Advances in Social Science, Education and Humanities Research, volume 195
International Seminar on Education Research and Social Science (ISERSS 18)

Big Data Analysis in General Education: Opportunities and Concerns

Wenjun Lyu
School of Economics
Shanghai University
Shanghai, China
Lvwj2000@163.com

Zhaoqing Feng
School of Economics
Shanghai University
Shanghai, China
JenniferF5116@163.com

***Abstract*—General education, as important part of higher education, is putting increasing attention on big data courses in universities. This paper discusses the opportunities and concerns about big data analysis in general education by stating predictive analytics, personalized learning planning and introducing machine learning techniques in these aspects. A basic framework of opportunities, an integrated application process, machine learning applications are introduced and performance prediction as well as course recommendation system are given combining the specialty of general education. Concerns are suggested from open online platform, universities, governments and institutions that more security is hard needed in cross-disciplinary learning environment as opportunities are multifaceted.**

Keywords—Big Data; Education Data Mining; Predictive Analysis; General Education

I. INTRODUCTION

Big data is a popular term describing the dataset with exponential growth, availability and use of information, both structured, semi structured and unstructured. It brings better students learning curves and courses recommendation with characters of high volume, velocity, variety and veracity, applications of big data in education. New analysis techniques are needed for processing data as the amount of data grows unprecedentedly fast as well the data complexity. Predictive analytics/forecasting models in a big data environment enable institutions to make right investment decisions for higher institutional impact. Machine learning algorithms are applied to build a mapping function for prediction analysis. Supervised classification models such as SVM, Logistics Regression, Random Forest and Gradient Boosting Decision Tree are the common approaches to get those decisions.

Big data is increasingly applied in education sectors as well mostly with the target of educational data mining and learning analytics. Education data expanded largely giving an unprecedented chance for related institutions research. Many education institutions have long standing data warehouses and have used analytics tools make personalized learning plans, and analyze potential students and manage the experience while they are enrolled. Meanwhile, criticism and concerns rise on this education digital ocean since regulatory gaps and privacy policy inadequacy. Abundant data sources and channels may cause tracking and hacking risk of students' individual sensitive information such as test scores, behavior problems and learning disabilities, and social platforms interactions which would easily lead to commercial fraud. Relevant policymaking is supposed to consider these "vicious hands" for unjustifiable purposes.

II. RELATED WORKS

Typical researches on big data in education are educational data mining and learning analytics. Educational data mining is defined as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in". Learning analytics has different definitions because of different available analytics tools with the focus on capitalizing on the modelling capacity of analytics (Ackerson, 2006) and student success is one of the top issues.

Yu-Sheng Su et al (2017) propose a mechanism including four steps to enable big data analytics: data collection, data storage, data analysis, and data application. Jiajun Liang et al (2016) describe a complete approach to cope with e-learning dropout prediction analysis and use several supervised classification models including SVM, LR, RF and GBDT models. Nazarenko et al (2017) introduced a data-driven approach to big Data processing. Tree based methods empower student predictive models with high accuracy, stability and ease of interpretation (Vyas & Gulwani, 2017). Similar research has been done using supervised learning algorithms to illustrate opportunities for big data analysis by Cen et al. Argument around data mining methods tells two steps of classification algorithm: 1) collecting all identified attributes and apply algorithms on them; 2) applied attribute selection algorithm and ranked all attributes on the basis of their occurrence (Vyas & Gulwani, 2017).

This paper contributes to combine the opportunities and concerns of big data analysis in general education based on the major related research topics with new analytical approaches to specify the unique role played by general education in higher education domain.

III. OPPORTUNITIES OF BIG DATA ANALYSIS IN GENERAL EDUCATION

Traditional teaching mode pays attention to the construction of disciplinary knowledge system and teachers'

dominance, emphasizes one-way transfer of knowledge in class. Although it solves the training problem of skilled talents, but it cannot adapt to the situation of each learner and neglects the learners' personalized needs. Learning analytics comes more from semantic networks, intelligent curricula and system interventions. Educational data mining is more derived from the results of educational software, student modeling and predictive curricula. Using human judgment is very important in learning analytics, and automated discovery is a tool for achieving this goal. Automated discovery is the key to using human judgment as a tool to accomplish this goal in educational data mining.

