Mehran University Research Journal of Engineering and Technology
Vol. 39, No. 4, 772 - 778, October 2020
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.2004.09

Role of FCBF Feature Selection in Educational Data Mining

Maryam Zaffar[1], Manzoor Ahmad Hashmani[2], K.S. Savita[3], Syed Sajjad Hussain Rizvi[2]
Mubashar Rehman[5]

RECEIVED ON 21.04.2019, ACCEPTED ON 26.07.2019

ABSTRACT

The Educational Data Mining (EDM) is a very vigorous area of Data Mining (DM), and it is helpful in predicting the performance of students. Student performance prediction is not only important for the student but also helpful for academic organization to detect the causes of success and failures of students. Furthermore, the features selected through the students' performance prediction models helps in developing action plans for academic welfare. Feature selection can increase the prediction accuracy of the prediction model. In student performance prediction model, where every feature is very important, as a neglection of any important feature can cause the wrong development of academic action plans. Moreover, the feature selection is a very important step in the development of student performance prediction models. There are different types of feature selection algorithms. In this paper, Fast Correlation-Based Filter (FCBF) is selected as a feature selection algorithm. This paper is a step on the way to identifying the factors affecting the academic performance of the students. In this paper performance of FCBF is being evaluated on three different student's datasets. The performance of FCBF is detected well on a student dataset with greater no of features.

Keywords: **Educational Data Mining, Filter Feature Selection Algorithms, Fast Correlation-Based Filter Student Prediction Model, Support Vector Machine.**

1. INTRODUCTION

Student performance prediction models have received a significant amount of contemplation from both the research community and the educational sector. Student performance prediction model tackles the problem of student's grades [1], Grade Point Average (GPA) [2], Cumulative Grade Point Average (CGPA) [3] and Pass/Fail Course [4]. Thus the only goal of students' performance prediction models in EDM is not to achieve the high accuracy prediction model but also to help the educational stakeholders in predicting the performance of students, in order to make proactive decisions, and develop the strategies to enhance the quality of education for the improvement of students' academic performance. As the students are the main assets of any community, and the main aim of any academic organization is to provide quality education to its students. Moreover, quality education supports in building the skillful and

[1] Department of Computer and Information Sciences, Universiti Teknologi Petronas, Malaysia.
Email: maryam.zaffar82@gmail.com (Corresponding Author)
[2] High Performance Cloud Computing Centre, Universiti Teknologi Petronas, Malaysia.
Email: manzoor.hashmani@utp.edu.my
[3] Centre for Research in Data Science, Universiti Teknologi, Malaysia. Email: vitasugathan@utp.edu.my
[4] Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi, Pakistan.
Email: dr.sajjad@szabist.edu.pk
[5] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia.
Email: mobashar@utar.edu.my

featureful students. This gives attention to analyze the student's data in such a way to figure out the features affecting the performance of students. A lot of research is being conducted on the development of students' performance prediction models. But the study of students prediction models is still inadequate in predicting the performance of students [5]. This leads us to work on student prediction models to trace a suitable method for the development of student performance prediction model, to make proactive decision for the betterment of student's performance.

There are different techniques available for student performance prediction model. There are two main approaches to predict academic success whereas, one is supervised, and another is the unsupervised method. According to [6] around 71.4% of research articles on students' performance prediction models are using the classification method. It is the top method for performance prediction models [7]. In the classification method, the target variable is clearly defined as that it is predicted as grades, GPA, CGPA, or students PASS/FAIL. This leads us to build the students' performance prediction model with the help of the classification method as to figure out the dominant features affecting the student's final results.

Feature selection can play a prominent role in enhancing the accuracy of a prediction model. In the student's prediction model, where the selected features play not only an important role in increasing the prediction accuracy but also the base for the strategic plans for the educational environment. [8] deduced that information gain attribute evaluator is the best feature selection technique to improve the effectiveness of student prediction model. Whereas, [9] claims CFS subset evaluator as the best feature selection method for predicting the final semester examination performance of students. According to [10] there is no common feature selection method which can be accurate for all datasets even for a common domain. So that there is a need to figure out the important feature selection methods for predicting the performance of students. The importance of feature selection methods in predicting students' performance, motivated us to check the performance f feature selection for students' performance prediction.

There are mainly two types of feature selection methods, filter, and wrapper feature selection. Filter feature selection is being used and recommended by different studies in EDM. Filter feature selection is divided further into different types. In this paper, we focus one of the most important filter feature selection algorithm that is FCBF.

