illustrates the architecture diagram of the proposed RCVEJBBDC technique to provide higher accuracy of performance student academic grade prediction with higher accuracy with the big dataset. A student performance data set used in this methodology has been collected from UCI Machine Learning Repository. With the collected data from the dataset, the attribute selection process is carried out in the first process before the academic prediction. Regularized canonical variate analysis is applied to find the more relevant attributes from the dataset for minimizing the complexity of performance prediction. After that, the student academic level prediction is said to be achieved using Emphasis Jenks Breaks Boost Data Clustering with higher

accuracy. These two processes of the proposed RCVEJBBDC technique are briefly described in the following subsections.

3.2. Regularized kernel canonical variate analysis based attribute selection

The proposed RCVEJBBDC technique initially performs the attribute selection to minimize the complexity of the student academic performance prediction. Since the dataset consists of many attributes, which inappropriate for clustering purposes and it also causes high dimensionality when considering the large amounts of student's characteristics. These kinds of problems are resolved by selecting significant attributes from the big dataset. The purpose of attribute selection is to find an appropriate subset of attributes and removes the irrelevant data that efficiently minimizes the dimensionality of attribute space in the dataset.

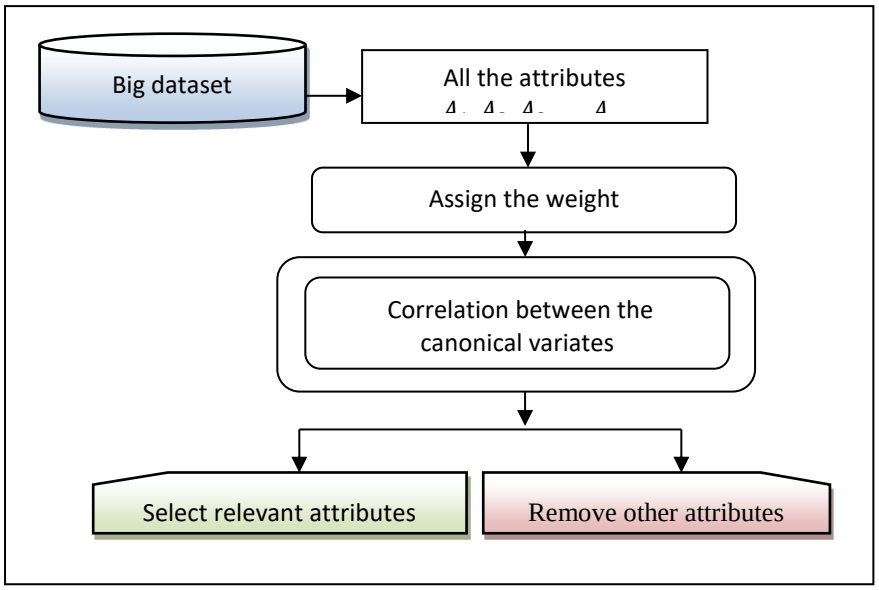

Figure 2 flow process of attribute selection using Regularized kernel canonical variate analysis

Figure 2 given above demonstrates that the Regularized kernel canonical variate analysis Set based feature selection. As shown in figure 2, set of attributes in a given dataset $D \in \{A_1, A_2, A_3, \ldots . A_m\}$. For each attribute, the Regularized canonical variate analysis is performed to find the relevant attribute subset and irrelevant attribute subset. Regularized canonical variate analysis is a method for finding a linear correlation between two variables.

Let us consider the attribute set $\{A_1, A_2, A_3, \ldots . A_m\}$ and assign the weight to the attributes are denoted as follows,

$$c_{vi} = \langle a_i, A_i \rangle \quad (1)$$

$$c_{vj} = \langle b_j, A_j \rangle \quad (2)$$

Where, c_{vi}, c_{vj} denotes canonical variates, a_i denotes a canonical weight of the attribute 'A_i', b_j denotes a canonical weight of the attribute 'A_j'. The correlation between the canonical variates is measured as follows,

$$\beta = arg\ max\ [k\langle c_{vi}, c_{vj}\rangle] \qquad (3)$$

$$k\langle c_{vi}, c_{vj}\rangle = \exp\left[-\frac{\|c_{vi}-c_{vj}\|^2}{2*d^2}\right] \qquad (4)$$

Where β denotes a canonical correlation coefficient, $arg\ max$ denotes an argument maximum function, $k\langle c_{vi}, c_{vj}\rangle$ denotes a radial basis kernel function, d denotes a deviation. The maximum correlated features are identified and select the one feature and remove the other one. In this way, the relevant features are selected from the dataset, and other features are removed. The algorithmic process of the Regularized kernel canonical variate analysis based feature selection is described as given below,