The development of big data can realize the optimization of teaching. Teachers can get the real information of every learner conveniently in school. The content, method and process of teaching can be customized according to the learners' condition for describing the characteristics of students, perceiving the students' learning needs, guiding the students' learning process, diagnosing the students' learning results, and so on. It summarizes learning styles and data from massive learners learning behavior through tracing and collecting the big data of learners, so that learners are likely to get the most suitable for their own development. This will not only improve teachers' work efficiency and students' learning efficiency, but also be able to truly teach students in accordance with their aptitudes and cultivate personalized and innovative talents that meet the needs of the information era.

Researchers discuss opportunities of big data applications in general education as follow aspects [6]:

- Increased possibilities for students to wide learning paths and individualize study proposals when providing guides in a professional way especially in a competitive education environment.
- Courses timely and proper adjustments to satisfy various learning needs of students with different learning styles and related trouble spots including problem solving, deep thinking, conceptual understanding and skill assessing.
- Techniques such to cut opportunities cost in the time-consuming process by information redundancy of the profound knowledge system.

A. Educational Data Mining for Performance Prediction

Educational data mining is a critical work for institutions to extract, analyze and manage vast volumes and dimensions of data. Nation longitudinal data systems can track individual students from pre-K through college and work. Therefore, predicting students' academic behaviors makes for outcomes of how well individuals and teams perform on different learning assignments. Big data processing includes crucial data resources and suitable analytical tools to make efficient data-driven approaches. These approaches could provide informed guidance, advice and early feedback that may help to improve student's knowledge retention, formal assessment outcomes and satisfaction from the educational experience and they can also be used to identify at-risk students who are likely to fail.

Although performance prediction can be complexed, procedures can be simplified as follows:

1) Define students, including their levels of knowledge, comprehensive ability, personality match, learning capacity, knowledge reserve and other respects.

2) Collect and extract student data in step one in a dataset and do preprocessing especially focusing on learning time and efforts which almost affect learning outcomes.

3) Draw thinking modes and learning styles from the step 2 which wields core influence to improve students learning achievement.

4) Integrate all learning messages and extract learning habits to perfect the fitted learning curves.

Furthermore, learning performance prediction can create equal chances for general education to advance student interest unequivocally. It is useful for subject's interest orientation setting for lowerclassmen in both professional academies and comprehensive universities.

B. Course Recommendation System

The secondary and higher education students are usually offered with a wide selection of courses and modules. They have a lot of choices in both formal classroom learning and informal platform-learning or e-learning [7]. How to make a correct choice may be a realistic dilemma for them. Course recommendation system derived by big data analysis can make a difference.

Driven by the education of big data technology, through deep excavation and analysis of education big data, teachers can quantify the learner's learning process and learning status. They can pay attention to the identification of the correlation, and emphasize the determination of causation. We can find out the actual problems in the education system, find the influencing factors and intervention strategies more quickly and accurately, discover the true new rules of education, and greatly expand the breadth of exploring educational laws.

In addition, big data in education can also analyze the data generated by learners in learning processes, predict their learning modes and learning ability, find out potential studying problems and systematically improve the model of teaching in colleges and universities [6]. Schools' information technology application ability and teaching service level will also benefit.

C. Learning Analytics and Personalized Learning

Varieties of interactions and feedbacks according to courses are critical to improve educational efficiency thus leading to the significance of learning analytics. Knowledge reserve and study habits, personalized and individualized learning plans prove the way to maximize learning results in consideration of different learners' thinking models.

Big data allow researchers in education industry to judge the feasibility in and both advantages and disadvantages from a completely new perspective. It demonstrates that every learning state can be noticed by using methods better than

traditional ways and it is possible for each student focusing on his own path for new study plans.

A description of the learning analytics system can be seven steps: learning demand understanding, educational data understanding, data preparation and preprocessing, model selection and planning, feature engineering and model building, evaluation, reframing and optimizing, and monitoring and reviewing. The dimensions of analytics in general education procedures including data gathering, storage and management systems, analytics system, visualization and covers pre-existing, machine-readable data [2].