The contribution of this paper is that it checks the performance of FCBF on three different student's datasets. To give ease to the new researchers to know the performance of FCBF on datasets with different number of instances and different number of features. According to best of Knowledge, this is the first article in EDM that performs the evaluation of a filter feature selection FCBF on three different student datasets.

The outline of the paper is as follows. Section 2 describes the methods used in this research, section 3 discusses and describes results of filter feature selection algorithm on three datasets, and the conclusion of the paper is presented in section 4.

2. METHOD

In this study, the performance of the FCBF filter feature selection is evaluated on three different datasets of students. FCBF is applied to three datasets. The Support Vector Machine (SVM) classification algorithm is applied to the chosen datasets. The SVM classification algorithm is used to find the predictions. At the end, findings are evaluated. Prediction accuracy, F-measure, Precision and Recall are taken as the performance evaluation measures. Fig.1 describes the flow of main methodology of the proposed research presented in the paper. Three benchmark datasets DS33, DS16. And DS2 of students' academic records are selected to check the performance of FCBF. These datasets contain different number of instances, features and also belong to different educational domains. These three datasets are given as an input to FCBF feature selection algorithm one by one to select the features from the dataset. The dataset with selected features is then trained through SVM classification algorithm, and at the end tested and evaluated through performance evaluation measures (precision, recall, f-measure, and accuracy).

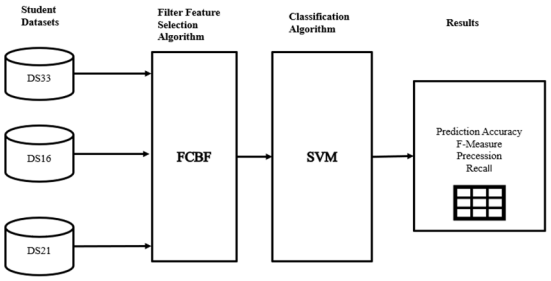

Fig. 1: Main Methodology

2.1 Description of Datasets

Three different student datasets are taken from different sources. The description of the three datasets is given below.

DS33: The DS33 datasets is a Portages secondary students school dataset. The dataset is has been used in different EDM studies [10-12]. This is dataset of 395 students taking Mathematics subject. The dataset includes 33 features having demographic, academic information and personal information of students.

DS16: The second dataset consists of 500 student's records. There are 16 features in the dataset including demographic details, academic details and behavioral features of the students. The dataset was previously used by Elaf [13].

DS21: The third dataset DS21 is collected from different colleges in India. The dataset consists of 300 students records and 21 features. The dataset was used by the study in [14].

2.2 Filter Feature Selection Algorithm

Feature selection is a significant pre-processing technique that is applied in machine learning methods. Feature selection is important in all other fields of research to make proper decision [15]. The filter feature selection is a type of feature selection that maximizes the evaluation function for getting the best feature subset through a search strategy [16]. There are three main stages of filters that are feature set generation, measurement and testing by a learning algorithm. The filter feature selection algorithms process quickly, and they calculate the information from the features so that their results will depend on measured information of the features [17]. The filter feature selection algorithms are chosen because they can accomplish better with any classification algorithm as they have a smaller amount of computational complexity [18].

FCBF was purposed by [19]. It is a multivariate feature selection method that attempts to discover the best feature subset based on goodness of features [19]. It starts with a full set of features and uses symmetrical uncertainty to calculate the dependence of features. Symmetrical Uncertainty (SU) is a normalized information theoretic measure which uses the values of entropy and conditional entropy to calculate the dependencies of features. FCBF is a correlation based feature subset selection method, which is faster than other subset selection methods [20]. In EDM, FCBF practiced ranking the features of graduate students in United States universities, to detect the factors of high dropout rate and low graduation rate of four-year college students [21]. Authors of reference [22] applied FCBF in pre-processing stage to predict the student interactions in the intelligent learning environment, furthermore, the study recommended that FCBF would be competent on selecting features from students dataset, as in this kind of datasets the correlation between features are very crucial.

2.3 Classification Algorithm

There are two main methods in data mining, one is supervised, and another is unsupervised. Classification is a type of supervised method. According to the existing literature on EDM, it is most frequently used in predicting the performance of students [23-25]. There are quite a lot of classification algorithms available that are being used in student performance prediction models such as Decision Tree, Neural Network, Naïve Bayes, Random Forest, Ada Boost and SVM. In this research work SVM as the classification algorithm is used for students' performance prediction.

