SOSC4300

A Literature Review of Educational Data Mining

WANG, Yiren

QIU, Yirong

LIU, Chang

Introduction

Education has long been seen as an important component of society. Education can increase the human capital inherent in the labor force and improve labor productivity, leading to a transitional period of economic growth to a greater level of balanced output (Mankiw et al. ,1992). What's more, education can promote economic growth by fostering the spread and distribution of the knowledge needed to understand. It leads to the successful implementation of new technologies developed by others (Benhabib and Spiegel, 1994).

Educational data mining (EDM) is a new trend in the field of data mining and Knowledge Discovery in Databases (KDD). As an interdisciplinary field of study, it focuses on mining useful patterns from educational information systems and is associated with the application of different strategies for analyzing data generated in educational environments in order to discover useful knowledge. The goal is to produce models that can lead to improved learning experiences and greater institutional efficiency. Scholars in the field specialize in revealing useful insights to help educational institutions do a better job of managing their students, or to help students better organize their education and outcomes to raise their performance. In the past few years, educational data mining has received a great deal of attention. When many data mining techniques have been proposed to extract hidden knowledge from educational data, the extracted knowledge starts to help the institutions to improve their teaching methods and learning processes. All these improvements help in improving student performance and overall educational output.

Predicting students' academic outcome is universally useful for educational institutions that seek to enhance students' performance. Given a certain dataset, we could apply statistical methods to reasonably predict the outcomes. With the result of prediction, teachers can give more personalized and targeted instructions to students. Ever since data mining and machine learning methods became popular in the social science discipline, they are widely used to predict students' academic performance. For example, Educational Testing Service (ETS) has examined (Wei, 2013) the relationships between scores and the various background variables in order to determine the extent to which background variables predict scores. One of the techniques researchers used was multilevel modeling. The researchers concluded that background variables were more precise predictors of scores at the administration level than at the individual level. The researchers therefore argue that the application of the statistical techniques identified in this research to examinees' backgrounds at the administration level could be used to continuously evaluate the performance of the Test of English for International Communication (TOEIC) tests across administrations. The study effectively proves the necessity and validity to predict learning outcomes (scores) with background variables. Our research topic is predicting student learning outcomes with machine learning methods. This literature review aims to look through the existing methods used in EDM and further review related works to show the advantages and disadvantages of these methods.

Traditional methods and limitations

A common traditional research method for predicting student performance is to conduct additional tests and use the results as a benchmark for prediction. Lindy, Gerald and Steve (2001) conducted a study of using oral reading rate to predict student performance on statewide achievement tests. A total of 77 students participated in the study. The students were presented with an intact passage to read and prompted to read aloud for 1 minute during which the student's correct words are tallied and errors are recorded. Students usually time in three different paragraphs and record the median of these three paragraphs. Data collected from these measures of reading rate are often used to inform a broad range of educational decisions including progress monitoring, prereferral identification of students, and classification decisions. More importantly, the reading rate is used to predict students' reading and math performance on statewide achievement tests. The results indicate that the Pearson correlation coefficient between correct words per minute and statewide reading assessment is 0.66 and between correct words per minute and statewide math assessment is 0.53.

This is a typical traditional study of predicting students' performance. Based on the performance of students in reading passages, the teachers predict the performance of these students on the statewide achievement tests correspondingly. However, this kind of prediction is not rigorous and the accuracy rate is low, because the result of the prediction is based on only one variable. This simple component demonstrates the limitations of the research data since the data was created for this specific research and it can only be used for this research topic. In addition, performing an additional test is time-consuming and needs to be performed frequently to follow

up, so the number of participants is very limited, which means that this method can only predict the performance of a small number of people, and cannot make large-scale predictions.

Educational Data Mining

Obviously, the oral reading rate does not apply to other aspects and the methodology works with low efficiency. Nowadays, a lot of educational institutions and social science departments restore huge volumes of data, including family background, enrollment, attendance, etc. Data mining research from these datasets yields teaching and learning processes. Traditional data mining algorithms cannot be directly applied to educational problems, as they may have a specific objective and function. This implies that a preprocessing algorithm has to be enforced first and only then some specific data mining methods can be applied to the problems. In order to create a predictive model, there are several data mining techniques, which are classification, regression, and clustering. The most commonly used technique for predicting student performance is classification. There are some methods under classification such as decision tree (DT), Naive Bayes (NB), artificial neural networks (ANN), and clustering.

