Computers, Materials & Continua CMC, vol.65, no.3, pp.2413-2423, 2020

MOOC Learner's Final Grade Prediction Based on an Improved Random Forests Method

Yuqing Yang[1, 3], Peng Fu[2, *], Xiaojiang Yang[1, 4], Hong Hong[5] and Dequn Zhou[1]

Abstract: Massive Open Online Course (MOOC) has become a popular way of online learning used across the world by millions of people. Meanwhile, a vast amount of information has been collected from the MOOC learners and institutions. Based on the educational data, a lot of researches have been investigated for the prediction of the MOOC learner's final grade. However, there are still two problems in this research field. The first problem is how to select the most proper features to improve the prediction accuracy, and the second problem is how to use or modify the data mining algorithms for a better analysis of the MOOC data. In order to solve these two problems, an improved random forests method is proposed in this paper. First, a hybrid indicator is defined to measure the importance of the features, and a rule is further established for the feature selection; then, a Clustering-Synthetic Minority Over-sampling Technique (SMOTE) is embedded into the traditional random forests algorithm to solve the class imbalance problem. In experiment part, we verify the performance of the proposed method by using the Canvas Network Person-Course (CNPC) dataset. Furthermore, four well-known prediction methods have been applied for comparison, where the superiority of our method has been proved.

Keywords: Random forests, grade prediction, feature selection, class imbalance.

1 Introduction

With the development of the innovative cloud computing technologies, Massive Open Online Course (MOOC) has explored in popularity over the past decade. MOOC allows anyone in the world to acquire knowledge by accessing course resources or watching

[1] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China.

[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China.

[3] Office of International Cooperation and Exchanges, Nanjing University of Finance & Economics, Nanjing, 210046, China.

[4] Jiangsu Guidgine Educational Evaluation Inc., Nanjing, 210046, China.

[5] Department of Electrical and Computer Engineering, University of California, Davis, CA 95616, USA.

[*] Corresponding Author: Peng Fu. Email: fupeng@njust.edu.cn.

Received: 03 June 2020; Accepted: 07 July 2020.

videos on the internet. Based on the vast amount of information collected from the learners and institutions, a growing number of investigators began to carry out relative researches on MOOC analytics [Moreno-Marcos, Alario-Hoyos, Muñoz-Merino et al. (2018); Pérez-Lemonche, Martínez-Muñoz and Pulido-Cañabate (2017); Jiang, Williams, Schenke et al. (2014)]. Among these studies, the predicting of MOOC learner's grade is popular and of great significance. Based on the grade predictions, teachers can improve their teaching methods or optimize the course schedule; meanwhile the learners can reflect on their learning process and improve their performance. In previous studies, different kinds of classification and prediction methods have been investigated [Zhang, Sun, Zheng et al. (2019); Zheng, Liu and Hou (2017); Zheng, Wang, Zhang et al. (2019)]. In literature Yang et al. [Yang, Brinton, Joe-Wong et al. (2017)], applied the time series neural networks (TSNN) to predict the MOOC learners' grades based on their behaviors. Ren et al. [Ren, Rangwala and Johri (2016)] proposed a multi-regression model (MRM) for the prediction of MOOC learner's performance. The MRM is a real-time model, which can track the participation of a student. Lopez et al. [Lopez, Luna, Romero et al. (2012)] firstly exploited the statistics and social network information with a Moodle module, and then final labels are predicted by using a clustering-based classification method. Gadhavi et al. [Gadhavi and Patel (2017)] attempted to establish a linear regression approach to predict the final scores of students. In Pérez-Lemonche et al. [Pérez-Lemonche, Martínez-Muñoz and Pulido-Cañabate (2017)], the authors applied the random forests and neural networks to predict the MOOC learner's grade. Chen et al. [Chen, Feng, Sun et al. (2019)] proposed an algorithm to predict the student's final score by taking full advantages of the decision tree and extreme learning machine. In the first step, the decision tree algorithm is adopted to choose features, and then the extreme learning machine helps to optimize the prediction accuracy. In addition to the researches mentioned above, a number of methods have been investigated for the prediction of the MOOC learner's grade, where the survey can be found in the literatures [Meier, Xu, Atan et al. (2015); Al-Shabandar, Hussain, Laws et al. (2017); Dalipi, Imran and Kastrati (2018)].

