AN ENHANCED APPROACH FOR EDUCATIONAL DATA MINING WITH PL-SVM CLASSIFIER

Nikki Kumari
M.tech (Computer Science and Engineering (CSE))
School of Engineering and Technology
Jaipur, India
nikki051998@gmail.com

Dr. Savita Shiwani
Head of the Department, CSE
School of Engineering and Technology
Jaipur, India
convener.cse@jnujaipur.ac.in

***Abstract*— the learning algorithms are widely adopted by educational domain in order to measure the overall performance of the students on the basis of their academic records. The usage of automatic mechanism can ease the decision deriving process regarding the reputation of the organizations on the basis of the student's performance. The traditional student data mining mechanisms uses the feature extraction and classification mechanism to improve the efficiency of the system. The PCA and LDA feature extraction technique is found to be better than other feature extraction mechanism. Therefore, the author in this work develops a novel student data mining approach by using the SVM classifier and a hybrid feature extraction mechanism by using PCA and LDA technique. The proposed work is implemented in MATLAB simulation platform in the terms of F-Measure, Precision and Recall. The performance of the proposed work is evaluated on the basis of various feature extraction techniques such as Principal Component, Cfs Attribute. Relief Attributes, Gain Ratio Attributes, Chi Squared and Filtered Attributes. On the basis of the observed facts, the proposed work is proved more efficient than the other mechanisms.**

Keywords—*Educational Data Mining, Knowledge Discovery, Student's performance, Feature Extraction, Classifiers.*

I. INTRODUCTION

With the advancements in the technology, the modern education also gets advanced as the e-learning is widely adapted by the institutes to teach the students. It becomes significant to collect the student's data in order to evaluate their performance in the organization and to discover the learning process. The incremental growth of educational data results leads to the requirement for establishing the research in educational data mining (EDM). In other words, the EDM can be defined as a rising discipline related to the developing mechanism for discovering the unique types of data that is gathered from educational setting and then to utilize it to understand the overall performance of the students.

A large number of researches by various authors have been conducted in this field. For this purpose, the feature extraction mechanism was applied to extract the principal features from the gathered student's data and then the classifiers are used to classify the extracted data. On the basis of the observations from previous work, it has been concluded that the mechanism implemented for student data mining was less efficient and hence led to the lower accuracy rate. Thus, a novel approach has been developed in this work to enhance the accuracy for the predicted data. For this purpose, the feature extraction is done by collaborating the two of the most prominent feature extraction mechanism i.e. Linear Discriminant Analysis (LDA) ad Principal Component Analysis (PCA). The Hybrid feature extraction is applied to extract the relevant data efficiently. Along with this, the Straight Vector Machine (SVM) is applied for classifying the extracted features so that the more accurate decisions can be derived.

II. FEATURE EXTRACTION

2.1 LDA (*Linear Discriminant Analysis*):

LDA approach is employed for face recognition. It is a statistical approach used to compare unknown patterns with known patterns. This method use variables like continuous independent and category based dependent variable. This approach also used PCA for low dimension representation. LDA use classes which are based on database by dividing the database into number of classes. On the basis of segmented classes LDA perform various operations. The classes are randomly created by using sample database. LDA is utilized in the cases where the unequal frequencies are present and the requirement is to evaluate dynamically generated information. It provides maximum distinction among within-class and between-class variance. The fundamental difference between PCA and LDA: PCA performs feature classification; LDA have major contribution in data classification. Hence, it provides greater understanding of the feature data distribution.

Mechanisms followed by LDA are:

i) Class dependent transformation: An increase in ratio of between-class variance and within-class variance is obtained in order to provide maximum attainable value. The objective behind this enhancement is that

ratio is directly related to separation factor. The higher value of ratio provides greater distinction.

ii) Class independent transformation: In this approach, between-class variance is not considered. It aims at maximizing ratio of overall variance to within-class variance. It utilizes a unique optimizing mechanism on the data. Such that all the data points are transformed, without considering their class identity. Every class is treated as a distinct class with respect to other classes.