Algorithm 1: Regularized kernel canonical variate analysis based feature selection
Input: Big dataset, attributes $A_1, A_2, A_3, \ldots . A_m$ **Output:** Selected attributes subset
Begin **Step 1: For each attribute** in dataset 'A_i' **Step 2:** Assign the weight c_{vi} , c_{vj} **Step 3**: **for each** canonical variates **Step 4**: Measure the canonical variates correlation **Step 5:** **If** (arg max $k\langle c_{vi}, c_{vj}\rangle$) **then** **Step 6:** Attributes are similar **Step 7:** **else** **Step 8:** Attributes are dissimilar **Step 9:** **end if** **Step 10:** Select attributes and remove the other attributes **Step 11: End for** **Step 12: end for**

Algorithm 1 given above describes Regularized kernel canonical variate analysis-based feature selection. Initially, the attributes are selected from the dataset. Then the weight is assigned to each attribute and obtains the canonical variates. Then the correlation between the canonical variates is measured using radial basis kernel function to find similar and dissimilar attributes. The maximum correlated features are identified and select the one feature and remove the other features. As a result, the selected feature subsets are used for classification and it helps to minimize the time and space complexity of the student performance prediction.

3.2 Weighted Emphasis Jenks Breaks Boost Data Clustering

Big data analytics is the process of collecting and examining a huge amount of data to determine useful information. In general, big data are used for predictive analytics with the selected features. The proposed technique uses the Emphasis Jenks Breaks Boost Clustering technique for student performance prediction. The Emphasis Jenks Breaks Boost is an ensemble meta-algorithm that provides improved clustering performance than any of the weak clusters alone. A weak cluster is a machine learning algorithm that provides the clustering outcomes with the probability of some error.

On the contrary, a boosting technique is a strong learner that accurately provides a better clustering performance with lesser error probability. The basic construction of the ensemble learning is shown in figure 3.

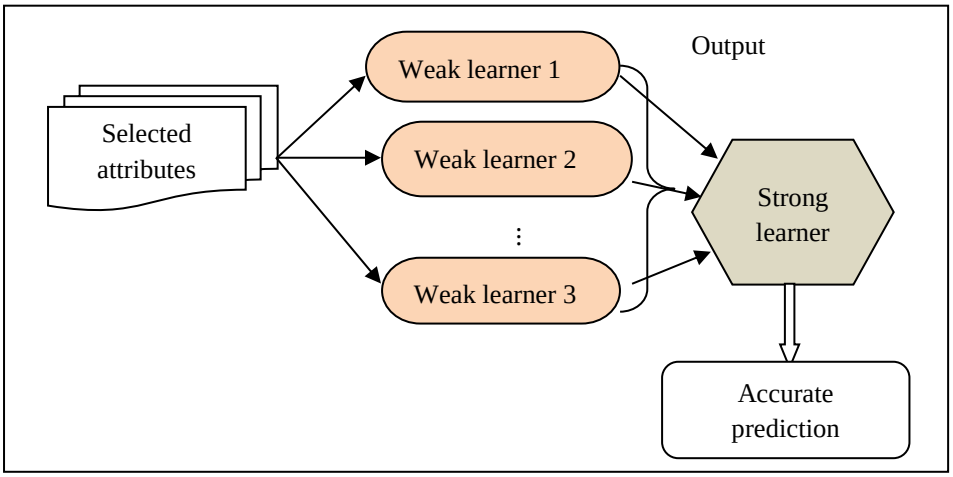

Figure.3 block diagram of Emphasis Jenks Breaks Boost clustering

Figure 3 demonstrates the block diagram of the boosting ensemble clustering technique to obtain accurate prediction with minimum time. The boosting ensemble technique considers the training set $\{X_i, Y_i\}$ where $X_i = A_1, A_2, \ldots, A_b$' denotes the selected features with sample data and Y_i indicates the clustering outcomes of the ensemble technique. As shown in figure 3, the boosting ensemble technique initially constructs 'b' weak learners $v_1, v_{2,} v_{3,} \ldots . v_b$ and their results are combined to obtain strong clustering results. The proposed boosting ensemble technique uses the weak learner as a Jenks breaks clustering to group the given input samples i.e. data $D_1, D_2, D_3, \ldots D_n$ into different clusters based on a mean and deviation. Initially, the numbers of clusters are initialized randomly. After that, the mean (i.e. centroid) is assigned for each cluster. The mean of the cluster is measured as given below,

$$\mu = \frac{\sum_{i=1}^{n} D_i}{n} \quad (5)$$

Where μ denotes a mean of the particular cluster is measured as the ratio of the sum of all the samples 'D_i' in the particular cluster to the total number of data 'n'. Then the Jenks breaks the clustering technique segment or partitions the samples into different clusters. The clustering technique measures the sum of squared deviations from the means as given below,

$$d = \sum_{i=1}^{n} [D_i - \mu]^2 \quad (6)$$

Where d denotes a deviation, μ denotes a mean, D_i denotes input samples or data. The clustering technique minimizes average deviation from the mean within the cluster, while deviation from the means of the other clusters. In other words, the clustering technique reduces the variance within-cluster and maximizes the variance between the clusters. The data closer to the mean or minimum deviation is grouped into the particular cluster.