Although many researchers have developed kinds of methods for the grade prediction, there are still two main challenges in this field: (1) How to select the most proper features to achieve better prediction results? (2) How to use or even improve the data mining methods for a better MOOC dataset analysis? In order to address these two challenges, we propose an improved random forests (IRF) method for the MOOC learner's final grade prediction. First, a hybrid indicator is proposed to measure the importance of features, and then a standard rule is then established for the feature selection; second, we improve the traditional random forests algorithm by embedding a Clustering-Synthetic Minority Over-sampling Technique (SMOTE) to solve the class imbalance problem. In experiment part, performance of the proposed IRF method has been verified by using the well-known Canvas Network Person-Course (CNPC) dataset.

2 Methodology

2.1 Feature selection with a hybrid indicator and a decision rule

When dealing with a prediction task, we may have a number of features for the classifier. However, it does not mean that more features will lead to higher accuracy. In practice, we

need to analyze the characteristics of features and then select the most useful features for the prediction. To solve this problem, we develop an indicator to measure the importance of the features. Furthermore, we establish a simple rule for the feature selection based on the proposed indicator. The developed indicator is composed of the measure Variance and Pearson Correlation Coefficient (PCC). The Variance can indicate the spread of the data in a feature, which can be written as:

$$\mathrm{V}_X = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \tag{1}$$

where X represents the data vector in a feature, n counts the number of the variables in X, μ is the mean value of these variables and V_X denotes the variance. The measure Variance indicates the inner statistic characteristics of a feature. In practice, the higher value of the variance means that the data in this feature is more diversified, thus the feature is more useful for the prediction. Another measure is PCC, which calculates correlations between the features as:

$$\mathrm{PCC}_{X,Y} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}} \tag{2}$$

where X and Y are two vectors, and n counts the total number of elements in the vector. Before the prediction, we firstly compute PCC values of the predicted feature versus the remaining features to find the relationship between them. In practice, we always use the absolute value of PCC to measure the correlations, where a higher value indicates a closer relationship. By combining the measures Variance and PCC, the developed indicator V-PCC can be written as:

$$\text{V-PCC}=\mathrm{V}\times\mathrm{abs}\left(\mathrm{PCC}\right) \tag{3}$$

where abs means calculating the absolute value. It is worth noting that to make the measures Variance and PCC in the same range, each feature should be normalized according to the following equation:

$$\hat{F} = \frac{\left(F - \min\left(F\right)\right)}{\left(\max\left(F\right) - \min\left(F\right)\right)} \tag{4}$$

where $\hat{F}$ is the normalized feature, $\min\left(F\right)$ denotes the minimum value and $\max\left(F\right)$ denotes the maximum value in F.

When we obtain the V-PCC value of each feature, the mean and median values are calculated to find the threshold for the feature selection. The rule is to set the minimum value of the mean and median as the threshold, and then the useful features are selected if their V-PCC values are no smaller than the threshold. For a better illustration, two examples are exhibited in Fig. 1. In Fig. 1(a), the features F_1 to F_6 have high V-PCC values, they should be all selected as the useful features, thus the mean value should be chosen as the threshold in this case; on the other hand, in Fig. 1(b), the feature F_1 has very high V-PCC value, and F_2 to F_6 have relative high values compared to F_7. As one

single feature is difficult to produce an accurate and stable prediction, it is better to select the features F_1 to F_7, thus the median value should be set as the threshold in this case.