SVM: SVM is a type of classification algorithm. It has been applied in a number of research works including face recognition, 3D (Three-Dimensional) object recognition, text and image classification and in EDM. It has an inimitable benefit of solving small-sample, on-linear, and high dimensional pattern recognition

problems [26]. SVM practices a Gaussian function. As an assistance, the complex relationship between the given data points can be captured. SVM is appropriate for feature selection hitches [27]. In this research, we have used SVM linear Kernel. Equation (1) presents the linear kernel whereas x_i is representing data points. The SVM linear kernel classification is very simple and training with the data with linear kernel of SVM is faster than any other kernel.

Linear Kernel: $K(\mathbf{x_i},\mathbf{x_j})= \mathbf{x_i}^T\mathbf{x_j}$ (1)

3. RESULTS AND DISCUSSION

In this section, we present the results of FCBF filter feature selection on all the three datasets DS33, DS16 and DS21. The results are evaluated on different evaluation measures. First, we show the previous results of FCBF using one-fold cross-validation on three different student. The datasets 1 and 2 have almost two same categories of features that are Demographic (DF), and Academic (AF), whereas dataset1 has Lifestyle Information (LF1), and dataset2 has behavioral features (BEF) category. Dataset2 has also included the features regarding parent's participation in the learning process (PPL). Whereas, the features of the third dataset has features regarding demographic, academic and socio-economic information of students. FCBF shows the highest accuracy on dataset1. Whereas, FCBF shows lowest on dataset 3, that have the lowest number of instances among all 3. The results show that the academic background is a very important category of features for predicting the performance of students. Whereas, student behavior and socio-economic factors also influence the performance of the student. This motivated us to check the performance of FCBF on three datasets using 10-cross-validation and through different measures.

3.1 Prediction Accuracy of three Datasets

Fig. 2 presents the comparison of prediction accuracy by using FCBF on three selected student's datasets. The results show that the FCBF shows better accuracy on the dataset DS33, whereas shows lowest performance on the dataset with less number of instances that is DS21 whereas accuracy is the ratio between all correct predictions.

The accuracy is defined as

$$\text{Prediction Accuracy} = \frac{TP+FN}{TP+FN+FP+FN} \quad (2)$$

Table 1: Previous results of FCBF on three Datasets using 1-Cross-Validation [28]

D	A	TNF	FE	NF
1	87.97	34	2	G2 (Second Period Grade), travel time (home to school travel time)
2	72.92	16	4	Relation, Visited Resources (BEF), Parent Answering Survey (PPL), Student Absence Days (AF)
3	45	21	3	GE (Gender), TNP (Class X%), FMI (Family Monthly Income)

where: D = Datasets, A = Accuracy, TNF = Total Number of Features, FE = Features Extracted, NF = Name of Features

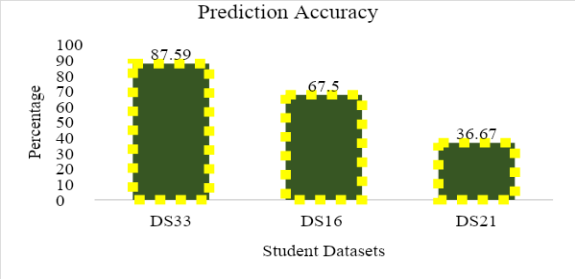

Fig 2: Comparison of Prediction Accuracy of FCBF on three datasets

3.2 F-Measure on Three Datasets

Fig. 3 demonstrates the comparison of F-Measure percentage on three student's datasets when we apply FCBF on them. It is observed through the results that the value of F-measure is 90% on DS33, whereas only 23% on DS21. But FCBF shows average results in terms of F-measure on DS16. F-measure is described through Equation (3).

$$F-\text{Measure} = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision}+\text{Recall})} \quad (3)$$

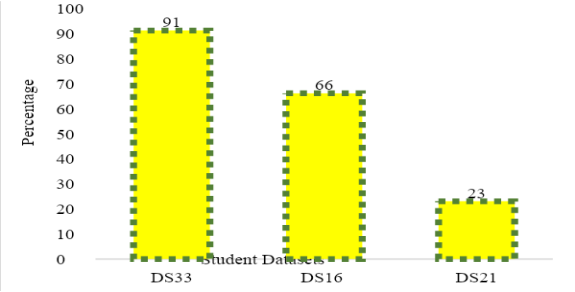

Fig. 3: Comparison of F-Measure of FCBF on three datasets

3.3 Precision on Three Datasets

Fig. 4 presents a comparison between the values of FCBF precision percentage on three different datasets of students. It is observed that the FCBF shows better results in terms of precision on DS33. Whereas the worst results are being observed on DS21.

