An Empirical Analysis on the Academic Early Warning for Online Learning

Gong Chen

Binzhou University, Shandong, China

Abstract

Information technology is used to construct virtual online learning environment, avoiding the limitations of time and space, and sharing high-quality learning resources. It is a new form of learning. In the process of human-computer interaction, the behaviors of learning are recorded in real time. A large amount of data is generated, providing a basis for studying online learning behavior. The oulad dataset is the research object and machine learning technology is used to predict the final learning result of students in this paper.

Keywords

Online learning; MOOC; Machine Learning.

1. Introduction

With the development of information technology, online learning has become an emerging form of learning. Online learning spreads information and knowledge through the Internet. People use mobile phones or computers to learn anytime and anywhere, making full use of fragmentation time and avoiding the limitations of time and space. In the new framework, a large number of high-quality educational resources are provided. The learners interact with framework through video, text, testing and forums. Taking the MOOC of Tsinghua University as an example, students can choose courses and lectures as needed by registering the platform, and consolidate their knowledge through the exercises provided by the courses. In the process, big datat is generated through the interaction between the student and the platform, such as watching video, answering questions and test conditions, etc., forming online learning big data. Through the analysis of the data, the teacher obtained the student's learning situation. In recent years, the voice of advocating personalized teaching has been increasing. As a leader of personalized teaching, learning behavior analysis technology is the focus of education researchers. Currently, the open learning behavior data sets are kdd cup 2010, OULAD, kdd cup 2015, and so on. Through the research and analysis of data sets, the use of machine learning algorithms to predict students' learning behaviors plays an important role in guiding students to complete their studies and adjust the resources of learning platform. In literature [1], the teacher's online learning is taken as an example to analyze the current situation of online learning. The learning input index is used as the categorical variable to predict the teacher's academic performance. In literature [2], students' learning behaviors are clustered, and then ten learning behavior patterns are obtained. Machine learning algorithms are an important method to predict the behavior of online learners. LR, SGD, RF and SVM are used respectively to predict changes in MOOC participation indicators in the literature [3]. In the literature [4], the accuracy of the random forest algorithm is higher in predicting the change of MOOC participation index compared the classical machine learning algorithms.

2. Machine Learning Algorithm

2.1 Decision Tree

Decision Tree is one of the classic algorithms of machine learning. As the name suggests, it is a tree structure. In the classification problem, in a node of a tree, according to a certain criterion, a certain feature is selected, and the data set is divided into two or more data subsets. In the next node, classification is continued based on a certain feature until the stop condition is reached. The criterion

for feature selection is that the purity of the split data set is higher than that before splitting, and the features are choosen among the features that can make the difference the most, that is, reduce the uncertainty before and after the split as much as possible. The methods of measuring the difference mainly include information gain (ID3), information gain ratio (C4.5) and Gini index (CART), etc. The information gain, information gain ratio, or Gini index is maximized to determine the splitting characteristics of the current node. If the stop condition is not set, the decision tree will grow completely, which may result in a decrease in prediction ability. Therefore, the minimum number of partition instances, partition threshold, or maximum tree depth is set for a better prediction result.

2.2 Random Forest

In order to solve the problem of decision tree over-fitting, local optimal solution and model instability, a random forest is constructed based on multiple decision trees. Each decision tree is a base classifier. If the random forest is consisted of N decision trees, there are N classification results, each decision tree has a voting right, and the prediction result is determined according to one-vote veto, minority-submissive majority or weighted. In essence, random forest is an integrated learning algorithm. The training samples are selected randomly and regressively during the growth of each tree. In addition, in order to enhance the generalization of the model, the selection of features is also random. An optimal feature is selected from all samples in the decision tree algorithm. However, a subset of features is randomly selected in the random forest algorithm, and then an optimal feature is selected from the subset.

