Science Journal of Education
2021; 9(2): 58-62
http://www.sciencepublishinggroup.com/j/sjedu
doi: 10.11648/j.sjedu.20210902.16
ISSN: 2329-0900 (Print); ISSN: 2329-0897 (Online)

Education Data Mining Application for Predicting Students' Achievements of Portuguese Using Ensemble Model

Shuai Zhang, Jie Chen, Wenyu Zhang*, Qiwei Xu, Jiaxuan Shi

School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou, China

Email address:

zhangshuai@zufe.edu.cn (Shuai Zhang), chenjzufe103@zufe.edu.cn (Jie Chen), wyzhang@e.ntu.edu.sg (Wenyu Zhang), xuqiwei@zufe.edu.cn (Qiwei Xu), shijiaxuan@zufe.edu.cn (Jiaxuan Shi)

*Corresponding author

To cite this article:

Shuai Zhang, Jie Chen, Wenyu Zhang, Qiwei Xu, Jiaxuan Shi. Education Data Mining Application for Predicting Students' Achievements of Portuguese Using Ensemble Model. *Science Journal of Education.* Vol. 9, No. 2, 2021, pp. 58-62. doi: 10.11648/j.sjedu.20210902.16

Received: March 13, 2021; **Accepted**: April 15, 2021; **Published**: April 26, 2021

Abstract: With the emergence of the massive educational data, education data mining techniques have extensively drawn considerable interest from scholars to explore the relationship between students' achievements and other factors. In this study, the data set about the students' achievements of Portuguese in two secondary education schools in Portugal is selected for education data mining, which involves the personal information, social and school related factors. To analyze the relationship between the students' achievements and other factors, this study proposed an ensemble model based on weighted voting for predicting the students' achievements of Portuguese in the final period. First, the raw data is preprocessed using some basic methods, including dummy coding, correlation analysis, standardization, and normalization. Second, the isolation forest algorithm-based outlier adaption is applied to deal with the data set to enhance the robustness of the ensemble model. Finally, two base classifiers, i.e. gradient boosting decision tree and extreme gradient boosting, are integrated to form the ensemble model. The experiments are presented for verifying the superiority of the proposed model by comparing with five base classifiers, including gradient boosting decision tree, adaptive boosting, extreme gradient boosting, random forest, and decision tree. The experimental results demonstrate that the ensemble model performs better than other base classifiers in classification, and prove the validity of the outlier adaption based on isolation forest algorithm.

Keywords: Ensemble Model, Education Data Mining, Prediction, Students' Achievements

1. Introduction

With the development of the education, massive useful educational data is emerged. Therefore, how to use these data effectively has been a critical issue. Recently, education data mining (EDM) has drawn a considerable research interest from scholars. EDM is an emerging discipline that extracts and analyzes the hidden knowledge from the educational data by using data mining techniques [1]. Moreover, EDM focuses on using data mining techniques to explore the educational data to learn more about the students and educational environment [2].

Generally, data mining techniques include classification, clustering, association rule mining, prediction, and so on. In this study, classification is used to analyze students' achievements of Portuguese in the final period from two secondary education schools in Portugal. First, the ensemble model based on weighted voting is proposed, which integrates two base classifiers, i.e. gradient boosting decision tree (GBDT) [3] and extreme gradient boosting (XGBoost) [4]. The raw data is preprocessed with some approaches, including dummy coding, correlation analysis, standardization, and normalization. Moreover, the isolation forest algorithm (IF)-based outlier adaption is applied in the ensemble model to improve the classification performance. Finally, the experiments are performed by comparing the proposed ensemble model with five base classifiers, including GBDT, adaptive boosting (AdaBoost) [5], XGBoost, random forest (RF) [6], and decision tree (DT) [7]. The experimental results demonstrate that the proposed model outperforms other base classifiers in predicting the students' Portuguese achievements in the final period, and the

validity of the IF-based outlier adaption has also been proved.

The remainder of this paper is organized as follows. Section 2 reviews the related work on EDM and ensemble methods. Section 3 describes the methodology of the proposed ensemble model. Section 4 presents the experiments and analyzes the experimental results by comparing the proposed model with the base classifiers. Section 5 presents the conclusions and directions for future work.