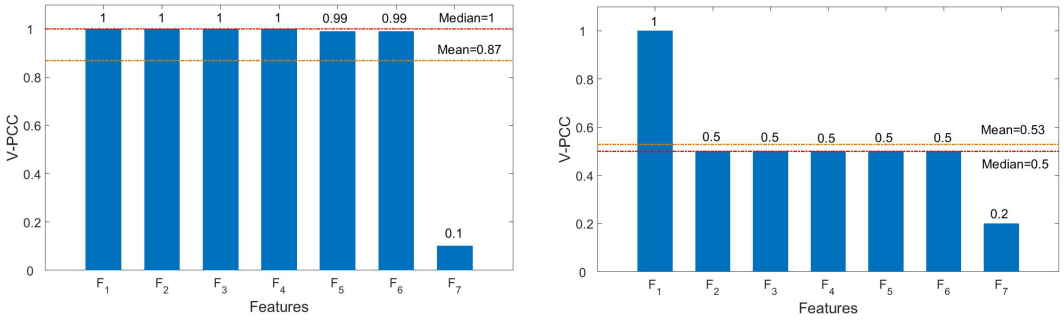


Figure 1: Two examples of the established feature selection rule

2.2 The improved random forests with Clustering-SMOTE technique

The traditional random forests algorithm always suffers from the class imbalance problem. To solve this problem, the SMOTE algorithm is usually applied to generate artificial samples and avoid the risk of over-fitting [Gong and Gu (2016)]. However, in the SMOTE algorithm, artificial samples may change the distribution of the within-class data, the reason is that it just randomly chooses a minority instance to oversample with uniform probability. To overcome this shortage, a Clustering-SMOTE technique is developed in this paper. Data clustering is a classical problem in the field of pattern recognition, and a lot of clustering approaches have been investigated in the past, such as the K-means [Arora and Varshney (2016)], Fuzzy-K-means [Blömer, Brauer and Bujna (2016)], DBSCAN (Density-Based Spatial Clustering of Application with Noise) [Schubert, Sander, Ester et al. (2017)], CFSFDP (Clustering by Fast Search and Find of Density Peaks) [Cheng, Yang and Kong (2018)] and so on. In the Clustering-SMOTE, the classes with less samples are first processed by using a clustering approach, then the cluster centers are utilized to generate the artificial samples, which can be formulated as:

$$X_{\mathrm{AS}} = \Theta_{\tau} + \mathrm{rand}(0,1) \times (X_{\mathrm{ORI}} - \Theta_{\tau}) \tag{5}$$

where X_{AS} and X_{ORI} are the artificial samples and the original samples, respectively. Θ_{τ} represents the τ th clustering center, and $\mathrm{rand}(0,1)$ means a random number between 0 and 1. In this case, the data samples can be balanced between-class and within-class. Once the data are processed by the Clustering-SMOTE technique, the random forests algorithm is applied for the subsequent prediction. Combining the feature selection and Clustering-SMOTE technique, the steps of the proposed IRF method can be summarized as follows:

Improved random forests (IRF) method

Step1: Feature selection
- Compute the Variance and PCC of the samples in each feature;
- Calculate the mean and median values of the V-PCC indicator, and then choose useful features according to the feature selection rule.

Step2: Imbalanced data processing
- Cluster the data in classes with less samples to formulate several class centers;
- Generate artificial samples according to Eq. (5).

Step3: Random forests
- Generate N tree bootstrap samples from the dataset;
- Grow a tree for each bootstrap data set, where the data are trained separately;
- Combining information from the N trees for the prediction based on voting.

3 Experiments

3.1 Experimental dataset introduction

In our experiment, the commonly used CNPC dataset is exploited for all our experiments. The CNPC dataset collects more than 32500 records to save the different features of learners' information and learning activity. In this study, we mainly use the features to predict their final grades, thus eight features are chosen in our experiment, including "Comp", "Nev", "Ndays, "Nfor", "Edu", "Age", "Exp" and "Ncont", where the detailed description of these features can be found in Yang et al. [Yang, Zhou and Yang (2019)]. It is worth noting that a large proportion of the records in CNPC are incomplete. Thus, only 1050 records are utilized for the analysis. Regarding to the prediction feature "grade", it has discrete values from 0 to 1. In order to facilitate the analysis and prediction, we group the grades into five classes according to the Eq. (6):

$$\text{grade}=\begin{cases}\text{C1} & 0\le \text{grade}<0.2\\ \text{C2} & 0.2\le \text{grade}<0.4\\ \text{C3} & 0.4\le \text{grade}<0.6\\ \text{C4} & 0.6\le \text{grade}<0.8\\ \text{C5} & 0.8\le \text{grade}\le 1\end{cases} \tag{6}$$

where C1 to C5 denote the five classes. For a better visualization, a pie chart is shown in Fig. 2. From Fig. 2 we can see that the class C1 has the most samples, while the class C4 owns the least samples. This is because in MOOC learning, a lot of learners may drop out and thus get a zero score. In order to quantitatively analyze the balance between classes, an imbalance ratio α is defined as:

$$\alpha=\frac{\text{majority}(C)}{\text{minority}(C)} \tag{7}$$

in which $\text{majority}(C)$ and $\text{minority}(C)$ represent the number of the samples in classes with the most and fewest samples, respectively. According to Eq. (7), the imbalance ratio equals to 19.9 of the experimental dataset, which further demonstrating the serious class imbalance within the adopted dataset.

