Effectiveness analysis of machine learning in education big data

Ya Zhou¹, Zhuoqing Song¹*
¹Guilin University Of Electronic Technology, Guangxi Key Laboratory of Trusted Software, City, Province, 541000, China
*Corresponding author’s e-mail: 19032303042@mails.guet.edu.cn

Abstract. In the education big data environment, the online data and information of the current education industry learning resources supply become complex and diverse. Different scenarios and different analysis environments face different learners. Combined with the current education field in China, prediction, evaluation and analysis play an important application value and broad space for education and teaching. Machine learning technology has strong learning ability. Breaking the limitation of time and space, deeply analyzing the rationality of data relationship and obtaining the effectiveness evaluation are the effective means to explore this field. As well as a way to summarize the general rules and trends of using machine learning in education big data and educational data mining.

1. Introduction
With the rapid development of information technology and artificial intelligence, the field of education gradually develops to intelligent education. The core element of smart education is to carry out personalized learning, similar to personalized learning, the underlying key support technology is machine learning. Therefore, it is of great significance to study the application of machine learning in the field of education big data. The era of artificial intelligence and big data has come, and the education system is also developing at a high speed, which is totally different from the improvement process of computer technology on education in the past. In 2012, the United Nations pointed out in big data for development: Challenges & opportunities that big data will have an impact on different areas of society[1]. In October of the same year, the U.S. Department of Education released "promoting teaching and learning through educational data mining and learning analysis", which is mainly aimed at K-12 schools and local colleges and universities. In 2018, China put forward the national big data strategy to realize digital China smart society. Smart education and education big data are an indispensable part. As an interdisciplinary subject involving many fields, machine learning belongs to the core category of artificial intelligence, which enables existing computers to acquire self-learning ability. Machine learning can be divided into supervised learning and unsupervised learning through whether the input data is marked or not. Its research is an interdisciplinary subject involving many subjects[2]. The application of machine learning education under the background of big data includes search and iterative optimization, which has good performance in data processing, can obtain the potential laws of data, also it can be applied to the location data analysis and prediction. This will get greater value. So it is imperative to use machine learning in education big data research and analysis[3].
2. General process of machine learning research method

When we get a machine learning problem, the usual process is divided into the following steps:

Figure 1. General process of machine learning.

"Data and features determine the upper bound of machine learning, and models and algorithms only approximate this upper bound.” Data is crucial to the whole machine learning project. In the case of education big data, exploratory data analysis should be carried out after the obtained data set to understand the distribution of data.

Figure 2. Dataframe read with pandas. Figure 3. Education recommendation system.

2.1. Data preprocessing

In most cases, the data set has many problems such as data missing, uneven distribution and so on. It is necessary to further process the collected data, including processing missing values, data normalization, data conversion, and so on. For example, in the table below, we can use user based collaborative filtering to calculate the cosine similarity between users (the similarity between student A and student B and student C is 0.38 and 0.32 respectively), and then calculate the weighted average to fill in the missing value of student's score on course i5. calculation cosine similarity, as in (1).

Table 1. Student rating.

|   | I1 | I2 | I3 | I4 | I5 | I6 |
|---|----|----|----|----|----|----|
| A | 4  | 5  | 1  |    |    |    |
| B | 5  | 5  | 4  | 3  |    |    |
| C |    | 2  | 4  |    |    |    |
| D | 4  |    |    |    |    |    |

\[
similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (B_i)^2}}
\]  

(1)

2.2. Data Set Segmentation

Generally, the sample is divided into three parts: training set, validation set and test set. The training set is used to estimate the model, the verification set is used to adjust the model parameters to obtain
the optimal model, and the test set is used to test the performance of the optimal model. Usually we use a ratio of 7:2:1.