2.2 PCA (Principal Component Analysis):

It is basically a conversion method that converts number of associate variables into unassociated variables. It works smoothly and effectively in case of image classification as well as image compression. Mathematical calculations and functions are used for conversion of associate variables. It leads to brief description of data set. Its basic functions are implemented in form of principal components. To hold the data as flexible form, first principal component is to be considered and second principal component is maintained perpendicular to the first component's subspace. However, for the maximum divergence of subspace perpendicular to first and second component, third principal component is used. Principal Component Analysis basically acts as backup to IHS based conversion method. The basic reason behind this same working is mutual relationship between both techniques of MS band as are known as PC1, PC2 and PC3 and so on. Its working is given below:

(1) Initially, the IR is provided to the PAN and MS value, further re-sampling the MS.
(2) Then, conversion of MS bands to components such as PC1, PC2, and so on is carried out.
(3) Histogram links are provided between PAN and PC1.
(4) Then restore the PC1 with PAN.
(5) At the end PAN is converted to left principal components.

III. CLASSIFIER

Support Vector Machines (SVM) was derived from the Statistical Learning Theory which was introduced by Boser, Guyon and Vapnik in COLT-92. This technique was introduced to do the 2-class classification by efficiently dividing the data by using the N-dimensional hyper-plane. Thus, they are the part of *linear classifiers* and these are the examples of supervised learning mechanism. When it is required to deal with the hyper-planes in which the data is classified by reducing the error due to empirical classification, In Support Vector Machines the margins are increased and also it can be said that the increased amount of separation is considered [18].

Apart from the separating hyper-plane, if we consider 2 more hyper-planes which are parallel to the separating one and pass through the closest data points on each side; we end up with the "support hyper-planes" and these closest data points are named "support vectors" [19]. Margin of a linear classifier can then be defined as the width of the area between the two parallel hyper-planes.

Selecting the maximum margin is one of the targets of the Support Vector Machine classifiers due to the empirical issues and the fact that the chance of having a misclassification decreases in case of a small error in the boundary's location. Along with this the separating hyper-plane must be placed at the similar distance from support hyper-planes so that the probability of error in classification distributed in equal amount on both sides.

IV. PROBLEM FORMULATION

Previously, a lot of work is done to predict the performance of student using different feature selection techniques. In recent studies, researchers use different feature selection techniques and the combination of classifiers to produce efficient prediction models. A research is required to identify the performance analysis in terms of prediction accuracy in combination of different feature selection algorithms with differently classifiers. Moreover, advanced classifiers are required to be used for classifications. The techniques employed in the existing work provide less prediction accuracy, efficiency and effectiveness. Considering these issues in the existing work, a novel approach is required to identify the prediction accuracy of different available feature selection algorithm in the context of classifiers being used on educational data.

V. PROPOSED WORK

After reviewing the issues in the existing work, a novel approach is proposed. Initially, feature selection will be performed using two techniques such as Principal Component Analysis and Linear Discriminate Analysis. The collaboration of these techniques can perform effectively in terms of prediction accuracy. The advantage of PCA and LDA is that it extracts the features and also normalizes the data in an effective manner. After extracting the features the next step is to perform classification of the extracted data and in order to do so the SVM is used as a classifier in proposed work. The SVM is found to be more advantageous in comparison to other classifiers because it is has four advantages as follows: Firstly it has a regularization parameter, which makes the user think about avoiding over-fitting. Secondly it uses the kernel trick, so you can build in expert knowledge about the problem via engineering the kernel. Thirdly an SVM is defined by a convex optimization problem (no local minima) for which there is efficient methods (e.g. SMO). Lastly, it is an approximation to a bound on the test error rate, and there is a substantial body of theory behind it which suggests it should be a good idea.

The methodology of proposed work is as follows:

Step 1. Start
Step 2. First step is to form the dataset from the available information of the students.
Step 3. In this step, the collected data is normalized as per the information of the students.

Step 4. After normalizing the data, the tested data set is selected from original dataset. The tested dataset is used for testing purpose.

Step 5. On selected testing dataset, the PCA feature extraction is applied to extract the principle components from the dataset. Then the LDA mechanism is applied to PCA extracted features.

