Educational Data Classification using Data Mining and Kernel Ensemble Classifier

Sheo Kumar

Abstract*: The success of students gives the good name for institution and it become popular. Due to the large number of student's database it is difficult to identify the performance and activities of each student. The educational data mining is used to identify the performance and status of the students individually. In this study, the Educational Data Classification (EDC) using data mining technique and kernel ensemble classification using Support Vector Machine (SVM) based kernels like linear, polynomial, quadratic and Radial Basis Function (RBF) is discussed. Initially the data preprocessing is made to remove the raw data into understandable format. The SVM kernels like linear, polynomial, quadratic and radial basis function based ensemble classifier is used for classification of student's data. The data mining is used for making final decision of student's performance in class like activities and interaction with electronic learning system. The performance of the system is evaluated by kalboard 360 database. The performance of the system is made by classification accuracy of 72.52% using SVM kernel ensemble classification

***Keywords*: Educational Data Classification, Data mining, SVM kernels, kalboard 360 database**

I. INTRODUCTION

The education system uses data mining to make the final decision and identify the status of the students. Educational data mining multiclass classification for imbalanced class handling and data is described in [1]. The data's are prepared using data acquisition method. Then the imbalance class is changed into balanced class in the data preprocessing by using majority, minority class, oversampling and under sampling. The one sided selection and synthetic minority oversampling technique are also used in preprocessing stage. SVM classifier is used to classify the data's. EDM analysis using classification is described in [2]. The classification techniques like SVM, J48 algorithm, random forest, naïve bayes classifier and multilayer perceptron is used for classification. The final decision is made by data mining method.

Data mining algorithm for classification to predict learns in EDM is discussed in [3]. The raw data's are clean the inconsistent and incomplete data. The data mining method is used to predict the output and analyzed. EDC using naïve bayes classifier is discussed in [4]. The data's are classified using naïve bayes classifier and the data mining algorithm is used for prediction.

Revised Manuscript Received on December 11, 2019.
***** Correspondence Author
Dr Sheo Kumar*, Professor, Department of Computer Science and Engineering, CMR Engineeringg College, Hyderabad, India. Email:sheo2008@gmail.com

Prediction of slow, average and fast learners is discussed. The input dataset are given to data transformation for data preprocessing. The preprocessed data is given to naïve bayes, J48, random tree and zeroR algorithm for classification. Educational data mining based on feature selection is discussed in [5]. The features like Chi Squared Attribute Eval, Cfs Subset Eval, Filtered Attribute Eval, Principal Components, Relief Attribute Eval and Gain Ratio Attribute Eval are selected. Classification is made by naïve bayes, multiple learning perceptron, bayesnet, decision table, simple logistic, J48 and random forest classifier are used to predict the performance of students.

Educational data mining based students prediction is described in [6]. The student's whole information are maintained and predicted easily by using the data mining. The classifiers like decision tree, regression and neural network is used for the prediction of students. EDC using particle swarm method and high dimensional educational data is described in [7]. Initially, the input text data's are preprocessed to remove the raw data. Then the particle swarm optimization technique is used for the classification of student's data.

EDC using association rule and data analytics is described in [8]. At first, the association rule is applied for admission data for knowledge identification. Decision tree classifier is used for the classification and data mining algorithm is used for prediction. Prediction of EDC for student's performance is discussed in [9]. The input student's database is preprocessed to remove raw data. Then it was given to c4.5 and improved id3 algorithm then evaluation classifier is used for classification. The prediction model is used for performance evaluation.

Performance analysis of students using classification techniques is described in [10]. Initially, the student's data is given to data preprocessing technique to remove missing values and categorization, the gain ratio is used for feature selection to select attributes. The classification is made by decision tree algorithm. Then data mining technique is used for prediction. High dimensional educational data for feature selection of hybrid scheme is described in [11]. The feature selection methods like wrapper and filter method is given to input student's data. Then the stopping criteria method is used for the performance evaluation and final result prediction is made by data mining method.

EDC using data mining and kernel ensemble classification is presented. The organization of paper is follows: Section 2 describes the methods and materials of EDC system used in this study.

The results and discussion of EDC system is described in section 3. The last section concludes the EDC system.