1. **Decision Tree**

A decision tree is a set of conditions arranged in a hierarchical framework. Due to its simplicity, many researchers have used this technique to transform it into a set of classification rules. C4.5 (Quadri and Kalyankar, 2010) and Classification And Regression Tree (CART) are the two popular decision tree algorithms. A study used decision tree algorithms to predict students' final

grades based on data about their usage of the Moodle system (Romero et al., 2008). Moodle is an online learning platform or course management system (CMS) that is frequently used by students. The real data was collected from seven Moodle courses at the University of Cordoba and the students were divided into two groups: 'passed' and 'fail'. The purpose of this research is to divide students with the same grades into different groups based on the activities carried out in the online course.

Another popular decision tree algorithm is ID3 decision tree. A study investigated 50 students who enrolled in a specific course program within 2007 to 2010 (Baradwaj and Pal, 2011). There were multiple performance indicators, such as 'Class Test Grades', 'Class Test Grades', 'Attendance' and etc.. ID3 is used to build the decision tree and if-then rules to eventually help teachers and students better understand and predict student performance at the end of the semester. ID3 decision tree is chosen in Baradwaj and Pal's study because it is a simple algorithm.

2. Naive Bayes

Naive Bayes is a simple probability classification technique, it is based on applying Bayes' theorem with strong independence assumptions between the features. Hien & Haddawy (2007) used Bayesian networks to predict CGPA of applicants based on their background at the time of admission. Today, educational institutions need a way to evaluate qualified applicants who graduated from various institutions. This research introduces a new method that integrates case-based components with predictive models. Case-based component searches for past students who are most similar to the assessed applicant. The challenge is to define the similarity

of cases in a manner consistent with the predictive model. This can be used in institutions with a good database of student and applicant information.

3. Regression

Another preprocessing algorithm in EDM is regression. A group of research teams in Morocco (Aissaoui et al., 2020) conducted multiple linear regression-based research to predict student's learning outcome. The researchers aimed to select among different machine learning methods to determine the most important variables that contribute to building a student performance prediction model. As this work mainly focuses on student modeling, it is one of the educational data mining objectives. The researchers concluded the various data mining techniques in former literature, such as classification and regression. He then analyzed the merits and drawbacks of each algorithm. While most of the current works tend to use all attributes to generate the prediction models, this paper only selects crucial variables. The author first decided the most important variables using multivariate adaptive regression splines (MARS) methods, and built seven different models based on the top five attributes. This study is a relatively simple but successful attempt of educational data mining.

4. Clustering

A cluster is a group of objects that are similar to each other within the group and not similar to objects belonging to other clusters. The clustering methods are usually sorted according to the approach used to implement the algorithm. Centroid-based clustering, graph-based clustering, grid-based clustering, density-based clustering, neural network-based clustering, etc. Therefore,

one can find algorithms implementing partitioning/hierarchical and hard/soft methods among these methods. Buehl & Alexander investigated students' epistemological beliefs about knowledge attainment and their learning processes (Buehl & Alexander, 2005). The aim of this study was to examine epistemological beliefs and students' achievement motivation. This study was distinctive because the authors did not examine whether individual beliefs were correlated or co-correlated to achievement and motivation. They tested different belief configurations related to students' beliefs about competence, achievement values, and learning of texts. The sample size was 482 undergraduate students, analyzed their beliefs about historical and mathematical knowledge, ability levels, and achievement values.The data were analyzed using Ward's minimum variance hierarchical clustering technique, which showed that students with different epistemological beliefs differed in their ability beliefs and achievement values. They recommend that future research could apply cluster analysis to analyze the different belief configurations related to various aspects of student learning.