D. Machine Learning Applications

Based on proper technologies, educational data mining can be efficient and precise. Machine learning is used to make predictive analytics with the process of classification, clustering and regression. Common algorithms include K-Means, decision tree, random forest, logistic regression, SVM, naïve Bayes, neural networks and so on [5].

- Extract the meta information about courses from online learning platform in behavior logs by web crawler.
- Data preprocessing and distinguish records and objects as key pairs.
- Detail and redefine the students learning feedback through feature engineering.
- Train and tune classification models with the original dataset.
- Get the best accuracy from machine learning models outcomes.

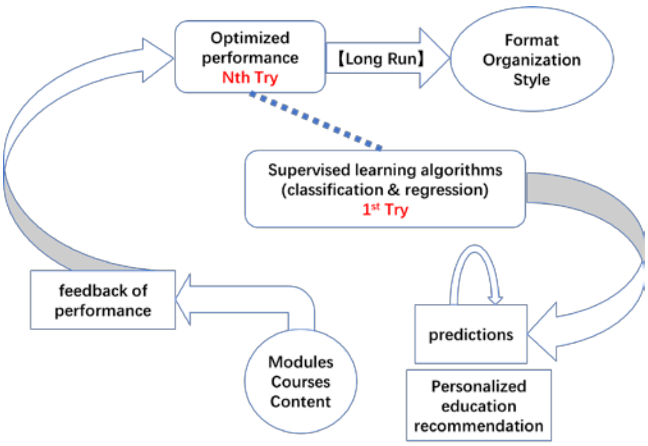

Fig. 1. Integrated Application Process of Supervised Learning Algorithms.

The popularity of big data open service platform among students in general education implies the necessity of selection from category and labels to obtain classifications. Multiple criteria and dimensions should be put into classification analysis such as datasets authenticity, data sources reliability and data quality timeliness [4].

IV. Concerns of Big Data Analysis in General Education

The content of the concerns and criticism of big data involves privacy and security considerations and other aspects. Privacy can be so circumstantial that a person may not necessarily care if his or her contact information is seen on Yellow Pages, but overly concerned if the same information is seen on a police database [1].

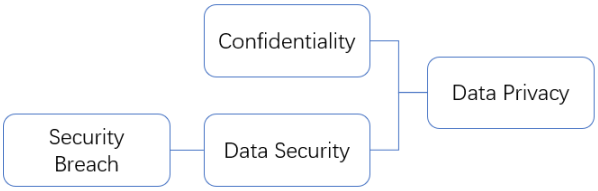

Fig. 2. Fully hosted cloud-based solution.

Data privacy means the ability of an organization or individual to determine what data in information systems can be shared. Confidentiality "pertains to the treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others in ways that are inconsistent with the understanding of the original disclosure without permission" after information shared. And essential data security is a means of "preventing unauthorized access to data and includes standards that can be followed to maintain proper access to data." A security breach may be intentional or unintentional, which stands for the leakage, theft, loss or other unauthorized access to data containing sensitive personal information that results in the potential compromise of the confidentiality of the data.

A. Education Policy

As mentioned, among the mounting challenges are data security, privacy protection as well as ethical boundaries of accessing personal digital data. The first of them is based on lack of data due to many countless activities, especially in class education and communications with lecturers and other students. It is easy to collect data on computer-based education, but it definitely is not so in case of words of mouth communication [5]. The term words of mouth mean words that are said by one person to another or to a group directly in personal conversation without any media resource usage. The meaning of the words of mouth in the education process is very important. However, verified data collection of this kind is practically impossible.

These online comments and messages, as the new sources of public opinion, are of great value for policymakers to understand problems and public needs, formulate policies to address them, evaluate policy effectiveness, and even engage the public in idea generation and problem solving (Charalabidis et al. 2015; Prpić et al. 2015; Schintler and Kulkarni 2014).

At present, examination data, student status data, teacher data, career data, funding data, population data and research data are all scattered among different agencies and government departments, making it hard to form big data. This is an issue that needs government departments to consider overall.

The disconnection among data silos are conducive to data security, primarily because the impact of data security breach in one database is localized without triggering the data breach of the entire data system. Therefore, policies and legislations are in dire need to prevent student learning data from being abused and misused, demarcating a fine line between using data to facilitate student learning and using data to deny the credit of students' further development. A recent debate was stirred by an experiment about emotional contagion on Facebook (Kramer et al. 2014).