$$H = \arg\min\ d \quad (7)$$

Where H denotes an output of weak clusters. In this way, all the data are grouped into the number of clusters based on mean and deviation. The weak learner has some training errors in the clustering outcomes. In order to obtain strong clustering results, the ensemble technique combines all the weak learner results as given below,

$$Y = \sum_{i=1}^{b} H_i \quad (8)$$

From (9), Y indicates ensemble results,H_i indicates an output of the weak learners. Then the weight is initialized to find the accurate results,

$$Y = \sum_{i=1}^{b} \omega_i H_i \quad (9)$$

From (9), 'ω_i' indicates the weight of the weak learner. The weight is a random integer number. The Emphasis Boost ensemble technique uses the weighted emphasis function for each input pattern according to a parameter (δ). The weighted emphasis function measures the quadratic error of each pattern as given below,

$$\varphi = \exp\left[\delta\left(\left(\sum_{i=1}^{b} \omega_i H_i - Y\right)^2 - (1-\delta)\left(\sum_{i=1}^{b} H_i\right)^2\right)\right] \quad (10)$$

Where φ denotes a weighted emphasis function, δ denotes a weighting parameter, Y denotes actual ensemble results, $\sum_{i=1}^{b} \omega_i H_i$ denotes predicted results of weak learner with the weight ω_i and the without weight$\sum_{i=1}^{b} H_i$. From the above equation, δ denotes a weighting parameter value substitutes 1 and get the final output,

$$\varphi = \exp\left[\left(\sum_{i=1}^{b} \omega_i H_i - Y\right)^2\right] \quad (11)$$

The emphasis function φ only gives attention to the quadratic error of each cluster. Based on the error value, the weak learner with minimum error is chosen as strong clustering results than the other clusters. As a result, accurate clustering results are obtained. Based on the clustering results, the student performance level is correctly predicted with higher accuracy. The algorithmic process of the Weighted Emphasis Jenks Breaks Boost Data Clustering is described as give below**,**

// **Algorithm 2 Weighted Emphasis Jenks Breaks Boost Data Clustering**
Input: Extracted features $X_i = A_1, A_2, \ldots, A_b$ and data $D_1, D_2, D_3, \ldots D_n$
Output: Improve the prediction accuracy
Begin **Step 1**: Construct 'b' weak learners **Step 2:** **Initialize** 'c' number of clusters **Step 3:** **for each** cluster **'c'** **Step 4:** Assign the mean value **'μ'** **Step 5:** **end for** **Step 6:** **for each** μ **Step 7:** **for each data** D_i **Step 8:** Measure the squared deviation 'd' **Step 9**: **end for** **Step 10:** **end for**

Step 11: Finds the minimum deviation
Step 12: Group the data into particular cluster
Step 13: Combine all weak learner results $Y = \sum_{i=1}^{b} H_i$
Step 14: **for** each weak learner results
Step 15: Initialize the weight ' ω_i'
Step 16: Measure the quadratic error ' β' based on the emphasis function
Step 17: Find the weak learner with minimum error
Step 18: **Return** (strong clustering results)
Step 19: **end for**
End

Algorithm 2 given above describes the step-by-step process of clustering to predict performance with higher accuracy and minimum error. The boosting ensemble technique initially constructs a set of weak learners as the Jenks Breaks clustering technique with the training samples. The weak learner finds the mean and deviation to group the data into a particular cluster. The ensemble technique combines the weak learner to obtain accurate strong clustering results. The weighted emphasis function measures the quadratic error of each clustering result. The strong clustering results find the weak learner with a lesser quadratic error. Finally, the strong learner finds accurate clustering results with a minimum error rate.

4. EXPERIMENTAL SETTINGS

Experimental evaluation of proposed RCVEJBBDC technique and existing methods CB-EMT [1], Hybrid CLDA, and ANN approach [2] is carried out using Java Language and Educational Process Mining (EPM): A Learning Analytics Data Set. The dataset is taken from the UCI machine learning repository [https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set]. The dataset comprises 13 attributes and 230318 instances. Educational Process Mining data set is constructed based on the recordings of 115 student's activities (i.e. behaviors) through a logging application while learning with an educational simulator. The data set consists of the different students' time series of activities during six various sessions. There are 6 folders containing the student's data. The associated task performed by the dataset is classification, regression, and clustering with the aim of predicting the final performance grade. The attributes are listed in table 1.