2. Related Work

2.1. EDM

To effectively extract the hidden knowledge in the educational data, many scholars pay attention to the EDM to analyze educational data. There are several EDM techniques that are extensively used in the educational environment, including classification, clustering, association rule mining, and prediction. Zhang et al. [8] utilized the association rule mining to discover the hidden knowledge between the career choices and academic performance. Francis and Babu [9] proposed the hybrid data mining method to predict the students' academic performance based on students' learning behavior, by integrating the classification and clustering techniques. Karthikeyan et al. [10] employed the hybrid model based on the J48 Classifier and Naive Baye's classification to analyze the students' academic performance. Mengash [11] predicted the academic performance of applicants by using data mining techniques to provide the reliable admission information for higher education institutions.

Previous literatures have proved that the EDM techniques can analyze the relationship between students' behaviors and the other factors to solve the educational issues. In this study, the data set about students' achievements of Portuguese in secondary education of two Portuguese schools is selected to analyze the hidden relationship between students' achievements and other factors such as the personal, social and school information. Moreover, the students' achievements of Portuguese in the final period are predicted.

2.2. Ensemble Methods

In the EDM, the ensemble methods that integrate several base classifiers have drawn considerable attention due to its better classification performance over the individual classifiers. Existing literatures have explored the ensemble methods in education issues. Kausar et al. [12] utilized the ensemble methods to analyze the students' learning performance, and demonstrated that the ensemble methods of bagging and stacking classifiers can enhance the classification performance. Sun et al. [13] proposed a multi-classification model, by combining the feature integration and the ensemble method to predict the education grants of students.

In the ensemble model, voting is popularly used to integrate the prediction results of several base classifiers. The voting strategies, which include majority voting [6] and weighted voting [14], have been extensively applied in the classification field. For example, Assi et al. [15] developed the ensemble methods based on majority voting to predict the choice behavior of student travel mode. Troussas et al. [16] utilized the ensemble classification methods through majority voting to identify the learning styles of students. Rao et al. [17] constructed the binary classification model based on weighted voting to assess the credit of the borrowers in rural. Xiao et al. [18] proposed the weighted voting-based ensemble method for identifying and classifying the pulmonary nodules.

Based on the successful application of the ensemble model in educational issues, this study proposed an ensemble model based on weighted voting to predict the students' achievements of Portuguese in the final period. The raw data is preprocessed by dummy coding, correlation analysis, standardization, and normalization, and the IF-based outlier adaption is applied to improve the performance of the proposed model in robustness.

3. Methodology

3.1. Dataset Exploration and Preprocessing

The data set used in this study concerns the students' achievements of Portuguese in two secondary education schools in Portugal. The data source is http://archive.ics.uci.edu/ml/datasets/Student+Performance. This data set includes 33 attributes and 649 samples, and the detailed description is illustrated in Cortez and Silva [19]. In Portugal, students in secondary education are evaluated through three periods, and the last period is the final grade. The purpose of this study is to predict the students' achievements of Portuguese in the final period through their personal information, social and school related factors.

In this study, the numerical value of the final grade to be predicted is converted into binary value. According to Cortez and Silva [19], if the final grade is equal or greater than 10, the student will pass; else, the student will fail.

In this study, the raw data is preprocessed with some basic methods to improve the classification performance, including dummy coding, correlation analysis, standardization, and normalization. The input variables are converted into dummy variables by dummy coding. The correlation between two explanatory features is evaluated by correlation analysis. If the correlation of any two explanatory features exceeds 0.97, one of the features is removed. Then, the numerical data is scaled into a specific range through standardization and normalization.

3.2. The Proposed Model

This section presents the ensemble model, which is shown in Figure 1. The detail of the model is as follows.