However, the topics of education data mining and learning analysis are all centered on the students' learning and cognitive development. The problems of education public opinion analysis and moral education evaluation are rarely involved in the education. The online data generated in nature reflects the reality of society and education. How to use the big data to understand the students' true academic burden, to understand the satisfaction of the society with education and to understand the moral level of the students need to cooperate with education researchers and data scientists to solve the problem.

B. Privacy Concerns

With the certainty of growing and innovating of the educational data volumes and big data analysis technologies, privacy concerns arise from hacking cases and private information leakage in open online platforms. Big data applied in hidden interconnectivity of cross sections indicate an attribution of deletion difficulty. It is close to impossible to eliminate the data from public data systems. Policymakers, researchers and education practitioners who have an advanced understanding of big data have proposed several aspects of countermeasures. For example, a proactive approach to address data privacy during the data collection phase can control the leakage of private information during data ingestion, management, integration, operation phases and enforce security by putting more requirements of authentication to data access.

Privacy concerns about general education in universities are more meaningful since it is an important link of undergraduate education to develop citizens and spread common values [3]. This issue can be stated by four aspects: learning platforms, colleges and universities, the government and research institutions. Learning platforms provide choices for students eager to get into other subjective areas; colleges and universities are the provision side of these student users.

Maintaining general education network and data security is supposed to firstly establish safety systems for most open online learning platform. And local colleges should pay more attention to publicity efforts for students' individual message protection. The government is responsible for providing data or organizing data collection, research institutes responsible for designing research frameworks and analysis results, and professional data analysis firms responsible for software development and providing data analysis models. This relates data security and storage issues, because big data will involve the student's personal information and the reuse of big data may lead to a tighter contract work system to ensure the safety of students' personal information, which should not be used for commercial purposes (such as learning product ads locating), but also to determine who will be responsible for the storage and maintenance of data, who have the right to secondary development and utilization of the data. Generally speaking, the non-profit organization authorized by the government or government is responsible for the data storage and management. If the commercial company needs to reuse the database after completing the tasks, it needs to obtain the government's authorization again to prevent the company from collecting the government commissioned by the government data for commercial development and commercial training.

V. CONCLUSION

The emerging highly flexible and scalable approaches to data processing and analysis allow us to extract new insights and meaningful information from educational data that can benefit students, teachers and the whole education ecosystem. Big data include traces of student-to-student interactions [2]. At present, many big data problems related to social sciences, such as public opinion analysis and sentiment analysis, have not been considered in the past and have just begun to be studied. With the help of big data reasoning, social sciences will take off the coat of "quasi-science" and really move into the hall of science.

The construction of a big data platform for education still needs state-led construction. However, big data technology has more bottom-up kinetic energy and a more maverick self-development force. In the future, personalized teaching based on big data, scientific evaluation, elaborate management, intelligent decision-making and precise scientific research will play an inestimable role in promoting education fairness, improving education quality and cultivating creative talents.

REFERENCES

[1] J. F. H. Barril, Q. Tan, "Integrating privacy in architecture design of student information system for big data analytics," IEEE Conference on Cloud Computing and Big Data Analysis, pp. 139-144, 2017

[2] F. Matsebula, E. Mnkandla, "A big data architecture for learning analytics in higher education," IEEE africon: Science, Technology and Innovation for Africa, pp. 951-956, 2017

[3] L. Lin. "A study on general education in local institutions of higher learning," 3rd International Conference on Management, Education Technology and Sports Science, METSS, 2016

[4] J. Shu, et al, "Exploration on college education big data open service platform," 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA, pp. 161-165 2017

[5] J. Liang, J. Yang, Y. Wu, C. Li, L. Zheng, "Big data application in education: dropout prediction in edx MOOCs," IEEE 2nd International Conference on Multimedia Big Data, pp. 440-443, 2016

[6] L. Cen, D. Ruta, J. Ng, "Big education: opportunities for big data analytic," International Conference on Digital Signal Processing, DSP, pp. 502-506, 2015

[7] M.S. Vyas, R. Gulwani, "Predictive analytics for e-learning system," International Conference on Inventive Systems and Control, ICISC 2017