Table 1 Attribute Description

S. No.	Features	Description
1	Session	Number of lab sessions from 1 to 6.
2	Student-ID	115 students' ID number (1,2,3….115)
3	Exercise	It shows the working Ex. The ID of the student. (Es_2_1 represents Session 2, Exercise 1).
4	Activity	The activities are grouped into 15 categories related to Exercise,

		Using Deeds Simulator, Using Text Editor, Working on Diagram, Working on Properties Window, Viewing Study Materials, Using Finite State Machine Simulator, Using Aulaweb, Blank, and other irrelevant Activities
5	Start-time	Starting date and time of a specific activity.
6	End-time	Ending date and time of a specific activity.
7	Idle-time	The duration of idle time between the start and end time.
8	Mouse-wheel	The volume of mouse wheel operations during an activity.
9	Mouse-wheel-click	Number of mouse wheel clicks during an activity.
10	Mouse-click-left	Number of left mouse clicks during an activity.
11	Mouse-click-right	Number of right mouse clicks during an activity.
12	Mouse-movement	Distance moved by the mouse movements during activity.
13	Keystroke	Number of key presses during an activity.

5. RESULTS AND DISCUSSION

The experimental results of the three different methods namely the RCVEJBBDC technique and existing methods CB-EMT [1], Hybrid CLDA, and ANN approach [2] are discussed with respect to various performance metrics such as prediction accuracy, false-positive rate, predation time, and space complexity. These metrics are evaluated with a number of data using a table and graphical representation.

5.1. Impact of prediction accuracy

Prediction accuracy is measured as the ratio of the number of student data (i.e. instances) are correctly predicted to the total number of data taken for the experimental evaluation. Therefore, the student performance grade prediction accuracy is calculated as follows,

$$PA = \left[\frac{ncp}{n}\right] * 100 \quad (12)$$

Where PA denotes a prediction accuracy, n denotes the number of the data, ncp denotes the number of data correctly predicted. Therefore, the overall prediction accuracy is measured in terms of percentage (%).

Table 1 Comparison of Prediction Accuracy

Number of data	Prediction accuracy (%)		
	RCVEJBBDC	CB-EMT	Hybrid CLDA and ANN approach
1000	89	85	83
2000	90	87	84

3000	88	83	80
4000	86	82	81
5000	89	85	82
6000	88	86	83
7000	90	87	84
8000	89	85	83
9000	90	87	84
10000	89	86	85

Table 2 reports experimental results of student performance prediction accuracy according to the number of student data ranged from 1000 to 10000 from the big dataset. Table 2 shows three clustering algorithms used in our comparison. The RCVEJBBDC technique in comparison with CB-EMT [1], Hybrid CLDA, and ANN approach [2] achieves an increase in performance of accuracy. Let us consider the 1000 data for conducting the experiment in the first iteration, 890 student data (i.e. instances) are correctly clustered and predict the performance level therefore the accuracy of the proposed RCVEJBBDC technique is 89%. Followed by, the 850 and 830 data are correctly clustered CB-EMT [1], Hybrid CLDA and ANN approach [2] and their prediction accuracy are 85% and 83% respectively. Similarly, other iterations are conducted with different counts of input data. For each clustering method, ten different results are observed. The observed results of the proposed RCVEJBBDC technique are compared to the accuracy of existing methods. The average comparison results confirm that the student performance prediction accuracy of the RCVEJBBDC technique is considerably increased by 14% when compared to CB-EMT [1] and 11% when compared to Hybrid CLDA and ANN approach [2] [2] respectively.