In another study(Kock & Paramythis, 2010), a discussion of problem-solving behavior, different types of behavioral patterns of learners, as well as how these patterns can be discovered automatically is discussed. The objective of the application of this approach is to identify patterns based on the directional and automatic clustering of the user's problem-solving sequence. The new and innovative aspect of this research is that the clustering has been applied under three different behavioral patterns. The first level (pattern-driven), which uses an established predefined problem-solving approach, aims to reveal these patterns in student behavior. Two clusters with a Trial and Error problem-solving style were found after clustering a dataset of eight clusters. In the second stage (dimension-driven), the system seeks to identify the

given dimensions and then contributes to finding the specific styles relevant to those same dimensions. The third level (open discovery) is aimed at automatic discovery of learning and dimension styles. A basic importance of this work is the use of a set of optimization metrics that are applied to the implemented clusters to identify whether the optimal clustering settings have been attained.

Clustering can provide insights about variables that are relevant to individual clusters. Educational data are often multi-level and non-independent, so researchers must carefully select clustering algorithms that demonstrate the research question in order to obtain valid and reliable results. Scholars also must be careful when reducing the data size because when representing data in the form of fewer clusters typically loses certain fine details similar to lossy data compression.

Advantages of Using Computational Social Science (CSS) Methods

Compared with traditional quantitative research methods, computational methods show several obvious advantages. Firstly, The data used in computational social science research is the most important component. CSS data includes multi-dimensional information such as social relationships, social interactions, text, audio and video, and most of them are recorded with accurate and precise time. Moreover, most of the data is generated by ordinary users in the natural environment and is not deliberately created. Traditional data such as questionnaires are customized for a particular research topic, so interference and bias may exist. CSS data provides multi-level information in an analytics-friendly way.

The second advantage is the efficiency and effectiveness of data collection. The powerful functions and ease of use of calculation algorithms increase the speed of retrieving digital traces and reduce costs. Traditional research usually requires a lot of time and energy, and the cost is relatively high, such as field surveys and follow-up studies , which are commonly used and time-consuming. But for CSS, obtaining thousands of digital records for a single research project is not a difficult task. In addition, digital traces can be accessed almost in real time, which makes it possible to track the emergence and evolution of social phenomena or human behaviors over a longer period of time.

The third advantage of CSS is the enhanced modeling technology in data analysis. CSS data has accurate time, and is multi-dimensional and large-scale, which helps to use enhanced modeling techniques to reveal complex mechanisms and subtle patterns in human communication. Such modeling techniques may include time series analysis, sequential modeling, mixed effects modeling, network analysis and spatial modeling (Peng, Liang & Zhu, 2019). The results and methods will also help to define future steps for the EDm research, including possible transformations of the dataset, tuning the classification algorithms' parameters, etc (kabakchie va, 2012).These enhanced modeling techniques allow CSS researchers to review classic theoretical issues through powerful research designs, or explore new research issues with new perspectives and methods. Also, CSS methods that are novel to the social sciences act as a function of the accessibility of powerful open-source tools (LSE, 2017). For instance, network analysis has long occupied social science researchers, yet a new wave of studies leverages advances in computational techniques for inferences in a degree of detail and complexity that was previously unthinkable.

Shortcomings of Using Computational Social Science (CSS) Methods

Although the CSS method has the above advantages, there are reasons to admit and discuss the limitations of the CSS method. CSS data on social and mobile media was discovered casually, rather than empirical research conducted by researchers for different purposes of attention, so researchers have very limited control over the representativeness of research objects related to the discovered data. A question worth pondering is whether users who generate digital traces on the Internet can represent a social group, which also means that researchers must be very cautious about the validity of the variables constructed from the discovered data.

Second, most data on the Internet is related to human behavior. They provide very limited information about the basic psychological characteristics of human behavior, such as attitudes, emotions, and motivations. Although there are computational algorithms that can infer psychological characteristics from behavioral traces, it requires a rigorous evaluation of the validity of the measurement of derived psychological characteristics based on certain facts.

The CSS method is easy to build models that connect different factors, but the mechanism remains unexplained. Focusing on a mere data-driven standpoint might cause ignorance of the fact that every data represents a unique human being. And when the dataset has too many observations, we might run into an overfitting issue in the regression model.

Resolving data quality issues in predicting student academic performance is one of the biggest efforts in EDM (Nuckowski, M. & Satyanarayana, A, 2016). Previous papers often focused on one classifier and no filtering on student data has been performed.