2.3. Model selection and training
After processing the data, select the appropriate machine learning model for data training. At first, we need to analyze the processed data to determine whether there are markers. If so, we should consider the model of supervised learning. Otherwise, it is divided into unsupervised learning for processing. After that, it analyzes whether the problem type belongs to classification problem or regression problem[4]. Only after determining the problem type, can we choose the appropriate model. In addition, the size of the dataset needs to be considered. If the data set sample is small and the training time is short, we usually consider some lightweight algorithms such as naive Bayes, otherwise we need to consider some heavyweight algorithms such as SVM. After confirming the model, it is a problem of optimization. Regularize the existing model, adjust the super parameters, modify the existing model repeatedly, train it, and evaluate it on the verification set until the model can perform well. For example, try to add dropout, different structures, add or remove layers, add L1 / L2 regularization, improve over fitting. Optimize iteration based on Feature Engineering: add new features, remove features without information.

2.4. Model Evaluation
After the construction of the model, we must evaluate the effect of the model, and continue to adjust the parameters, characteristics or algorithms of the model according to the evaluation results to achieve satisfactory results. However, the simplest and most commonly used index to evaluate a model is accuracy. However, if accuracy is used as evaluation index without any premise, accuracy often cannot reflect the performance of a model. For a model, we need to judge its performance from different aspects. Taking binary classification as an example, confusion matrix can reflect the performance of the model more comprehensively. Many indicators can be derived from the confusion matrix. Precision and recall are a pair of contradictory measures. Generally speaking, when the precision ratio is high, the recall rate is often low; when the recall rate is high, the precision rate is often low. Usually only in some simple tasks, it is possible to make both high. P of P-R curve is precision and R is recall. With P as abscissa and R as ordinate, P-R curve can be drawn. For the same model, different P-R values can be obtained by adjusting the classification threshold, thus a curve can be obtained (the ordinate is p, the abscissa is R). Generally, precision decreases and recall increases as the classification threshold changes from large to small (P is considered as greater than the threshold).

3. Opportunities and challenges under education big data
With the advent of the era of artificial intelligence, the technology applied in the field of education has not yet matured, which is both an opportunity and a challenge.

3.1. Large scale personalized learning recommendation system
The business nature of technology for education involves data, automation and personalization. In this era of rich content and channels, personalized learning recommendation has become an indispensable part, and accurate recommendation is based on the content that users (students) are interested in. In the modern intelligent education and teaching system, MOOC can provide more diversified teaching, including video, homework, examination, discussion and comment[5]. Taking NetEase cloud classroom as an example, it is one of the online practical skills online teaching platforms. The platform uses machine learning to analyze users' interests and recommends courses with high similarity to users. For different types of students, NetEase cloud conducts distributed processing, real-time tracking and recommendation, which greatly improves user satisfaction, and at the same time, recommends more courses on the platform to users. At the same time, association rules can also be used for personalized recommendation of learning platform, and the correlation between courses is the principle. Through the user (student) class observation records, rules are mined to discover the interest of students of the
same category, so as to realize modeling and course recommendation. However, association is not equal to causality. Common concepts include support, confidence and lift. For further research and elaboration, the definition and description of association rules are described as follows:

Definition1: support refers to the probability that two courses studied by the same student user occur simultaneously in the total user group.

\[
\text{Support}(A \cap B) = \frac{\text{freq}(A \cap B)}{N} \tag{2}
\]

Among them, freq \((A \cap B)\) is the number of events that occur simultaneously in course A and Course B, and N is the total record of transaction data set.

Definition2: confidence is the probability that users learn course B while learning course a. If the confidence level is high, it reflects that users are likely to study course B while learning course a. the calculation formula is as follows:

\[
\text{Confidence}(A \rightarrow B) = \frac{\text{freq}(A \cap B)}{\text{freq}(A)} \cdot \frac{\text{support}(A \cap B)}{\text{support}(A)} \tag{3}
\]

Definition3: lift can well measure the effectiveness of association rules and whether it has promotion effect. For example, when users study course a, the number of times they study course B at the same time is higher than that of course B alone, which shows that course a can improve course B. Generally speaking, if the promotion degree is greater than 1, the rule is valid, and less than 1 is invalid.