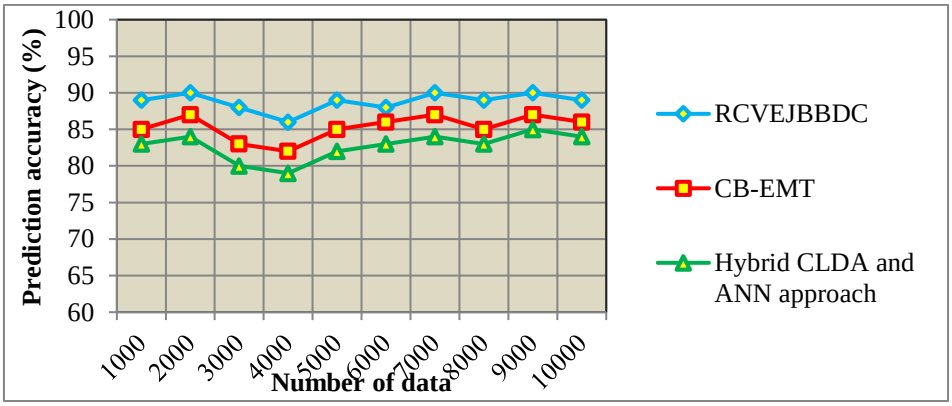

Figure 4 graphical representation of Prediction accuracy

Figure 4 depicts the experimental results of student performance prediction accuracy according to the number of data. The numerous input data are taken on the horizontal axis and prediction accuracy is observed on the vertical axis in terms of percentage. As shown in the graphical plot, there various colors of lines such as blue, red, and green indicate the prediction accuracy of three methods namely RCVEJBBDC, CB-EMT [1], Hybrid CLDA, and ANN approach [2] respectively. Among the three methods, the proposed RCVEJBBDC is capable of increasing the prediction accuracy. The reason behinds this improvement is to apply the Emphasis Jenks Breaks Boost Data Clustering technique. The proposed clustering technique is an ensemble clustering technique that uses the Jenks Breaks

Clustering as a weak learner to group similar data into different clusters based on the mean and deviation. Therefore, the ensemble technique combines the weak learner and obtains the strong clustering results with higher accuracy. Based on the clustering results, the student grade levels such as low, medium, and high are correctly predicted.

5.2 Impact of the false-positive rate

The false-positive rate is measured as the ratio of a number of student data (i.e. instances) are incorrectly predicted to the total number of data taken for the experimental evaluation. The false-positive rate is measured as given below,

$$FPR = \left[\frac{nicp}{n}\right] * 100 \quad (13)$$

Where FPR denotes a false positive rate, n denotes the number of the data, $nicp$ denotes the number of data incorrectly predicted. Therefore, the false positive rate is measured in terms of percentage (%).

Table 2 Comparison of the false-positive rate

Number of data	False-positive rate (%)		
	RCVEJBBDC	CB-EMT	Hybrid CLDA and ANN Approach
1000	11	15	17
2000	10	13	16
3000	12	17	20
4000	14	18	21
5000	11	15	18
6000	12	14	17
7000	10	13	16
8000	11	15	17
9000	10	13	15
10000	11	14	16

The false-positive rate of three different methods namely RCVEJBBDC, CB-EMT [1], Hybrid CLDA, and ANN approach [2] are reported in table 2. The experimental results illustrate that the false positive rate of the prediction using the RCVEJBBDC technique is minimal than the other existing methods. Let us consider 1000 data to measure the false positive rate in the first iteration. 110 data are incorrectly grouped and the false positive rate is 11% using the RCVEJBBDC technique. Similarly, 150 and 170 data are incorrectly grouped and the false-positive rates are 15% and 17% using CB-EMT [1], Hybrid CLDA, and ANN approach [2] respectively. Similarly, ten different results are observed for various counts of the input data. The false-positive rates of the proposed RCVEJBBDC technique are compared to the existing methods. The average of ten results indicates that the false positive rate is significantly reduced by 24% and 35% when compared to existing methods.