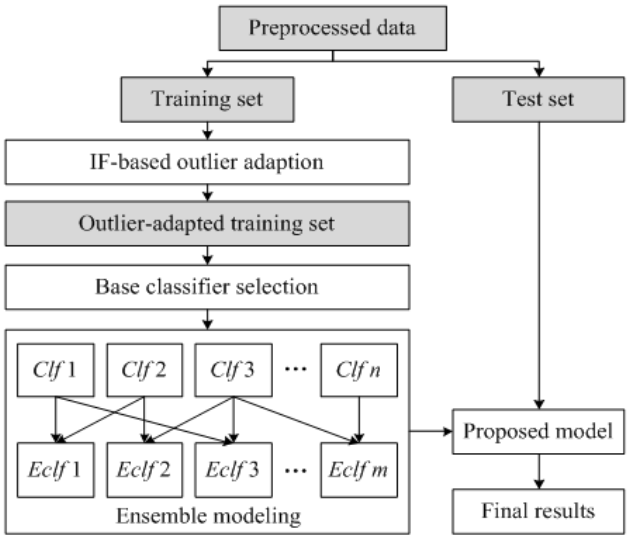

Figure 1. *The framework of the proposed model.*

3.2.1. IF-based Outlier Adaption

IF [20] is an unsupervised outlier detection method to identify outliers. It uses multiple binary trees to isolate the data points, and obtain the outlier score by measuring the average path length of the data point through multiple sampling. Wei et al. [21] indicated that the outliers detected by IF can reflect the real objective function, and boosting detected outliers to the training set can decrease the possibility of model overfitting. Therefore, IF is used to identify the outliers in the preprocessed data effectively in this study, and the IF-based outlier adaption [21] is employed to boost the outliers in the training set so as to improve the performance of the proposed model in robustness.

3.2.2. Ensemble Model

The ensemble model that integrates several base classifiers based on majority voting, weighted voting, and other voting strategies can improve the accuracy and robustness by contrast to the individual classifiers [22]. This study proposed an ensemble model based on weighted voting to obtain the better classification ability. The five popular base classifiers, including GBDT, AdaBoost, XGBoost, RF, and DT, are evaluated. Through evaluation, the n (<= 5) base classifiers (*Clf* 1, *Clf* 2, …, *Clf n*) with better performance are selected to construct m candidate ensemble models (*Eclf* 1, *Eclf* 2, …, *Eclf m*). By comparing the performance of the candidate ensemble models, this study selects the ensemble model with the best performance as the proposed model.

4. Experiment

In this section, the evaluation indicators for measuring the classification performance of the classifiers are introduced, and the experimental results of the classifiers in predicting students' achievements of Portuguese are analyzed. The data set was divided into the training set and test set with the proportion of 4:1. The experiments were executed in 5 replications, and the classification performance was measured by the average value. The experiments were implemented using Python programming language on a personal computer with a 3.2 GHz AMD Ryzen 7 2700 Eight-Core processor with 16 GB of RAM.

4.1. Evaluation Indicators

The confusion matrix has been extensively used to evaluate the classification performance. The confusion matrix [23] includes four elements: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). This study utilizes four evaluation indicators to assess the classification ability of the classifiers, including accuracy, AUC, F-score, and Brier score. The description of evaluation indicators is as follows.

Accuracy denotes the proportion of the samples that are correctly classified to the total samples, and is calculated by Equation (1).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

AUC is a typically used evaluation indicator in the binary classification, and is calculated as the area under the receiver operating characteristic curve [23]. The value of AUC ranges from 0 to 1, and the higher AUC value of the classifier represents the better classification performance.

F-score is defined as the harmonic mean of precision and recall. F-score is calculated by Equation (2). Precision and recall are calculated by Equations (3) and (4).

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Brier score [24] denotes the mean squared difference between the predicted probability and the actual value, and is calculated by Equation (5). The lower Brier score represents that the corresponding classifier has a better performance.

$$Brier\ score = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{5}$$

where N denotes the total number of the samples; p_i and y_i indicate the predicted probability and the actual value of the sample i, respectively.

4.2. Experimental Results Analysis

In this study, five base classifiers without IF-based outlier adaption were selected as baselines of comparison, which is shown in the Table 1. The average rank [21] is a ranking result with comprehensive consideration of the four evaluation indicators. The value is marked in bold if the average rank or evaluation indicator of the corresponding base classifier is better or same as the other base classifiers. From Table 1, it can be observed that the average rank of

GBDT is same as that of RF, but GBDT performs better than RF in AUC and Brier score. In terms of AUC, the value of DT is obviously lower than that of other base classifiers. Thus, the base classifiers except DT were selected and composed to construct six candidate ensemble models (*Eclf* 1, *Eclf* 2, *Eclf* 3, *Eclf* 4, *Eclf* 5, and *Eclf* 6) through weighted voting, which is illustrated in the Table 2.