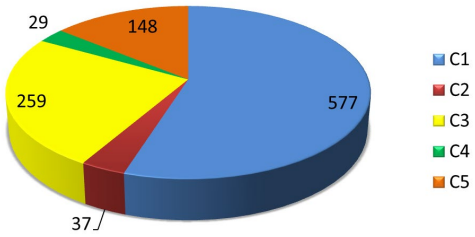


Figure 2: The number of samples in each class with the experimental dataset

3.2 Parameter setting

3.2.1 Set the number of features for prediction

In our experiment, we have eight features for the prediction. In practice, some features have close correlations to the final grade, and they are helpful for the grade prediction. On the other hand, some features will not increase the prediction accuracy or even make the results worse. By using Eqs. (3) and (4), we calculate the Variance of each feature and PCC of the grade and other features. The values of final indictor V-PCC for the eight features are shown in Fig. 3. From this chart we can find that the feature "Ncont" has the highest V-PCC value, while the feature "Nev" has the lowest. We compute the mean and median value of the indicator V-PCC, they equal to 0.0161 and 0.0013, respectively. Based on the feature selection rule, the median value is set as the threshold. Accordingly, four important features are selected for the grade prediction in our experiment, including "Ncont", "Age", "Comp", and "Edu".

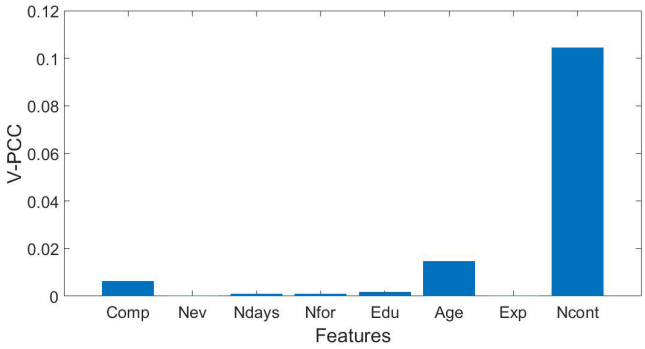
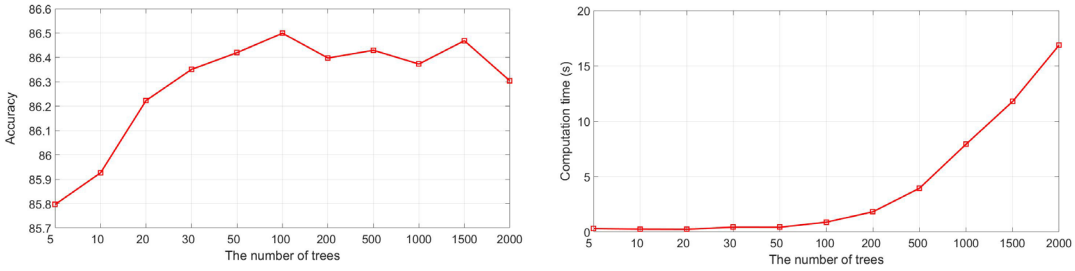


Figure 3: V-PCC value of each feature in the experimental dataset

3.2.2 Set the number of trees in IRF

In the proposed IRF method, the number of trees may affect the final prediction accuracy. In this experiment, we set the number of trees from 5 to 2000 for analysis, where the prediction performances with ten repetitions of 10-fold cross-validation are shown in Fig. 4. Note that in our experiment, we simply apply the K-means clustering algorithm in the clustering-SMOTE step. Fig. 4(a) shows the prediction accuracy with different number of trees. From this curve we can see that with the increasing of the number of trees, the prediction accuracy increases. But it is worth noting that, increasing the number of trees

would bring no significant performance gain when the number of trees is more than 100. Fig. 4(b) records the computational costs of the IRF method. It is clearly that with the increase of the number of trees, the computational cost increases greatly. By comprehensively considering the accuracy and computational cost, the number of trees is set as 100 in our experiment.