II. METHODS AND MATERIALS

EDC based on data mining and kernel SVM is shown in figure 1. At first the data preprocessing is made for the input student data to transform the raw data. Then classification is made by SVM kernels like linear, polynomial, RBF, quadratic classifier. The student's activities and performance is accurately predicted by data mining technique.

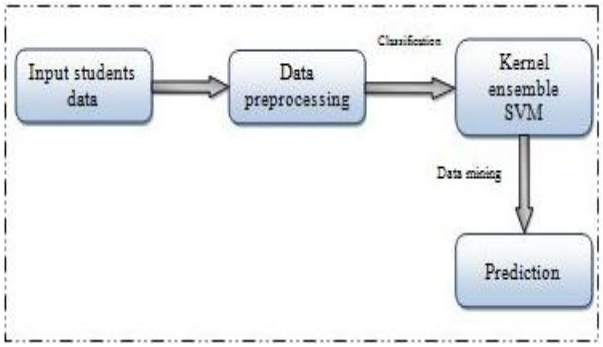

Fig. 1. EDC using data mining and kernel SVM

A. Students Data Preprocessing

The data preprocessing is a necessary step involved in data mining. It is process of transforming raw data into clear format. The data is inconsistent, incomplete, lacking in certain behaviors and it also contains many errors. These errors are solved by data preprocessing technique. It also transfers the raw data in efficient and useful format. The data preprocessing involves in three steps include: data cleaning, data transformation and data reduction. Figure 2 shows the steps involved in data preprocessing.

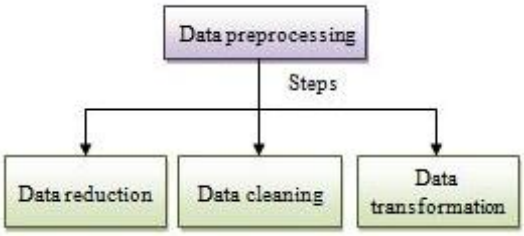

Fig. 2. Data preprocessing steps

Data cleaning is used to clean the missing parts and irrelevant data. The noisy data and missing data are handled by data cleaning. The meaningless data is known as noisy data, and it can be generated due to data entry errors and fault data collection is known as noisy data. If some data's are missed in a data is known as missing data. These noisy and missing data errors are cleared by using data cleaning method. The data transformation has four different ways to transform data it includes: normalization, attribute selection, discretization, concept hierarchy generation is shown in figure 3. These four ways are used to transform the data for suitable data mining process.

The huge amount of data is used in data mining technique, in some cases the huge volume of data is became harder. The data reduction technique is used to reduce this type of data. It reduces the data storage and increases storage efficiency and cost analysis. The data preprocessing technique is also used in clinical data applications [12] and safety equipment linkage system [13]. In this study the data preprocessing technique is used to reduce the raw data into understandable format.

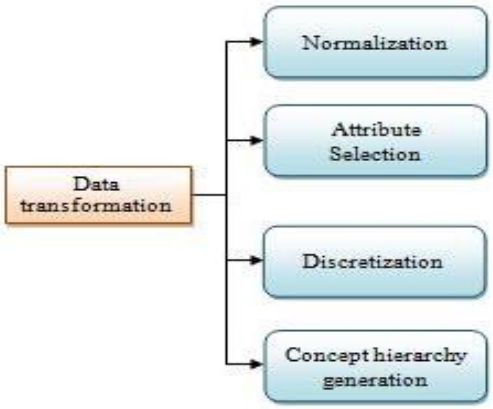

Fig. 3. Different ways of data transformation

B. Kernel Ensemble Classification

In this study, the ensemble classification using SVM kernels like linear, polynomial, quadratic and RBF are made for EDC system. The binary classification is made by SVM. Consider, the training sets of image in class K, $\{k_1, k_2, k_3, ...k_n\}$ where $K \subset P^n$. The δ is the mapping function, and then $\delta : K \to P$ is the feature space. The equation is given by,

$$\min\left[\frac{1}{2}\|k\|^2 + \frac{1}{mn}\sum_{l=1}^{n}\beta_l - h\right] \quad (1)$$

where $k.\beta \geq p - \beta$, let $l = 1,2,...n$ and $\beta_l \geq 0$. β is the slack variable, h is the bias and l is the number of samples. The SVM kernel functions like linear, polynomial, quadratic and RBF is used for ensemble classification of student's data. SVM kernel ensemble classification is also used in hyper spectral chemical plume detection [14], Cancer classification from gene expression [15] and plant leaf recognition [16]. Figure 4 shows the SVM kernel ensemble classification.