Table 1. *The classification performance of five base classifiers without IF-based outlier adaption.*

Base classifier	Average rank	Accuracy	AUC	F-score	Brier score
GBDT	1.50	0.89846	0.95577	0.82236	0.07053
RF	1.50	0.90308	0.95481	0.82788	0.07891
XGBoost	3.25	0.89231	0.94541	0.81457	0.08354
AdaBoost	4.00	0.89077	0.95347	0.81338	0.21096
DT	4.75	0.87692	0.85696	0.80311	0.12308

Table 2. *The composition of six candidate ensemble models.*

Ensemble model	GBDT	AdaBoost	RF	XGBoost
Eclf 1	√	√	-	-
Eclf 2	√	-	√	-
Eclf 3	√	-	-	√
Eclf 4	-	√	√	-
Eclf 5	-	√	-	√
Eclf 6	-	-	√	√

Then, the classification performance of six candidate ensemble models without IF-based outlier adaption was compared in the Table 3, and each ensemble model was ranked based on the evaluation indicators. The value is marked in bold if the average rank or evaluation indicator of the corresponding ensemble model is better or same as the other ensemble models. It can be observed that *Eclf* 3 outperforms the other ensemble models; thus, this study selects *Eclf* 3 as the proposed ensemble model, which integrates the GBDT and XGBoost based on weighted voting.

Moreover, the classification performance of the five base classifiers and proposed ensemble model with the IF-based outlier adaption applied were compared. The average rank and evaluation indicators of each classifier with IF-based outlier adaption are shown in the Table 4. The value of the average rank is marked in bold if the corresponding classifier has a better or same average rank as others; and the value of evaluation indicator is marked in bold too if the evaluation indicator of the corresponding classifier is better or same as the one without IF-based outlier adaption. Table 4 demonstrates that *Eclf* 3 with IF-based outlier adaption has the best performance in predicting the students' achievements on basis of the comprehensive evaluation. By comparing Table 4 with Tables 1 and 3, it can be seen that each classifier with IF-based outlier adaption performs better than the corresponding one without IF-based outlier adaption. In summary, it is proved that the ensemble model has the competitive performance by comparing with five base classifiers, and the validity of the IF-based outlier adaption in improving the robustness of the classifier has also been verified.

Table 3. *The classification performance of six candidate ensemble models without IF-based outlier adaption.*

Ensemble model	Average rank	Accuracy	AUC	F-score	Brier score
Eclf 3	1.75	0.90615	0.95603	0.83778	0.07426
Eclf 4	3.00	0.90154	0.95682	0.82655	0.11853
Eclf 1	3.25	0.89846	0.95793	0.82543	0.10473
Eclf 6	3.50	0.89846	0.95295	0.82568	0.07645
Eclf 2	3.75	0.89692	0.95539	0.81926	0.07172
Eclf 5	5.75	0.89231	0.95250	0.81457	0.10583

Table 4. *The classification performance of the proposed model and base classifiers with IF-based outlier adaption.*

Classifier	Average rank	Accuracy	AUC	F-score	Brier score
Eclf 3	1.50	0.91556	0.96741	0.84943	0.06423
GBDT	2.00	0.90963	0.96844	0.83854	0.06234
XGBoost	3.00	0.90963	0.95866	0.84091	0.07141
AdaBoost	4.00	0.90667	0.96522	0.84012	0.21015
RF	4.75	0.90519	0.95790	0.83084	0.07538
DT	5.75	0.88000	0.85979	0.81068	0.12000

5. Conclusion

EDM has been a popular tool to analyze the educational data with a large size and promote the development of education. This study proposed the ensemble model to predict the students' achievements of Portuguese in the final period, by integrating GBDT and XGBoost based on

weighted voting. The raw data is preprocessed by some basic methods, including dummy coding, correlation analysis, standardization, and normalization. Moreover, the IF-based outlier adaption is adopted to enhance the performance of the classifiers in robustness. By comparing the performance of the proposed ensemble model and five base classifiers, it is demonstrated that the proposed ensemble model outperforms the base classifiers. In addition, it is proved that the IF-based outlier adaption can effectively enhance the classification performance.

