eduDiag: Machine learning visual diagnosis based on student behavior and performance prediction

Tingting Zhang[1], Jiansu Pu[1] ,Yuwei Zhang[1], Yulu Xia[1], Haixing Dai[2] ,Jingwen Zhang[1], Hui Shao[1] , and Shaolun Ruan[1]

[1] University of Electronic Science and Technology of China
[2] University of Georgia, Athens, Georgia, United States

ABSTRACT

eduDiag, a visual analytic system, is presented to analyze student behaviors with academic performance based on smart card data. We used the random forest model to predict students' GPA scores, and selected the characteristics of students' behavior data in school, including access to the library, canteen consumption records, fetching water records and bathing records. However, the results of the model are not very satisfactory, we need visual technology to help us more easily analyze our data and diagnose our model more conveniently, so as to quickly find a breakthrough to improve the performance of the model.

Keywords: Visual analysis, education data, student performance, spatial temporal features, big data.

INTRODUCTION

Interactive model analysis, the process of understanding, diagnosing, and refining a machine learning model with the help of interactive visualization, is very important for users to efficiently solve real-world artificial intelligence and data mining problems [1,3].The measurement of students' school performance plays an important role in the evaluation of educational quality.

Therefore, research in recent years has paid considerable attention to the investigation of academic achievement. The data analysis of smart card has a wide application prospect. It can help track students' locations when using access control to restrict access to buildings, dormitories, libraries, and other facilities. The analysis of student card data is helpful to understand students' behaviors, especially the spatial and temporal characteristics, and insight into the behavioral differences between different groups of students, and can be combined with the analysis of performance, whether different behavior patterns will produce different performance benefits? On the one hand, these data sets provide valuable analysis resources for students' behavior analysis, on the other hand, they also pose a direct challenge to extract useful knowledge from large-scale spatio-temporal data. In order to solve these problems, we take advantage of the visual and interactive technology to explore and analyze the behavior of the students, and the random forest model to forecast the performance of machine learning, analysis the relationship between the behavior and performance, and the visual technology of the model of the predicted results show that the data preprocessing or characteristics can be analyzed and further on the choice of whether there is a problem, which leads to the prediction result is not ideal.

In this report, we propose a visual analysis system, namely eduDiag, which can interactively explore the relationship between behavioral data and performance based on the random forest prediction result data, and diagnose the problems in the model, providing help for the improvement of the model in the next step. Four views, namely model performance view, prediction data view, behavior sequence view and performance trend view, are proposed to visualize behaviors at different scales and to try to identify specific patterns Shared by different groups. Our advanced visual analysis systems are able to convey large amounts of information in a more efficient way without requiring much cognitive effort. In the experimental part, the performance was predicted by using the behavioral data of students at a certain level, and the predicted situation was displayed in a system, which diagnosed the places where the model could be adjusted for us, thus proving the effectiveness of the system.

VISUALIZATION DESIGN SYSTEM IMPLEMENTATION

The system consists of four modules :(1) prediction results module, (2) prediction classification module, (3) student behavior module and (4) academic performance module. The prediction results module shows the final accuracy, recall and accuracy of the performance prediction with random forest. In the prediction classification module, the collected prediction data are presented and compared with the data of different categories. Students can be divided into A, B and c. the three matrices in each block in the prediction classification module represent the size of the three types of data, and the small matrix in the matrix represents the number of data assigned to this class. If there are grid lines in the small matrix, the data are misclassified. The next three scatter plots in different colors represent the projected data of this class. From top to bottom, they represent the predicted students of class A, class B and class C, and one dot represents one student. The student behavior module shows the behavior patterns of students entering and leaving the library, consuming food in the cafeteria, bathing and fetching water in the dormitory, and the data will be used as the input characteristics of the random forest. The academic performance module is to draw the trend of students' performance predicted to be different from each other ,through the sankey diagram, which shows the trend of students' math performance in each category. This design approach provides more detail to diagnose areas in the model that can be optimized.