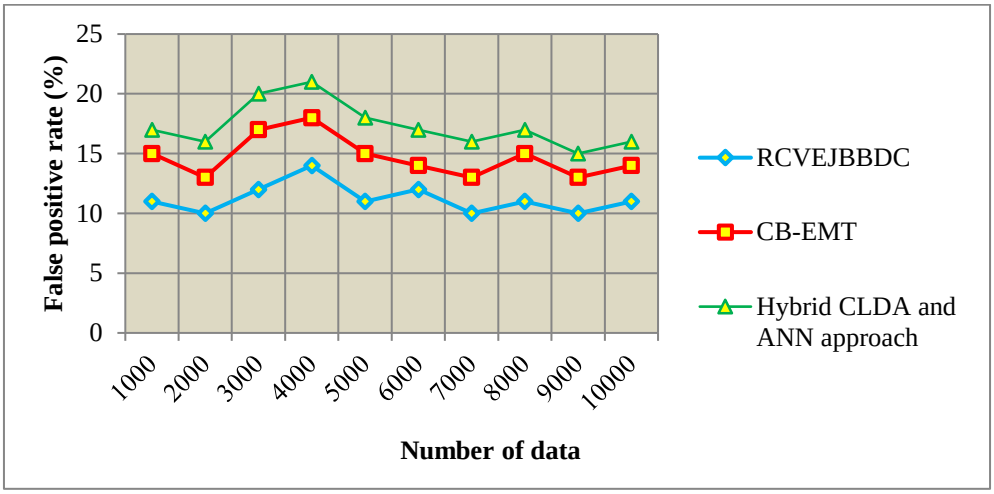

Figure 5 graphical representation of the false positive rate

Figure 5 depicts the performance results of the false positive rate with respect to the number of data. As illustrated in the above graphical plot, the proposed RCVEJBBDC technique attains higher prediction accuracy resulting it reduces the false positives. The reason for this improvement is to apply the ensemble clustering technique The boosting ensemble technique initially constructs a set of weak learners as the Jenks Breaks clustering technique with the training samples. The weak learner finds the mean and deviation to group the data into a particular cluster. The ensemble technique combines the weak learner to obtain accurate strong clustering results. The weighted emphasis function measures the quadratic error of each clustering result. The strong clustering results find the weak learner with a lesser quadratic error. Finally, the strong learner finds the accurate clustering results with a minimum false positive rate.

5.2.Impact of prediction time

Prediction time is measured as the amount of time taken by an algorithm to predict the student performance level in terms of grade based on the clustering process. The prediction time is mathematically calculated as follows,

$$PT = n * time\ (predict\ one\ student\ data\ ) \quad (14)$$

Where PT denotes a prediction time, n represents a number of student data, $time$ denotes a time for predicting the one student data. Prediction time is measured in terms of milliseconds (ms).

Table 3 Comparison of Prediction time

Number of data	Prediction time (ms)		
	RCVEJBBDC	**CB-EMT**	**Hybrid CLDA and ANN approach**
1000	20	25	28
2000	26	28	32
3000	30	32	36

4000	33	36	40
5000	38	40	43
6000	43	45	48
7000	44	47	49
8000	48	50	52
9000	52	54	56
10000	55	58	60

Table 5 describes the performance analysis of student performance prediction time versus the number of data taken from the big dataset. The time taken for predicting the student performance in terms of grade level is significantly reduced using the RCVEJBBDC technique than the other two existing techniques. Let us taken '1000 data for experimentation, the time consumption of RCVEJBBDC technique for predicting the student performance is $20ms$', whereas '$25ms$ and '$28ms$ time consumed by existing CB-EMT [1], Hybrid CLDA and ANN approach [2]. Therefore, the overall obtained results of the proposed RCVEJBBDC technique are compared to conventional clustering techniques. The average of ten results illustrates that the prediction time is significantly decreased by 7% and 14% when compared to state-of-the-art methods.

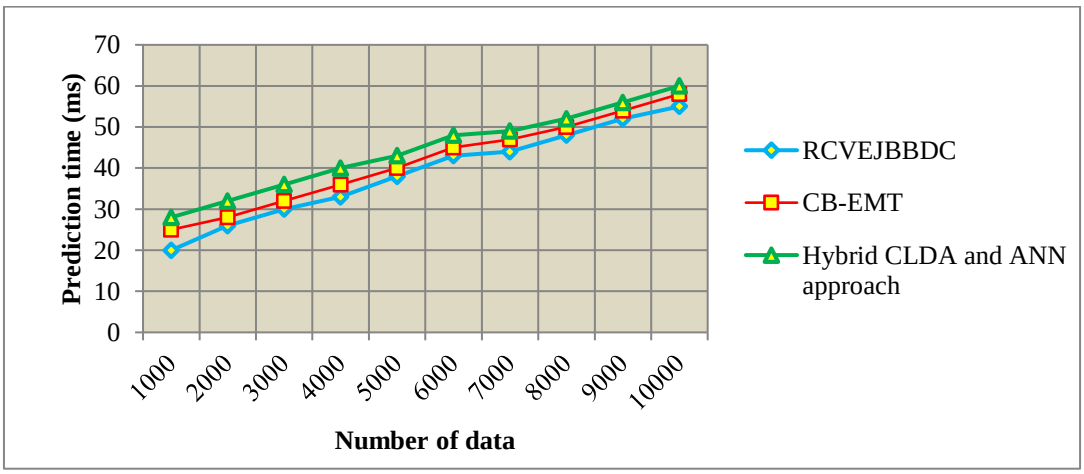

Figure 6 graphical representation of prediction time

The performance results of student performance prediction time along with the number of data are revealed in figure 6. As shown in the graphical representation, the prediction time of all the methods is gradually increased while increasing the number of data for each run. From the graphical results, the proposed RCVEJBBDC technique minimizes the prediction time. This is due to the application of Regularized kernel canonical variate analysis-based feature selection. For each attribute in the dataset, the weight is assigned and obtains the canonical variates. Then the correlation is measured using radial basis kernel function to find similar and dissimilar attributes. The maximum correlated attributes are identified and select the one feature and remove the other features. As a result, the selected attributes are used for clustering processing resulting in minimizing the time consumption of accurate prediction.