However, this study still has some limitations, and can be extended in several directions. For example, other outlier detection methods can be considered and compared to achieve a better classification performance. More evaluation indicators can also be used to comprehensively evaluate the classifiers.

Acknowledgements

The work has been supported by Zhejiang Higher Education Teaching Reform Research Project of China (No. JG20190294).

References

[1] Romero, C., and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33 (1), 135–146.

[2] Baker, R. S. J. D. (2010). Data mining for education. *International Encyclopedia of Education*, 7 (3), 112–118.

[3] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29 (5), 1189–1232.

[4] Chen, T. Q., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, pp. 785–794, August 13–17.

[5] Freund, Y., and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *In Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy, pp. 148–156, July 3–6.

[6] Breiman, L. (2001). Random forests. *Machine Learning*, 45 (1), 5–32.

[7] Li, X., Ying, W., Tuo, J., Li, B., and Liu, W. (2004). Applications of classification trees to consumer credit scoring methods in commercial banks. *In Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, Hague, Netherlands, pp. 4112–4117, October 10–13.

[8] Zhang, L. B., Tan, X. W., Zhang, S., and Zhang, W. Y. (2019). Association rule mining for career choices among fresh graduates. *Applied and Computational Mathematics*, 8 (2), 37–43.

[9] Francis, B. K., and Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of Medical Systems*, 43 (6), 1–15.

[10] Karthikeyan, V. G., Thangaraj, P., and Karthik, S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. *Soft Computing*, 24 (24), 18477–18487.

[11] Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, 8, 55462–55470.

[12] Kausar, S., Oyelere, S. S., Salal, Y. K., Hussain, S., Cifci, M. A., Hilcenko, S., Iqbal, M. S., Zhu, W. H., and Xu, H. H. (2020). Mining smart learning analytics data using ensemble classifiers. *International Journal of Emerging Technologies in Learning*, 15 (12), 81–102.

[13] Sun, Y., Li, Z. L., Li, X. W., and Zhang, J. (2021). Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction. *Applied Artificial Intelligence*, 35 (4), 290–303.

[14] Schapire, R. E. (1999). A brief introduction to boosting. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 1401–1406, July 31–August 6.

[15] Assi, K. J., Shafiullah, M., Nahiduzzaman, K. M., and Mansoor, U. (2019). Travel-to-school mode choice modelling employing artificial intelligence techniques: A comparative study. *Sustainability*, 11 (16), 4484.

[16] Troussas, C., Krouska, A., Sgouropoulou, C., and Voyiatzis, I. (2020). Ensemble learning using fuzzy weights to improve learning style identification for adapted instructional routines. *Entropy*, 22 (7), 735.

[17] Rao, C. J., Liu, M., Goh, M., and Wen, J. H. (2020). 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers. *Applied Soft Computing*, 95, 106570.

[18] Xiao, N., Qiang, Y., Bilal Zia, M., Wang, S. H., and Lian, J. H. (2020). Ensemble classification for predicting the malignancy level of pulmonary nodules on chest computed tomography images. *Oncology Letters*, 20 (1), 401–408.

[19] Cortez, P., and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *In Proceedings of the 5th Future Business Technique Conference*, Porto, Portugal, pp. 5–12, April 9–11.

[20] Liu, F. T., Ting, K. M., and Zhou, Z. H. (2008). Isolation forest. *In Proceedings of the 8th IEEE International Conference on Data Mining*, Pisa, Italy, pp. 413–422, December 15–19.

[21] Wei, S., Yang, D. Q., Zhang, W. Y., and Zhang, S. (2019). A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning. *IEEE Access*, 7, 99217–99230.

[22] Lin, W. Y., Hu, Y. H., and Tsai, C. F. (2011). Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42 (4), 421–436.

[23] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8), 861–874.

[24] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1), 1–3.