(a) (b)

Figure 4: Prediction performance with different number of trees. (a) Prediction accuracy, (b) computation time

3.3 Experimental results

To evaluate the performance of the proposed IRF, four widely used prediction methods are compared, including K Nearest Neighbor (KNN) [Zheng, Liu and Hou (2017)], Discriminant Analysis Classifier (DAC) [Al-Shabandar, Hussain, Laws et al. (2017)], Decision Tree (DT) [Albán and Mauricio (2018)], and Random Forests (RF) [Hardman, Paucar-Caceres and Fielding (2013)]. In this experiment, we randomly select 60% samples of the dataset for training process, and the remaining 40% for testing. To assess performance, we use ten repetitions of 10-fold cross-validation. The four compared methods are implemented first, where the prediction accuracy values are recoded in Tab. 1. In addition, the confusion matrix has been used for a clearer observation of the prediction performance. Fig. 5 shows the confusion matrixes of predictions with different methods, where the x-axis and y-axis record the predicted and true class values, respectively. From Tab. 1 and Figs. 5(a)-5(d) we can see that the RF method obtains the highest accuracy, while the DAC method has the worst performance. For a more fair comparison, we also compared these methods with feature selected, i.e., only four most important features are adopted for the prediction. Tab. 2 shows the prediction accuracy of IRF and the compared methods with selected features, and Figs. 5(e)-(i) exhibit the corresponding confusion matrixes. By contrasting the results recorded in Tabs. 1 and 2, we can find that all the four compared methods produce better performances with the selected features, which also prove the effectiveness of the feature selection approach. For a better visualization, a bar chart (Fig. 6) has been designed to record the prediction accuracy values of all the methods. From Tabs. 1-2 and Figs. 5-6, we can have the conclusion that compared to the four methods with the all features and selected features, the proposed IRF method produces the highest class prediction accuracy.

Table 1: Prediction accuracy with the four compared methods

Method	KNN	DAC	DT	RF
Accuracy (%)	71.90	60.71	79.29	82.86

Table 2: Prediction accuracy with IRF and four compared methods with feature selected