2.3 Naïve Bayes

The classification is done by using a known probabilty in the Naive Bayes algorithm. For a given sample, the class is determined according to the probability. It is a linear classifier that can perform multi-classification tasks and adapt to small sample sizes. However, the features are required to be independent of each other. When the features are related to each other, or when linear problems are encountered, the Naive Bayes classifier has poor performance. Its general expression is as shown in Equation 2.1.

$$P(A|B)=P(B|A)P(A)/P(B) \quad (2.1)$$

2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) belongs to supervised algorithm. Firstly, the distance between samples and different categories is calculated, such as Euclidean distance. Secondly, the category of samples is judged according to the distance. The problem of multi-classification can also be solved by the algorithm. However, it is time consuming and not suitable for unbalanced data sets.

3. Online Learning Behavior Prediction

3.1 Introduction of OULAD

The OULAD data set [5][6] is composed of the seven online course data of the UK Open University in 2013 and 2014. It includes personal information and learning behavior data about 40,000 learners, which are recorded in seven data files. Among them, the student's personal basic information and learning results are recorded in studentInfo.csv. The date of student registration and deregistration is recorded in Student Registration table. The course information table includes course modules and course information. The assessments table contains information about assessments in module-presentations, and the Student Assessments table contains the result of the assessments. The VLE table contains information about the materials available, and the Student VLE table records the interaction between the student and the platform, including the interaction type, date and number of times.

3.2 Prediction of Learning Result

The machine learning algorithm is used to predict the final results of studenting in this paper. The results of learning are divided into four categories: Pass, Fail, Withdraw, and Distinction. Before training the model, The data is divided into training set and test set, which are 80% and 20% of the

total data. To characterize the performance of the model, the indicators used in this paper are accuracy, accuracy, recall, and F1-score. They definitions are as follows:

(1)Accuracy is the ratio of the number of correctly classified samples to the total number of samples.

Accuracy = (TP + TN)/(P+N) (3.1), Where TP is true positives and TN is true negative.

(2)Precision is the proportion of true positive samples in the samples that is predicted to be positive.

P=TP/TP+FP (3.2), Where FP is false positive.

(3)Recall is the probability that the positive example in the sample is correctly predicted.

R=TP/TP+FN (3.3),Where FN is false negative.

(4) The F1 value is the weighted mean of the precision value and the recall rate.

In the case where the parameters are not adjusted and the model default parameters are used, the performance indicators are shown in Table 3.1.

Table 3.1 comparison of algorithm performance

algorithm	Accuracy	Precision	Recall	F1-score
DT	0.96	0.96	0.96	0.96
RF	0.95	0.95	0.95	0.95
Naïve Bayes	0.26	0.47	0.26	0.20
KNN	0.59	0.34	0.59	0.43

4. Conclusion

In this paper, the classical machine learning algorithms Naïve Bayes, KNN, Decision Tree and Random Forest are used to predict the learning results according to the online learning situation of students. In the case of unadjusted optimization model, the advantage of Decision Tree and Random Forest is relatively obvious, but the performance index is still not high. In the subsequent work, by preprocessing the data set and optimizing the model parameters, the purpose of further improving the prediction accuracy will be achieved.

References

[1]Zhang Si Liu Qingtang Lei Shijie Wang Yaru.Study of Learners' Engagement in Online Learning Space——Big Data Analytics of e-Learning Behavior[J].China Educational Technology, 2017(04):24-30+40.

[2]Ferguson R, Clow D.Examining engagement:Analysing learner subpopulations in massive open online courses (MOOCs) [C].ACM, Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, 2015:51-58.

[3]Bote-Lorenzo M L. Predicting the decrease of engagement indicators in a MOOC[C]// International Learning Analytics & Knowledge Conference. 2017.

[4]Al-Shabandar R , Hussain A , Laws A , et al. Machine learning approaches to predict learning outcomes in Massive open online courses[C]// 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017.

[5]Kuzilek, J. et al. Open University Learning Analytics dataset. Sci. Data 4:170171 doi: 10.1038/ sdata.2017.171 (2017).

[6]SHI Wanruo NIU Xiaojie ZHENG Qinhua.Empirical Study on the Influencing Factors of Activity-Centered Online Courses Learning Outcomes: Take OULAD as an Example[J].Journal of Open Learning, 2018,23(06):10-18.