\[
\text{Lift}(A \rightarrow B) = \frac{\text{support}(A \cap B)}{\text{support}(A) \cdot \text{support}(B)} \tag{4}
\]

\[
\text{Lift}(A \rightarrow B) = \frac{\text{confidence}(A \rightarrow B)}{\text{support}(B)} \tag{5}
\]

The core step of association rule mining algorithm is to generate frequent itemsets, among which Apriori algorithm is the most representative. It is divided into two steps: the first step is to iteratively calculate the frequent itemsets in all transactions, that is, the item sets whose support degree is not lower than the threshold set by the user; the second step is to construct rules that satisfy the minimum confidence of users by using frequent itemsets. Table 2 shows the transaction data set. We set the minimum support degree as 50% and the minimum confidence level as 50%, and calculate the association rules by Apriori algorithm.

Table 2. Transaction data set.

| TID | Course                        |
|-----|-------------------------------|
| 10  | Data analysis, AI Engineer, Web |
| 20  | Python, AI Engineer, ML in Action |
| 30  | Data analysis, Python, AI Engineer, ML in Action |
| 40  | Python, ML in Action           |

The calculation process of our application of Apriori algorithm is shown in the figure 4. The occurrence times of Data analysis, AI engineer, Web, Python and ML in Action are 2,3,1,3,3 respectively. First of all, find out all 1 item sets, scan the data set, calculate the number of occurrences of 1 candidate set, so as to get their support. The calculated results form L1 set. The support degree of web security is 25%, which is lower than the support threshold we set earlier, so remove it to get the set K1. Then, the elements in K1 are combined to form a 2-itemset, and the support of the candidate itemset is calculated to obtain the L2 set, which will be lower than the support threshold \{Data analysis, Python\}, \{Data analysis, ML in Action\} remove to form K2 set. The elements in K2 are combined to form a 3-itemset, and the support degree of the candidate is calculated to obtain L3 set, and the threshold value lower than the support degree is removed. For frequent itemsets \{AI Engineer, Python, ML in Action\}, its nonempty proper subsets are \{AI Engineer\}, \{Python\}, \{ML in Action\},
According to this, the association rules are generated and the confidence degree is calculated. The result is shown in the table 3. It can be seen that the confidence values are more than 50%, so \{AI Engineer; Python; ML in Action\}. It is the final association rule.

### 3.2. Interaction between machine learning and educational data

Machine learning is also the core of the data field. It studies and analyzes the computer simulation of human learning behavior. Educational data mining uses machine learning, statistical learning and other technologies and methods to analyze and process educational data, and realize the interaction between them. The main methods are as follows: (1) Clustering. Using unsupervised learning to analyze the education data of unknown categories Data property, which divides the data set into different subsets. (2) Prediction. Using multivariate regression analysis or multivariate prediction to develop a single variable prediction model. In the experiment, the situation of learners' participation in tests and discussions is analyzed to predict whether the learners are successful in the course. (3) Relationship Mining. Find the correlation between the variables in the data set. The experiment is based on the relationship mining to find the relationship between students' learning activities and learning performance in the process of online learning, so as to improve the presentation of learning content and improve the quality of teaching. (4) Model Construction. Construct an effective interpretation model through clustering and correlation mining of educational data. For example, K-means algorithm is used to obtain the hierarchical label of learning events, and the original unsupervised learning samples are transformed into supervised learning samples. The interaction between machine learning and educational data is more integrated with intelligent education, so that the interaction between learners and the system can be realized intelligently. Through the field of education big data, we can guide students to learn what and how to learn, and deepen the characteristics of students to realize personalized service. At the same time, we should also consider the complexity characteristics of different students, and finally provide the most suitable learning content for them.