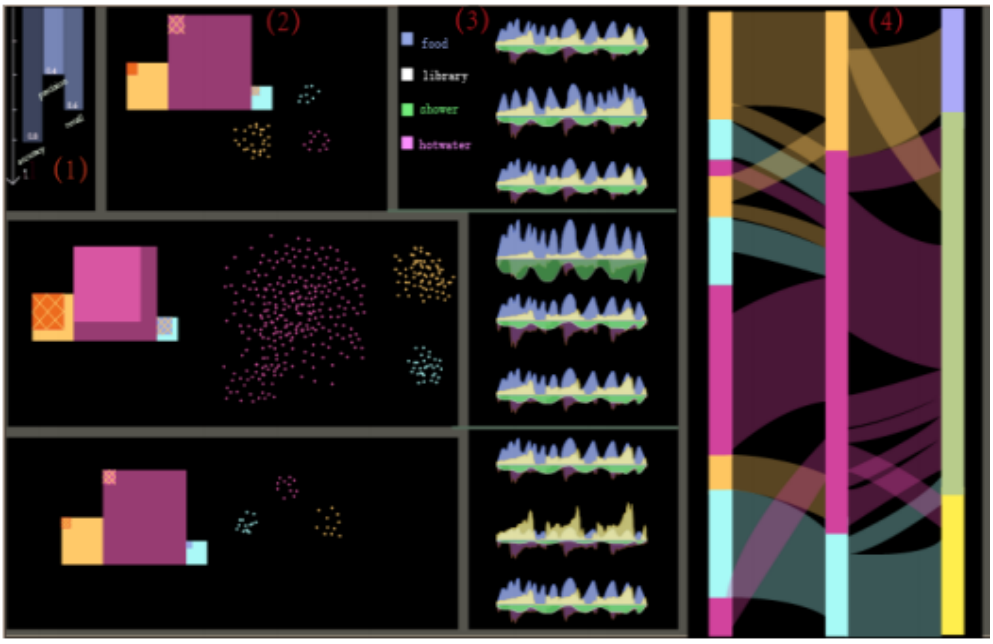

Figure 1: (1) prediction module, (2) prediction classification module, (3) student behavior module and (4) academic performance module

Case Study

We are authorized to use a set of anonymous data from the education big data institute of University of Electronic Science and Technology of China. In this study, analysis of consumption data in smart card data, library check-in data (including time, location, and encrypted student ids) was associated with academic records extracted from students' final gpa from 2009-2012.In this section, 1, 000 top students and 1, 000 bottom students were selected to verify the effectiveness of the system. The following conclusions: Let's look at module (1) first. The prediction accuracy of the model reached 80%, but the recall and accuracy of the model were not very good, only 60% and 40%.See module (2), the inside of the said three rectangular size according to the classification of grades A, B, C three kinds of the size of the number of students, we can see that most class B students, account data model class imbalance problems, and then looked at the second line, it is by mistake classification for the prediction of class B, A, C in many data have been misclassification in class B, this is because in the learning process of the model, model A strategy that in order to guarantee the improvement of accuracy, in order to make models don't learn, so I need to solve the problem of unbalanced class. and to the characteristics of different weights, Modules (3) and (4) show the behavior patterns and academic performance trends of students. We can see that the behaviors of students predicted to be class B are well differentiated between bathing and getting the hot water. Thus, different weights could be given to bathing and getting the hot water for training model, so as to achieve better classification results.

Conclusion

This paper presents eduDiag, a visual diagnostic analysis system for machine learning model based on the prediction of student behavior and academic performance based on smart card data. In this article, complex designs are flexible in scale and are integrated into graphics to aid statistical analysis. Based on the data test of a grade, the experimental results verify the effectiveness and efficiency of the proposed visual analysis task. In addition, according to the result analysis, our system can effectively diagnose the problems in data preprocessing and the problems in the model.

However, due to time and energy constraints, there is still much room for improvement in our work. Through the case study, we found that although the algorithm combined with time series can help us predict students' performance. Because the data set is sparse, it is not good for the prediction results of the model.In the future, we plan to design better methods to divide the data, better solve the class imbalance, and strive for more satisfactory model results.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61502083 and 61872066). We would like to thank all the participants involved in the studies for their valuable feedback, the reviewers for their constructive comments.

References

[1] Shixia Liu, Xiting Wang, Mengchen Liu, Jun Zhu. Towards Better Analysis of Machine Learning Models: A Visual Analytics Perspective Visual Informatics. 1(1): 48-56, 2017.

[2] Xun Zhao, Yanhong Wu, Dik Lun Lee, Weiwei Cui. iForest: Interpreting Random Forests via Visual Analytics,IEEE Transactions on Visualization and Computer Graphics, 2019.

[3] Liu Jiang, Shixia Liu, Changjian Chen. Recent Research Advances on Interactive Machine Learning. Journal of Visualization. 1-17, 2018.