Step 6. Next step is to initiate the training and testing of extracted data. For this purpose, the SVM is used.

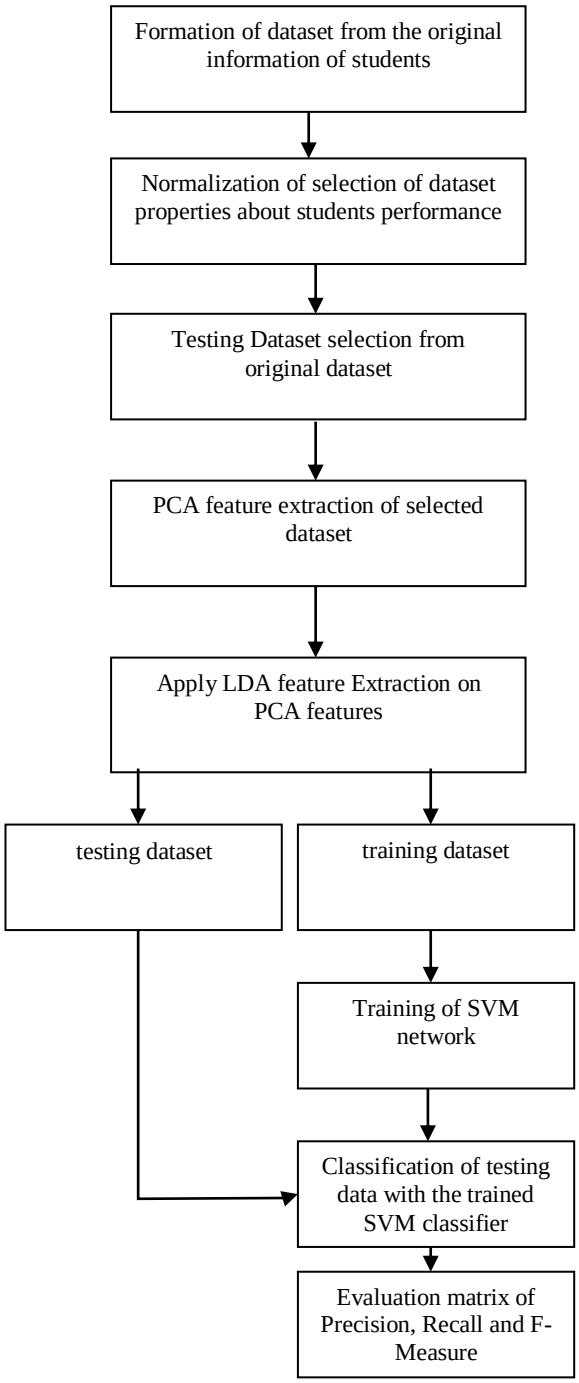

Figure 1 Methodology of Proposed Work

In this step, first of all the training of SVM is done and then the classification of tested data with respect to the trained data is performed.

Step 7. After this, the evaluation of the proposed work is done in the form of precision, F-Measure and recall.

VI. EXPERIMENTAL RESULTS

In this study, the SVM classifiers are used to classify the student's data. Before classification, the LDA and PCA are applied for feature extraction. The performance of the proposed work is evaluated in the terms of Precision, Recall and F-measure. Precision is known as positive predictive value. It is measured as follows:

$$Precision = \frac{TP}{(TP + FP)} \ldots\ldots (1)$$

Where TP is true positive, FP is false positive.

Recall defines that how many relevant items are elected. The formulation given below is used for evaluating the recall for proposed work:

$$Recall = \frac{TP}{(TP + FN)} \ldots\ldots (2)$$

F-Measure is a performance matrix that is used to evaluate the harmonic mean of precision and recall. The formulation is as follows:

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision + Recall)} \ldots\ldots (3)$$

The comparison of proposed work is done with traditional classification algorithms in terms of F-Measure, Precision and Recall. The traditional feature selection algorithms such as Principal Component Analysis, Gain Ratio, Cfs Subset, Chi Squared, Relief Attribute etc. The graph in figure 2 defines the comparison of classification algorithms on the basis of Principal component analysis.