#### Table 3. Confidence calculation of course association rule

| Rules                           | Confidence |
|---------------------------------|------------|
| \{Python\} → \{AI Engineer; ML in Action\} | 0.50/0.75=0.67 |
| \{AI Engineer\} → \{Python; ML in Action\} | 0.50/0.75=0.67 |
| \{ML in Action\} → \{Python; AI Engineer\} | 0.50/0.75=0.67 |
| \{AI Engineer; ML in Action\} → \{Python\} | 0.50/0.50=1    |
| \{Python; ML in Action\} → \{AI Engineer\} | 0.50/0.75=0.67 |
| \{Python; AI Engineer\} → \{ML in Action\} | 0.50/0.50=1    |
4. Integration of Machine learning and intelligent education solutions

In the environment of education big data, the purpose of using machine learning program is knowledge discovery and education data mining. Smart education combines modern computer technology and appropriate education concepts, and is also based on education big data. In the 1990s, data mining (DM) technology has been applied in education and laid a foundation Machine learning in the field of education research[8], now, machine learning mainly in the depth of education data mining, to build a good platform for intelligent education, so effective machine learning is no doubt a sharp tool for the realization of intelligent education. The process is shown in the figure.

In the existing intelligent education platform, massive data information will be generated every moment, but not all the data are analyzable[9]. We usually need to carry out noise reduction, missing value processing, data cleaning to ensure the reliability of the data, and then build a model for data mining work, such as predicting the learning performance and behavior of learners through regression tree[10]. Finally, the results of the model are evaluated, including both data description and theoretical support. Knowledge discovery from the model is fed back to the system, iterative cycle model, and finally improve learning. In depth education of big data background, machine learning is still in its
infancy in this field[11]. It is imperative to promote the integration of technology and education in cross fields. Scene-based learning is also a major trend. Smart campus and smart classroom in Guilin University of Electronic Technology and technology have filled in this lack of mature cases.

5. Summary
Through the research and analysis of machine learning in education big data, the purpose of this paper is to enhance our multi-dimensional understanding of big data and artificial intelligence, strengthen digital literacy, learning ability, problem solving ability and so on. Since the application of machine learning in the field of education big data is still in the early stage of development, neither the data processing method nor the model is mature. In this regard, some suggestions are put forward to promote the deep integration of intelligent education and machine learning. At the same time, the overall planning of education big data should be carried out, and further efforts should be made for the systematic comprehensive research from top to bottom.

Acknowledgments
This work was supported by National Natural Science Foundation of China (No. U1811264).

References
[1] Big data for development: challenge & opportunities[db/ol]2012-05-01
[2] Graesser A C, Hu X, Sottilare R. Intelligent tutoring systems[J]. International handbook of the learning sciences, 2018: 246-255.
[3] Zhou Qing, Mou Chao, Yang Dan. Review of educational data mining[J]. Journal of software, 2015, (11): 3026-3042.
[4] Mehwish Naseer, Wu Zhang, Wenhao Zhu. Early Prediction of a Team Performance in the Initial Assessment Phases of a Software Project for Sustainable Software Engineering Education [J]. Sustainability, 2020, 12(11).
[5] Sahebi S, Brusilovsky P. Student Performance Prediction by Discovering Inter-Activity Relations[J]. International Educational Data Mining Society, 2011:192-200
[6] Information Technology - Data Mining; Findings from National Changhua University of Education in Data Mining Reported (Applying data-mining techniques for discovering association rules) [J]. Information Technology Newsweekly,2020.
[7] Romero C, Ventura S, Pechenizkiy M, Baker RS. Handbook of Educational Data Mining[M] CRC Press, 2011.
[8] Yu Fang, Liu Yanshen. Research on the application of "user centered" Education Data Mining[J]. E-education Research,2018,39(11):69-77.
[9] Gang Li, Fang Wang. Research on art innovation teaching platform based on data mining algorithm[J]. Cluster Computing,2019,22(6).
[10] Zheng Yafeng, Fu Qian, Zhao Yaning. Research status and theme structure of artificial intelligence education in China[J]. Digital Education,2020,6(01):21-26.
[11] Hu Yongxiang. Research on machine learning under the background of big data[J]. Information and computer,2019,31(24):15-16.