5.3 Impact of Space complexity

Space complexity is measured as the amount of memory consumed by the algorithm to predict the student performance level based on the clustering process. The space complexity is formulated as given below,

$$SC = n * Mem\ (predict\ one\ student\ data\ ) \quad (15)$$

Where SC denotes a space complexity, n represents a number of student data, Mem denotes a memory consumed for predicting the one student data. The overall space complexity is measured in terms of Megabytes (MB).

Table 4 comparison of Space complexity

Number of data	Space complexity (MB)		
	RCVEJBBDC	CB-EMT	Hybrid CLDA and ANN approach
1000	18	21	23
2000	20	22	24
3000	24	27	30
4000	28	30	32
5000	33	35	38
6000	35	37	39
7000	39	41	43
8000	41	43	48
9000	43	47	50
10000	45	48	52

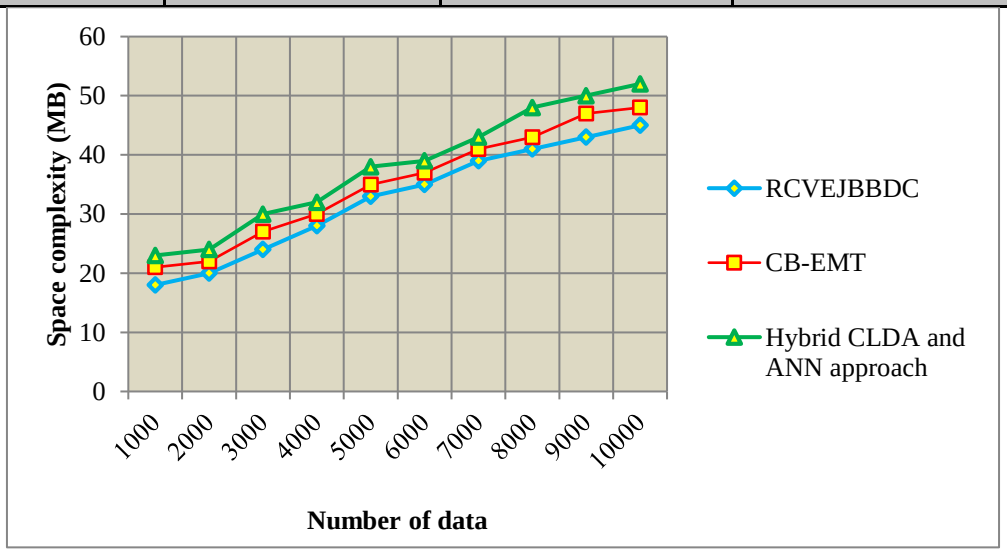

Figure 7 graphical representation of space complexity

Table 4 and figure 7 show the graphical illustration of the space complexity of predicting the student grade level using three different methods namely RCVEJBBDC technique and existing CB-EMT [1],

Hybrid CLDA and ANN approach [2]. As shown in figure 7, a linear increasing trend is to be observed for all the methods while increasing the number of data. Among the three methods, the RCVEJBBDC technique is significantly minimized than the other two methods. Besides, from the sample calculation provided above the table, with '1000' data are considered to perform the experimentation, the memory consumption for predicting the student grade level using the RCVEJBBDC technique was found to be '$18MB$' and memory consumption of [1], [2] was found to be '$21MB$', $23MB$'respectively. The overall comparison results indicate that the average value of space complexity of the RCVEJBBDC technique is considerably reduced by 8% and 15% when compared to conventional methods. This is because by applying Regularized kernel canonical variate analysis based feature selection and other attributes are removed. This helps to minimize the dimensionality of the dataset. Therefore, the proposed RCVEJBBDC technique uses the lesser memory space for predicting the student grade level.

6. CONCLUSION

A novel RCVEJBBDC technique is developed for accurate student performance prediction consisting of two phases namely attribute selection and clustering. The attribute selection process is carried out using Regularized Canonical Variate analysis based on the radial basis kernel function. With the selected relevant features, the clustering process is carried out using the Emphasis Jenks Breaks Boost technique. The ensemble clustering technique uses the Jenks Breaks clustering as a weak learner to group the data into the clusters based on the mean and deviation. The clustering weak learner results are combined to make a strong by minimizing the quadratic error. This helps to improve the accurate prediction and minimizes the false positive rate. The comprehensive experimental evaluation is conducted with different performance metrics such as prediction accuracy, false-positive rate, prediction time, and space complexity. The observed result indicates that the proposed RCVEJBBDC technique offers better performance than the baseline approaches, having higher prediction accuracy, and lesser false positive rate, time as well as space complexity.