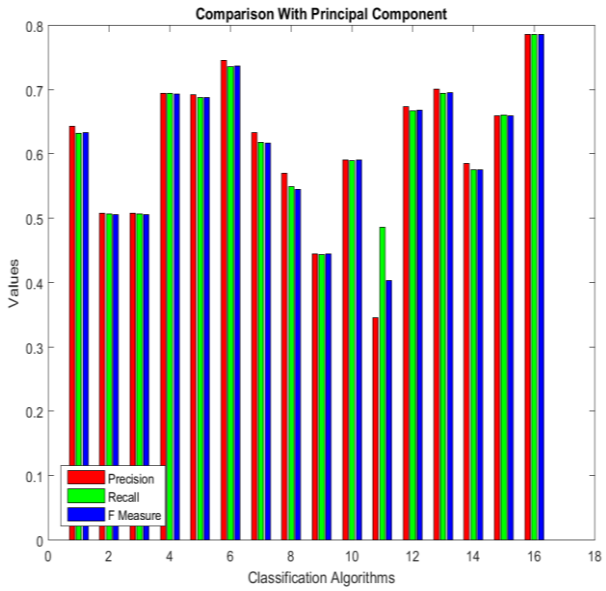

Figure 2 Comparison analysis of Principal Component

On the basis of the observations, it is concluded that the precision, F-Measure and Recall of the proposed work is higher than others. Similarly, the graph in figure 3 defines the comparison of classification algorithms with gin ratio and the observations of the proposed work are better than others.

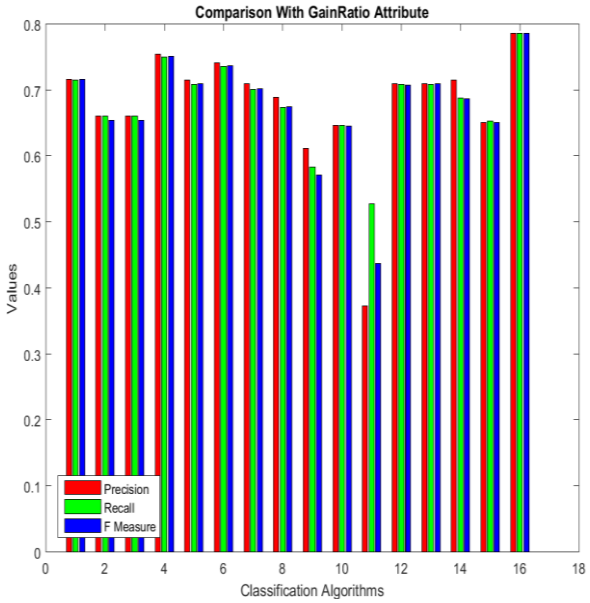

Figure 3 Comparison analyses of Gain Ration Attributes

The graph of figure 4 delineates the performance of proposed work and traditional work with respect to the various classifiers. The performance is measured in the terms of Precision, F-Measure and Recall. On the basis of the graph it is observed that the precision of 11th classifier ("DS") is 0.373, F-Measure of same is 0.528. The Recall of 9th classifier ("OneR") is 0.571. Thus, it can be concluded that the F-Measure and Precision of 11th classifier is lower whereas the recall of 9th classifier is lower and in comparison to these results, the proposed work has better and higher values corresponding to precision, F-Measure and Recall i.e. 0.785714.