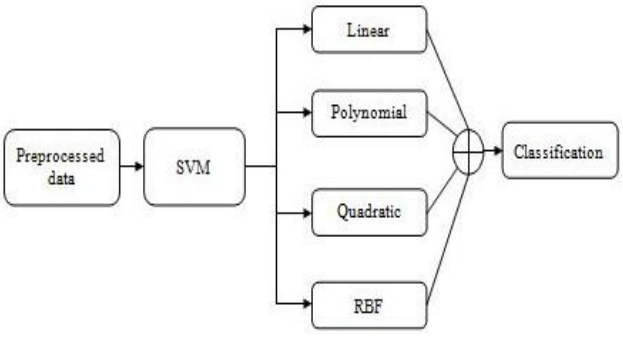

Fig. 4. Kernel ensemble classification SVM

C. Prediction using Data mining

The data mining is used to extract the data from large dataset. In education the data mining is refers to tools, techniques and research for automatic extraction of data based on student's learning and educational settings.

Data mining is a process of finding new information from the large amount of data and later it can be used. The data mining is also used in analysis models of technical and economic data [17] and safety monitoring system of coal mine [18] In this study, the classified data's are used to predict the performance of students identify the status of each students

III. RESULTS AND DISCUSSION

The performance of the system is evaluated by e-Learning system called Kalboard 360 using Experience API web service (XAPI). The database is based on educational domain. The dataset consist of 16 different features with 480 student's records. The database is collected through the tool called API (xAPI). The features include demographic, educational and psychological features. The performance of the system is measured by classification accuracy. Table 1 shows the performance of EDC system.

SVM kernels	Student's activity accuracy (%)			
	Raised hands	Visited resources	Announcement	Discussion
Linear	63.00	67.50	69.00	72.50
Polynomial	69.00	68.00	72.50	76.50
Quadratic	72.50	71.00	73.00	78.00
RBF	76.00	73.00	78.00	81.00
Ensemble classification (%)	70.12	69.87	73.12	77.00
			Average (%)	**72.52**

From the table 1 it is observed that the overall classification accuracy of 72.52 % obtained by using the SVM kernels. The activity of the students is high in discussion and its classification accuracy is 77.00 %. The minimum classification accuracy of 69.87% obtained in visited resources and also the classification accuracy of visited resources and announcements is 69.87% and 73.12%. Figure 5 shows the performance metrics of EDC system.

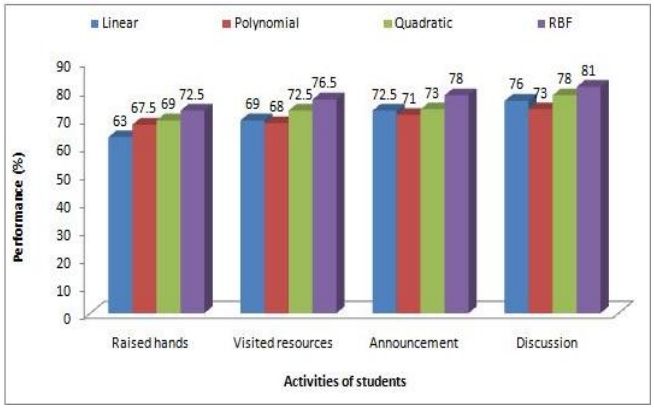

Fig. 5. Performance metrics of EDC system

From the above figure it is clearly observed that the highest classification accuracy is produced only in RBF kernel comparing with other kernels. The lowest classification accuracy is produced by linear kernel.

IV. CONCLUSION

A novel method for EDC system using data mining and SVM based kernel ensemble classification is discussed. Initially, the data preprocessing is made to remove raw data in the database and change into readable format. The ensemble classification is performed using the SVM kernels like linear, polynomial, quadratic and RBF. Then the activity of the students is predicted by using data mining technique. The data mining technique is used to predict the accurate activity of the students. The performance of the system is evaluated by using Kalboard 360 database. The EDC system produces the overall classification accuracy of 72.52% using SVM kernel ensemble classification.