Precision is the fraction of the retrieved instances that belong to the target class. The precision formula is presented through Equation (4).

$$\text{Predcision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

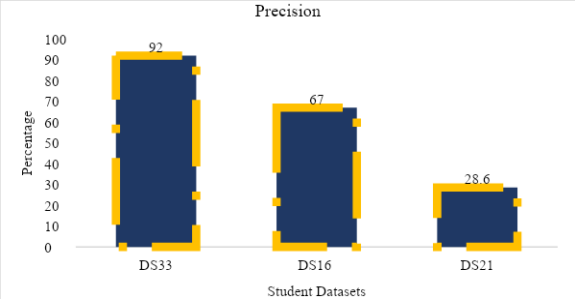

Fig. 4: Comparison of Prediction of FCBF on three datasets

3.4 Recall on Three Datasets

The Fig. 5 shows a comparison of the recall measure of FCBF. The results show that FCBF performs out class on DS33, whereas better results are not observed on DS21. However, it shows 66% Recall value on DS16.

The recall formula is presented through Equation (5).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

Fig. 5: Comparison of Recall of FCBF on three datasets

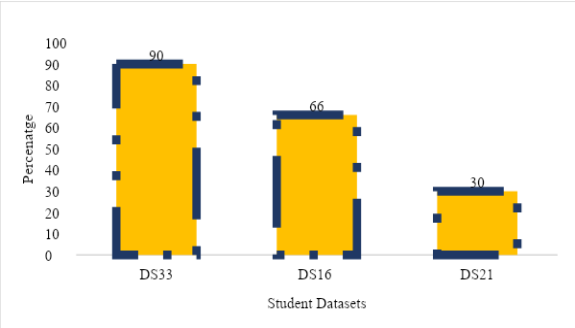

Fig. 5: Comparison of Recall of FCBF on three datasets

4. CONCLUSION

Student performance prediction is a very important area of research because this area is not only an interesting field for the researchers in EDM but also it is beneficial for all the educational stakeholders. Feature selection helps EDM in developing a high accuracy students prediction model. In this paper, we have evaluated the performance of FCBF. The performance of FCBF in terms of accuracy, f-measure, precision, and recall shows out class results on DS33. Whereas, perform not up to the mark on DS21. The results deduced that FCBF performs satisfactorily with a student dataset of large number of features. Moreover, FCBF does not give good results on a dataset with less number of instances. So, it is recommended to use FCBF feature selection on a dataset with large number of features. In future, we will evaluate different feature selection algorithms on student's dataset to evaluate their performance.

ACKNOWLEDGEMENT

This research work is supported by University Teknologi Petronas (UTP), Malaysia.

REFERENCES

[1] Al-Barrak, M.A., Al-Razgan, M.S., "Predictin Students' Performance through Classification: A Case Study", *Journal of Theoretical and Applied Information Technology*, Vol. 75, pp. 2, 2015.

[2] Aziz, A.A., Ismail, N.H., Ahmad, F., and Hassan, H., "A Framework for Students' Academic Performance Analysis Using Naïve Bayes Clasifier", *Journal Teknologi (Sciences & Engineering)*, Vol. 75, No. 3, pp. 13-19, 2015.

[3] Buniyamin, N., Bbin Mat, U., and Arshad, P.M., "Educational Data Mining for Prediction and Classification of Engineering Students Achievement", *Proceedings of the 7th IEEE International Conference on Engineering Education*, pp. 49-53, 2015.

[4] Ramanathan, L., Dh, S., and Kumar, S., "Predicting Students' Performance Using

Modified ID3 Algorithm", *International Journal of Engineering and Technology*, Vol. 5, No. 3, June, 2013.

[5] Shahiri, A.M., and Husain, W., "A Review on Predicting Student's Performance Using Data Mining Techniques", *Procedia Computer Science*, Vol. 72, pp. 414-422, 2015.

[6] Del Río, C.A., and Insuasti, J.A.P., "Predicting Academic Performance in Traditional Environments at Higher-Education Institutions Using Data Mining: A Review", *Ecos de la Academia*, Vol. 7, pp 185-201, December, 2016.

[7] Thakar, P., "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", arXiv Preprint arXiv:1509.05176, 2015.

[8] Ramaswami, M., and Bhaskaran, R., "A Study on Feature Selection Techniques in Educational Data Mining", arXiv Preprint arXiv:0912.3924, 2009.