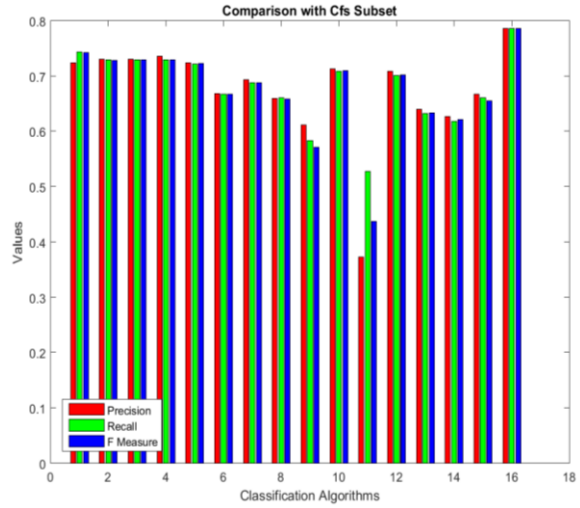

Figure 4 Comparison analyses of Cfs Subset

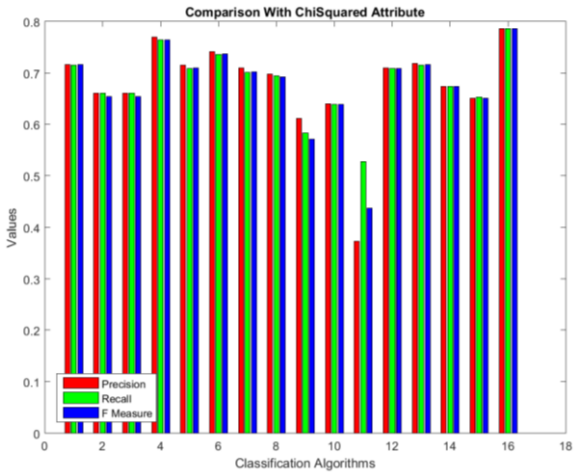

Figure 5 Comparison analyses of ChiSquared Attributes

The comparison for Chi Squared feature selection algorithm with other classifiers and proposed work is shown in figure 5. The graph proves that the proposed work also outperforms in this case as well. Similarly, the graph in figure 6 and 7 represent the comparison analysis for filtered attribute and relief attribute.

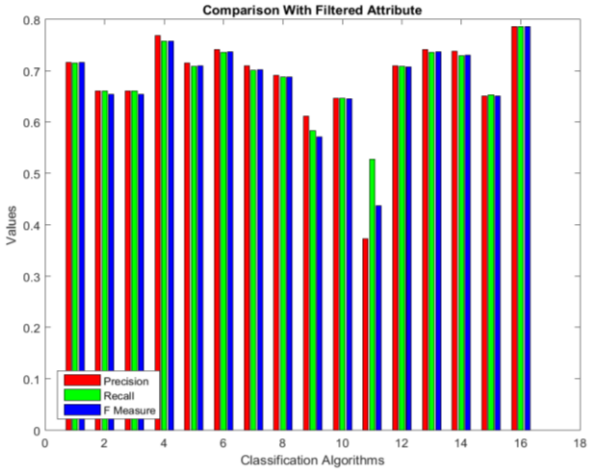

Figure 6 Comparison analyses of Filtered Attributes

The lowest Precision, F-measure and Precision in figure 6 are 0.373, 0.528 and 0.437 respectively.

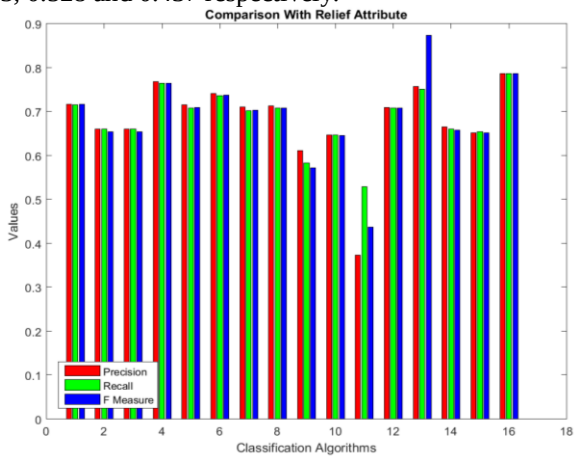

Figure 7 Comparison analyses of Relief Attributes

The graph in figure 7 delineates the performance with respect to the F measures. The facts and figures that are gathered from above defined graphs are summarized in table 1.The data in table 1 describes the overall performance of the various considered feature selection mechanisms.