Method	KNN-SF	DAC-SF	DT-SF	RF-SF	IRF
Accuracy (%)	75.24	76.67	82.14	84.52	86.67

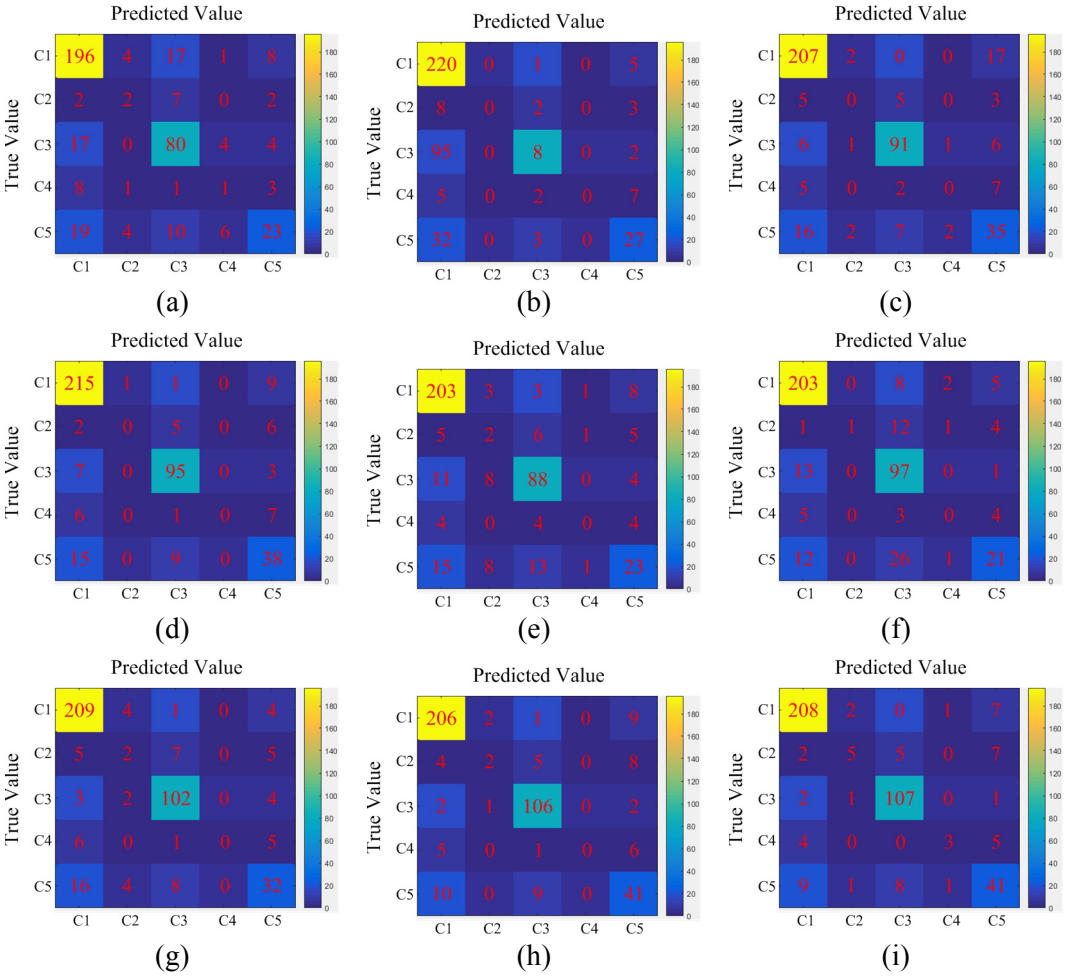


Figure 5: Confusion matrixes of different prediction methods. (a) KNN, (b) DAC, (c) DT, (d) RF, (e) KNN-SF, (f) DAC-SF, (g) DT-SF, (h) RF-SF, (i) IRF

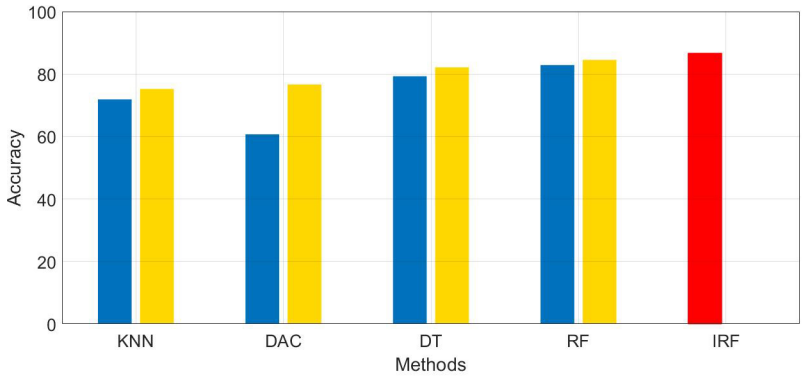


Figure 6: Prediction accuracy with the compared methods and the proposed IRF

4 Conclusion

This paper proposes an improved random forests method for the prediction of MOOC learner's final grade. First, an indicator is defined to measure the importance of the features, and a rule is established to select the most proper features for the grade prediction; then, a clustering-SMOTE technique is developed and embedded into the traditional random forests algorithm to improve the prediction accuracy with class imbalance data. Experimental results with CNPC MOOC dataset prove the effectiveness of the proposed method. Finally, a comparison of IRF and another four widely used prediction algorithms proves the superiority of the proposed method.

Funding Statement: This work was in part supported by the National Natural Science Foundation of China under Grant No. 61801222, and in part supported by the Fundamental Research Funds for the Central Universities under Grant No. 30919011230, and in part supported by the Jiangsu Provincial Department of Education Degree and Graduate Education Research Fund under Grant No. JGZD18_012. In addition, the authors would like to thank Prof. Feng Chen and Jianwei Zhou from Nanjing Medical University for the invaluable discussions about the online course study in high school.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Albán, M.; Mauricio, D. (2018): Decision trees for the early identification of university students at risk of desertion. *International Journal of Engineering & Technology*, vol. 7, no. 44, pp. 51-54.