REFERENCES

[1] Ammar Almasri, Rami S. Alkhawaldeh & Erbuğ Çelebi, "Clustering-Based EMT Model for Predicting Student Performance", Arabian Journal for Science and Engineering, Springer, 2020, Pages 1-12

[2] Sakshi Sood & Munish Saini, "Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation", Education and Information Technologies, Springer, 2020, Pages 1-16

[3] Şeyhmus Aydoğd, "Predicting student final performance using artificial neural networks in online learning environments", Education and Information Technologies, Spriger, Volume 25, 2020, Pages 1913-1927

[4] Edward Wakelam, Amanda Jefferies, Neil Davey, Yi Sun, "The potential for student performance prediction in small cohorts with minimal available attributes", British journal of educational technology, Wiley, Volume 51, Issue 2, 2020, Pages 347-370

[5] Bindhia K. Francis & Suvanam Sasidhar Babu, " Predicting academic performance of students using a hybrid data mining approach", Journal of Medical Systems, Springer, volume 43, 2019, Pages 1-15

[6] Erbug Celebi, and Rami S. Alkhawaldeh, "EMT: ensemble meta-based tree model for predicting student performance, Scientific Programming, Hindawi, Volume 2019, February 2019, Pages 1-13

[7] Magdalena Cantabella, Raquel Martinez-España, Belen Ayuso, Juan Antonio Yanez and Andres Munoz, "Analysis of student behavior in learning management systems through a big data framework", Future Generation Computer Systems, Elsevier, Volume 90, January 2019, Pages 262-272

[8] Hua Tang, Yueting Xu, Aiju Lin, Ali Asghar Heidari, Mingjing Wang, Huiling Chen, Yungang Luo and Chengye Li, "Predicting Green Consumption Behaviors of Students Using Efficient Firefly Grey Wolf-Assisted K-Nearest Neighbor Classifiers", IEEE Access, Volume 8, February 2020, Pages 35546 - 35562

[9] Hajra Waheed, Saeed-Ul Hassan, Naif RadiAljohani, Julie Hardman, Salem Alelyani, Raheel Nawaz, "Predicting academic performance of students from VLE big data using deep learning models", Computers in Human Behavior, Elsevier, Volume 104, 2020, Pages 1-34

[10] Hassan Zeineddine, Udo Braendle, Assaad Farah, "Enhancing prediction of student success: Automated machine learning approach", Computers & Electrical Engineering, Elsevier, Volume 89, January 2021, Pages 1-10

[11] Roshani Ade, "Students performance prediction using hybrid classifier technique in incremental learning", International Journal of Business Intelligence and Data Mining, 2019, Volume 15 Issue 2, Pages 173 – 189

[12] Abdullah Alshanqit and Abdallah Namoun, "Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification", IEEE Access, Volume 8, 2020, Pages 203827 – 203844

[13] Xizhe Wang, Xiaoyong Mei, Qionghao Huang, Zhongmei Han , Changqin Huang, "Fine-grained learning performance prediction via adaptive sparse self-attention networks", Information Sciences, Elsevier, Volume 545 , 2021, Pages 223–239

[14] Hanan Abdullah Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems", IEEE Access, Volume 8, 2020, Pages 55462 – 55470

[15] L. Ramanathan, G. Parthasarathy, K. Vijayakumar, L. Lakshmanan & S. Ramani, "Cluster-based distributed architecture for prediction of student's performance in higher education", Cluster Computing, Springer, Volume 22, 2019, Pages1329-1344

[16] Mudasir Ashraf, Majid Zaman, Muheet Ahmed, "An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches", Procedia Computer Science, Elsevier, Volume 167, 2020, Pages 1471-1483

[17] Alberto Rivas, Alfonso González-Briones, Guillermo Hernández, Javier Prieto, Pablo Chamoso, "Artificial neural network analysis of the academic performance of students in virtual learning environments", Neurocomputing, Elsevier, Volume 423, 2021, Pages 713-720

[18] Lubna Mahmoud Abu Zohair, "Prediction of Student's performance by modelling small dataset size", International Journal of Educational Technology in Higher Education, Springer, Volume 16, 2019, Pages 1-18

[19] Sahar Al-Sudani & Ramaswamy Palaniappan, "Predicting students' final degree classification using an extended profile", Education and Information Technologies, Springer, Volume 24, 2019, Pages 2357–2369

[20] Yeongwook Yang, Danial Hooshyar, Margus Pedaste, Minhong Wang, Yueh-Min Huang and Heuiseok Lim "Predicting course achievement of university students based on their procrastination behaviour on Moodle", Soft Computing, Springer, 2020, Pages 1-17