[9] Velmurugan, T., and Anuradha, C., "Performance Evaluation of Feature Selection Algorithms in Educational Data Mining", *Performance Evaluation*, Vol. 5, pp. 2, 2016.

[10] Abid, A., Kallel, I., Blanco, I.J., and Benayed, M., "Selecting Relevant Educational Attributes for Predicting Students' Academic Performance", *Proceedings of the International Conference on Intelligent Systems Design and Applications*, pp. 650-660, Springer, 2017.

[11] Pagnotta, F., and Amran, H., "Using Data Mining to Predict Secondary School Student Alcohol Consumption", Department of Computer Science, Vol. 8, pp. 8, University of Camerino, 2016.

[12] Cortez, P., and Silva, A.G.M., "Using Data Mining to Predict Secondary School Student Performance", *Proceedings of 5th Annual Future Business Technology Conference*, pp 5-12, 2008.

[13] Amrieh, E.A., Hamtini, T., and Aljarah, I., "Mining Educational Data to Predict Student's Academic Performance Using Ensemble Methods", *International Journal of Database Theory and Application*, Vol. 9, No. 8, pp. 119-136, 2016.

[14] Hussain, S., Dahan, N.A., Ba-Alwi, F.A., and Ribata, N., "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 9, pp . 2, 2018.

[15] Adil, S.H., Raza, K., and Hashmani, M.A., "A Hybrid Cuckoo Algorithm for loT Scheduling Problem Using Extended Basic Period and Power of Two Policy", *Mehran University Research Journal of Engineering and Technology*, Vol. 35, No. 2, pp. 229, April, 2016.

[16] Liu, J., Lin, Y., Lin, M., Wu, S., and Zhang, J., "Feature Selection Based on Quality of Information", *Neurocomputing*, Vol. 225, pp. 11-22, 2017.

[17] Hsu, H.-H., Hsieh, C.-W., and Lu, M.-D., "Hybrid Feature Selection by Combining Filters and Wrappers", *Expert Systems with Applications*, Vol. 38, No. 7, pp. 8144-8150, 2011.

[18] Gnana, D.A.A., Balamurugan, S.A.A., and Leavline, E.J., "Literature Review on Feature Selection Methods for High-Dimensional Data", *International Journal of Computer Applications*, Vol 136, No. 1, 2016.

[19] Yu, L., and Liu, H., "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", *Proceedings of 20th International Conference on Machine Learning*, pp. 856-863, 2003.

[20] Senliol, B., Gulgezen, G., Yu, L., and Cataltepe, Z., "Fast Correlation Based Filter (FCBF) with a Different Search Strategy", *Proceedings of the 23rd IEEE International Symposium on Computer and Information Sciences*, pp. 1-4, 2008.

[21] Gopalakrishnan, A., Kased, R., Yang, H., Love, M.B., Graterol, C., and Shada, A., "A Multifaceted Data Mining Approach to Understanding What Factors Lead College Students to Persist and Graduate", *Proceedings of the IEEE Conference on Computing*, pp. 372-381, 2017.

[22] Mavrikis, M., "Modelling Student Interactions in Intelligent Learning

Environments: Constructing Bayesian Networks from Data", *International Journal on Artificial Intelligence Tools*, Vol. 19, No. 06, pp. 733-753, 2010.

[23] Dekker, G.W., Pechenizkiy, M., and Vleeshouwers, J.M., "Predicting Students Drop Out: A Case Study", *International Working Group on Educational Data Mining*, 2009.

[24] Bhardwaj, B.K., and Pal, S., "Data Mining: A Prediction for Performance Improvement Using Classification", arXiv Preprint arXiv:1201.3418, 2012.

[25] Algur, S.P. Bhat, P., and Ayachit, N.H., "Educational Data Mining: RT and RF Classification Models for Higher Education Professional Courses", *International Journal of Information Engineering and Electronic Business*, Vol. 8, No. 2, p. 59, 2016.

[26] Fradkin, D., and Muchnik, I., "Support Vector Machines for Classification", *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Vol. 70, pp. 13-20, 2006.

[27] Cheng, C.-H., and Liu, W.-X., "An Appraisal Model Based on a Synthetic Feature Selection Approach for Students' Academic Achievement", *Symmetry*, Vol. 9, No. 11, pp. 282, 2017.

[28] Zaffar, M., Hashmani, M.A., and Savita, K., "Comparing the Performance of FCBF, Chi-Square and Relief-F Filter Feature Selection Algorithms in Educational Data Mining", *Proceedings of the International Conference of Reliable Information and Communication Technology*, pp. 151-160, Springer, 2018.