REFERENCES

1. Y. Pristyanto, I. Pratama, and A.F. Nugraha, "Data level approach for imbalanced class handling on educational data mining multiclass classification", *International Conference on Information and Communications Technology*, 2018, pp. 310-314.
2. C. Jalota, and R. Agrawal, "Analysis of Educational Data Mining using Classification", *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing*,2018, pp. 243-247.
3. V. Mhetre, and M. Nagar, "Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA", *International Conference on Computing Methodologies and Communication*, 2017, pp. 475-479.
4. A. Dangi, and S. Srivastava, "Educational data classification using selective Naïve Bayes for quota categorization", *IEEE International Conference on MOOC, Innovation and Technology in Education*, 2014, pp. 118-121.
5. M. Zaffar, M.A. Hashmani, and K.S. Savita, "Performance analysis of feature selection algorithm for educational data mining", *IEEE Conference on Big Data and Analytics*, 2017, pp. 7-12.
6. T. Devasia, T.P. Vinushree, and V. Hegde, "Prediction of student's performance using Educational Data Mining", *International Conference on Data Mining and Advanced Computing*, 2016, pp. 91-95.
7. A.A. Yahya, and A. Osman, "Classification of high dimensional Educational Data using Particle Swarm Classification" *International Conference on Computer Systems and Applications*, 2014, pp. 34-41.
8. P. Rojanavasu, "Educational Data Analytics using Association Rule Mining and Classification" *International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*, 2019, pp. 142-145.
9. R. Patil, S. Salunke, M. Kalbhor, and R. Lomte, "Prediction System for Student Performance Using Data Mining Classification", *International Conference on Computing Communication Control and Automation*, 2018, pp. 1-4.
10. V. Shanmugarajeshwari, and R. Lawrance, "Analysis of students' performance evaluation using classification techniques", *International Conference on Computing Technologies and Intelligent Data Engineering*, 2016, pp. 1-7.
11. U. Ali, K.S. Arif, and U. Qamar, "A Hybrid Scheme for Feature Selection of High Dimensional Educational Data", *International Conference on Communication Technologies*, 2019, pp. 71-75.
12. Q. Ang, Z. Liu, W. Wang, and K. Li, "Explored research on data preprocessing and mining technology for clinical data applications", *International Conference on Information Management and Engineering*, 2010, pp. 327-330.
13. X. Cheng, "Research on Data Preprocessing Technology in Safety Equipment Linkage System", *International Conference on Computational and Information Sciences*, 2013, pp. 1713-1716.
14. P. Gurram, and H. Kwon, "A full diagonal bandwidth gaussian kernel SVM based ensemble learning for hyperspectral chemical plume detection", *International Geoscience and Remote Sensing Symposium* , 2010, pp. 2804-2807.
15. S. Begum, D. Chakraborty, and R. Sarkar, "Cancer classification from gene expression based microarray data using SVM ensemble", *International conference on condition assessment techniques in electrical systems*, 2015, pp. 13-16.

16. Phiros Mansur, "Plant Leaf Recognition System Using Kernel Ensemble Approach", *International Journal Of Advances In Signal And Image Sciences*, Vol. 4, No.1, 2018, pp. 30-36.
17. J. Ming, L. Zhang, J. Sun, and Y. Zhang, "Analysis models of technical and economic data of mining enterprises based on big data analysis", *International Conference on Cloud Computing and Big Data Analysis*, 2018, pp. 224-227.
18. C.A. Shao, Q. Wu, and X. Guan, "The Research on Safety Monitoring System of Coal Mine Based on Spatial Data Mining", *International Workshop on Knowledge Discovery and Data Mining*, 2009, pp. 126-129.

AUTHORS PROFILE

Dr Sheo Kumar is associated with CMR Engineering College as Professor and Head, Dept of Computer Science & Engineering. He has 24 years of academic experience including 5 years of research. He has one Patent in his name. He is very active in research & review. His research area includes data mining and data science.