Finally, in CSS research, protecting user privacy is very important, and it has become an increasingly important issue to strike a balance between protecting the commercial interests of social media and protecting the rights of researchers. There is no doubt that the privacy of users and the commercial interests of the media platform should be respected and fully protected. However, they should not harm true public research.

Conclusion

The analysis of student data and information to make better decisions or improve student performance is an interesting area of research. The main focus is to analyze and understand the educational data of students, indicating their educational performance and generating specific rules, classifications and predictions to help students in the future. Using Computational Social Science Methods, researchers can deal with multi-dimensional information efficiently and apply complex statistics models to educational situations. The suitable and optimal CSS method should be sought during in terms of accuracy and precision, as well as considering the feature and nature of variables, since the shortcomings do occur in some occasions.

References

Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of students ' performance in E- learning environment using random Forest. International Journal of Innovative Computing, 7(2), 1–5.

Aissaoui, O, E. et al. (2020). A Multiple Linear Regression-Based Approach to Predict Student Performance.

Cortez, P., Silva, A. (2008). Using data mining to predict secondary school student performance. In: Proceedings of 5th Annual Future Business Technology Conference, pp. 5–12.

Baradwaj, B.K. and Pal, S., (2011). Mining Educational Data to Analyze Students' Performance. *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 6, 2011.

Benhabib, J. and Spiegel, M. M. (1994). The role of human capital in economic development: Evidence from aggregate cross-country data. Journal of Monetary Economics 34(2), 143–174.

Dutt, A., Ismail, M, A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya.

Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. Journal of Medical Systems, 43(6). https://doi.org/10.1007/s10916-019-1295-4.

Hien, N.T., & Haddawy, P. (2007). A decision support system for evaluating international student applications. *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, F2A-1-F2A-6.

Iatrellis, O. et al. (2020). A two-phase machine learning approach for predicting student outcomes. Education and Information Technologies. https://doi.org/10.1007/s10639-020-10260-x

Lee, K. (2018). Machine learning approaches for learning analytics: Collaborative filtering or regression with experts ? Korea, 1–11.

Lindy Crawford, Gerald Tindal & Steve> Stieber (2001) Using Oral Reading Rate to Predict Student Performance on Statewide Achievement Tests, *Educational Assessment, 7:4,* 303-323, DOI: 10.1207/S15326977EA0704_04

Mankiw, N. G., Romer, D., and Weil, D. (1992). A contribution to the empirics of economic growth. Quarterly Journal of Economics 107(2), 407–437.

Quadri, M., & Kalyankar, N. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. *Global journal of computer science and technology, 10*.

Rai, S. et al. (2020). Machine Learning Approach for Student Academic Performance Prediction. *Evolution in Computational Intelligence.*

Romero, C., Ventura, S. & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384. Elsevier Ltd. Retrieved November 8, 2020 from https://www.learntechlib.org/p/66484/.

Tai-Quan Peng, Hai Liang & Jonathan J. H. Zhu (2019) Introducing computational social science for Asia-Pacific communication research, *Asian Journal of Communication,* 29:3, 205-216, DOI: 10.1080/01292986.2019.1602911

Wei, Y. (2013). *Monitoring TOEIC Listening and Reading Test Performance Across Administrations Using Examinees' Background Information.* Powers, Donald E. (ed.) The Research Foundation for the TOEIC Tests: A Compendium of Studies: Volume II. Princeton, NJ: Educational Testing Service, Sep 2013, p11.1-11.28
Retrieved from
https://www.ets.org/research/policy_research_reports/publications/chapter/2013/jrou

M.M.BuehlandP.A.Alexander,‘‘Motivation and performance differences in students’ domain-specific epistemological belief profiles,’’ Amer. Educ. Res. J., vol. 42, no. 4, pp. 697–726, Aug. 2005.

M.KockandA.Paramythis,‘‘Towards Adaptive Learning Support The basis of behavioural patterns in learning activity sequences,’’ presented at the 2nd Int. Conf. Intell. Netw. Collaborative Syst. (INCOS), Nov. 2010, pp. 100–107.