Table1 Overall Performance evaluation of proposed work with respect to other classification algorithms

Techniques	Precision	F-Measure	Recall
PCA	0.5995	0.6032	0.5975
Gain Ratio	0.6707	0.6745	0.6670
Cfs Subset	0.6669	0.6731	0.6662
Chi Squared	0.6697	0.6759	0.6683
Filtered Attribute	0.6707	0.6745	0.6670
Relief Attribute	0.6707	0.6745	0.6670
Proposed Work	0.785714	0.785714	0.785714

VII. CONCLUSION

The education data mining or student data mining plays a vital role to evaluate the performance of a student in the institute. This is done with an objective to find out the overall performance of the students with respect to the given parameters. The student data mining utilizes the concepts of data mining to measure the performance. This is done by using various feature extraction and classification techniques in past. Some of the feature extraction and classification techniques such as PCA, ANN and KNN are widely adopted for this purpose. This study develops a mechanism for student data mining by using the hybrid feature extraction mechanism and SVM classifier. Here hybridization is done by using the LDA and PCA feature extraction techniques. The performance evaluation is done on the basis of traditional feature extraction technique and it is observed that the proposed work outperforms the traditional classifiers. The overall precision, F-measure and recall of the proposed work is 0.785714 respectively. In future, this work can be enhanced in future by working on it in order to reduce the complexity of the mechanism and to increase the accuracy level of generated decisions.

REFERENCES

[1] Raheela Asif, Agathe Merceron, Syed Abbas, Alic Najmi, Ghani Haidera, Analyzing undergraduate students' performance using educational data mining, ELSEVIER, vol 113, Pp 177-194, 2017.

[2] Charoula Angeli, Sarah K.Howard, Jun Mab Jie, Yangb Paul, A. Kirschnercd, Data mining in educational technology classroom research: Can it make a contribution?, ELSEVIER, vol 113, Pp 226-242, 2017.

[3] Evandro B. Costaa Baldoino , Fonsecaa Marcelo, Almeida Santanaa Fabrísia, Ferreirade Araújob, Joilson Regod, Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses, ELSEVIER, vol 73, Pp 247-256, 2017.

[4] AlejandroPeña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, ELSEVIER, vol 41,Pp 1432-1462, 2014.

[5] ManolisChalarisStefanosGritzalisManolisMaragoudakisCleoSgouropoulouAnastasiosTsolakidis, Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques, ELSEVIER, vol 147, Pp 390-397, 2014.

[6] Surjeet Kumar Yadav, Data Mining Applications: A comparative Study for Predicting Student's performance, ijitce, vol 1(12), Pp 13-20.

[7] Surjeet Kumar Yadav, Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, WCSIT, vol 2(2), Pp 51-56, 2012.

[8] SAYALI RAJESH SUYAL,"Quality Improvisation Of Student Performance Using Data Mining Techniques, vol 4(4), Pp 1-4, 2014.

[9] Paulo Cortez, Using Data Mining To Predict Secondary School Student Performance, 2008.

[10] Muluken Alemu Yehuala, Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre Markos University), vol 4(4), Pp 91-95.

[11] Surjeet Kumar Yadav, "Data Mining Applications: A comparative Study for Predicting Student's performance", ijitce, vol 1(12), Pp 13-20.

[12] Surjeet Kumar Yadav, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", WCSIT, vol 2(2), Pp 51-56, 2012.

[13] Brijesh Kumar Baradwaj, "Mining Educational Data to Analyze Students‟ Performance", ijacsa, vol 2(6), Pp 63-70, 2011.

[14] Ajinkya Kunjir," Recommendation of Data Mining Technique in Higher Education", IJCER, vol 5(3), Pp 29-35, 2015.

[15] Jayashree M Kudari," Survey on the Factors Influences the Students‟ Academic Performance", IJERMT, vol 5(6), Pp 30-37, 2016**,**

[16] K. Amarendra "Research on Data Mining Using Neural Networks" Special Issue of International Journal of Computer Science & Informatics (IJCSI), , Vol.- II, Issue-1, 2, Pp2231–5292.

[17] Anand V. Saurkar, "A Review Paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engi neering, Volume 4, Issue 4,Pp 98-101, 2014