Al-Shabandar, R.; Hussain, A.; Laws, A.; Keight, R.; Lunn, J. et al. (2017): Machine learning approaches to predict learning outcomes in massive open online courses. *Proceedings of the International Joint Conference on Neural Networks*, pp. 713-720.

Arora, P.; Varshney, S. (2016): Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, vol. 78, pp. 507-512.

Blömer, J.; Brauer, S.; Bujna, K. (2016): A theoretical analysis of the fuzzy k-means problem. *Proceedings of the 16th International Conference on Data Mining*, pp. 805-810.

Chen, J.; Feng, J.; Sun, X.; Wu, N.; Yang, Z. et al. (2019): MOOC dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Mathematical Problems in Engineering*, vol. 2019, pp. 1-12.

Cheng, C.; Yang, J.; Kong, X. (2018): Research on sampling method of CFSFDP clustering algorithm and its criteria for determining the best sample size. *Proceedings of the 2nd International Conference on Advances in Artificial Intelligence*, pp. 24-28.

Dalipi, F.; Imran, A. S.; Kastrati, Z. (2018): MOOC dropout prediction using machine learning techniques: Review and research challenges. *Proceedings of the IEEE Global Engineering Education Conference*, pp. 1007-1014.

Gadhavi, M.; Patel, C. (2017): Student final grade prediction based on linear regression. *Indian Journal of Computer Science and Engineering*, vol. 8, no. 3, pp. 274-279.

Gong, C.; Gu, L. (2016): A novel SMOTE-based classification approach to online data imbalance problem. Mathematical Problems in Engineering, vol. 2016, pp. 1-14.

Hardman, J.; Paucar-Caceres, A.; Fielding, A. (2013): Predicting students' progression in higher education by using the random forest algorithm. *Systems Research and Behavioral Science*, vol. 30, no. 2, pp. 194-203.

Jiang, S.; Williams, A.; Schenke, K.; Warschauer, M.; O'Dowd, D. (2014): Predicting MOOC performance with week 1 behavior. *Proceedings of the 7th International Conference on Educational Data Mining*, pp. 273-275.

Lopez, M. I.; Luna, J. M.; Romero, C.; Ventura, S. (2012): Classification via clustering for predicting final marks based on student participation in forums. *Proceedings of the 5th International Conference on Educational Data Mining*, pp. 148-151.

Meier, Y.; Xu, J.; Atan, O.; Van der Schaar, M. (2015): Predicting grades. *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 959-972.

Moreno-Marcos, P. M.; Alario-Hoyos, C.; Muñoz-Merino, P. J.; Kloos, C. D. (2018): Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies*, vol. 12, no. 3, pp. 384-401.

Pérez-Lemonche, Á.; Martínez-Muñoz, G.; Pulido-Cañabate, E. (2017): Analysing event transitions to discover student roles and predict grades in MOOCs. *Proceedings of the International Conference on Artificial Neural Networks*, pp. 224-232.

Ren, Z.; Rangwala, H.; Johri, A. (2016): Predicting performance on MOOC assessments using multi-regression models. *Proceedings of the 9th International Conference on Educational Data Mining*, pp. 484-489.

Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; Xu, X. (2017): DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1-21.

Yang, T. Y.; Brinton, C. G.; Joe-Wong, C.; Chiang, M. (2017): Behavior-based grade prediction for MOOCs via time series neural networks. *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 716-728.

Yang, Y.; Zhou, D.; Yang, X. (2019): A multi-feature weighting based K-means

algorithm for MOOC learner classification. *Computers, Materials & Continua*, vol. 59, no. 2, pp. 625-633.

Zhang, G.; Sun, H.; Zheng, Y.; Xia, G.; Feng, L. et al. (2019): Optimal discriminative projection for sparse representation-based classification via bilevel optimization. *IEEE Transactions on Circuits and Systems for Video Technology* (In Press).

Zheng, Y.; Liu, R.; Hou, J. (2017): The construction of high educational knowledge graph based on MOOC. *Proceedings of the 2nd Information Technology, Networking, Electronic and Automation Control Conference*, pp. 260-263.

Zheng, Y.; Wang, X.; Zhang, G.; Xiao, B.; Xiao, F. et al. (2019): Multi-kernel coupled projections for domain adaptive dictionary learning. *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2292-2304.