Student Network Behavior Analysis and Relevance Research Based on Optimal Decision Tree Algorithms

Geng Han

1 Beihang School, Beihang University, Beijing, 100191, China
*Corresponding author's e-mail: hangeng@buaa.edu.cn

Abstract. Based on the optimal decision tree algorithm, this paper proposes a student network behavior analysis model. Through training and predicting the academic performance of Beihang Grade 2013 undergraduates, it is found that students' online behavior has a profound impact on students' academic performance. To maintain a good learning state, students must strictly limit the time spent on the Internet in idle time, effectively control the time spent on the internet, and ensure that their daily sleep is not affected by the Internet behavior. The regularity of students' life is positively correlated with their achievement ranking.

1. Introduction
Contemporary college students have active thinking and rich ideas, which need the close attention of student management staff to better help them grow [1]. People's thoughts and psychology must be reflected through some kind of behavior. Most of these can be depicted and represented by data, thus forming the big data of people's thoughts and behaviors, which contains vivid characteristics of the times and rich educational value [2]. The aim of measuring, evaluating and improving the quality of ideological and political work is precise. The methods and approaches depend on big data analysis, that is, big data of education and big data of students. Classified data research on students' behavior is an important way to describe students' learning, life and ideological dynamics. With the continuous accumulation of data reflecting students' behavior, it provides conditions for students to better understand themselves by using data mining technology.

Academic performance is an important evaluation index of university learning quality. There has not been clear correlation between students' academic performance and students' behavior. Based on the decision tree algorithm, this paper proposes a student network behavior analysis model. Training and predicting the academic performance of all undergraduates of Beihang Class 2013, and constantly modifying the model. Based on the model, important behavioral characteristics of students are found and studied, and some universal conclusions are drawn.

2. Analysis of students' network behavior
Forecasting is the core value of big data. Therefore, the big data analysis of students is to study the contents closely related to students' education and training, including early warning of academic achievement and learning status, focus analysis of students' social hotspots, monitoring and dynamic prediction of network public opinion, tracking analysis of students' online behavior, capture and evaluation of students' daily life status, automatic mapping and diagnosis of students' psychological status, etc. [2] Most of the data sources involved in these studies can be obtained from the Internet.
According to the online registration records and traffic data of 21:00-24:00 hours, students' rest situation can be obtained. The normal class time is chosen to analyze the online behavior. According to the online registration records, IP login address and traffic data, the students' class situation can be got. Statistical analysis of the correlation between the total length of time spent on the Internet and academic performance of students can get the impact of online behavior on academic level. Therefore, students' daily work and rest, regularity of life, classroom learning quality and other behavioral characteristics can be characterized through the analysis of students' typical time on-line data, and students' individual learning and life status can be assessed as a whole.

Please follow these instructions as carefully as possible so all articles within a conference have the same style to the title page. This paragraph follows a section title so it should not be indented [2-3].

3. Decision tree algorithms
Data mining is essentially a process of knowledge discovery in databases. Generally, it refers to the process of discovering and extracting rules hidden in a large amount of complex data which have practical significance and learning value.

Decision tree method belongs to classification algorithm, and it is also one of the most intuitive and effective algorithms in classification algorithm. The decision tree first chooses attributes to establish the root node, and then generates a complete tree step by step. In each step, an attribute with the maximum entropy value is selected as the node of the branch of the tree. The decision tree completes when all attributes become branch nodes of the tree or the data of the same attributes belong to the same classification. After the decision tree is established, each leaf node is a specific class label, and the path from the root node to each leaf node is a specific classification process. At present, ID3 algorithm, C4.5 algorithm and XGBoost algorithm are widely used in decision tree algorithm. [3]

4. Development of student network behavior analysis model based on optimal decision tree algorithms
Taking the data of network behavior of 3199 students of Beihang Class 2013 for four years as samples, this paper uses decision tree algorithm to train and predict the average score. 80% of the total number is used for training and 20% for prediction. The error analysis of prediction data and statistical data is carried out. This paper designs "the difference between real value and predicted value" as the evaluation index of accuracy, and calculates that the accuracy of the difference within 8 points is 92%. The XGBoost decision tree algorithm is used to get the best results at present. The flow chart is shown in Figure 1. The difference between the real value and the predicted value is shown in Figure 2. The lightGBM algorithm and XGBoost algorithm are compared in Figure 3.

The data of students' surfing the internet is analysed and characterized. A total of 32 features are designed, of which the typical features are shown in Table 1.

![Figure 1. Technical framework of correlation analysis algorithm between student network behavior and academic performance](image-url)
Figure 2. Comparison of the difference between the real value and the predicted value

Figure 3. Comparisons of lightGBM algorithm and XGBoost algorithm

Table 1. Typical characteristics of data on Internet access

| Characteristic Design of Online Data                  | Unit   |
|------------------------------------------------------|--------|
| Total time spent online                              | day    |
| The total number of bytes on the Internet            | GB     |
| Average bytes per day                                | GB     |
| Average daily Internet time                         | minute |
| Number of Internet stays up late                     | times  |
| Average daily earliest logon time                    | minute |
| Average Internet bytes from Monday to Friday         | GB     |
| Average Internet bytes on Saturday and Sunday        | GB     |

XGBoost is a lifting method. Different from the traditional GBDT (Gradient Boosting Decision Tree), the traditional GBDT only uses the first derivative information when optimizing, while XGBoost expands the loss function with the second Taylor expansion and adds the first and second derivatives into the loss function. Regular terms are added to the loss function to control the
complexity of the model and prevent the model from over-fitting. The derivation process of XGBoost method is as follows [4,5,7]:

\[
\hat{y}_i = \phi(X_i) = \sum_{k=1}^{K} f_k(X_i), f_k \in F
\]  

(1)

Whereas \( F = \{ f(x) = \omega_q(x) \} \) \( q : R^m \rightarrow T, \omega \in R^T \) is a function space representing a decision tree, \( T \) represents the number of leaf nodes of a decision tree. The loss function is:

\[
L(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k)
\]

(2)

\[\Omega(f) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2\]

The goal is to minimize the loss function:

\[
\tilde{L}(\alpha) = \sum_{i} l(\hat{y}_i, \hat{y}_i^{(t-1)} + f_i(x_i)) + \Omega(f_i)
\]

The Taylor expansion is used to approximate the loss function:

\[
L_t \approx \sum_{i} [l(\hat{y}_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_i)
\]

Define \( g_i = \partial_{\hat{y}_i^{(t-1)}} l(\hat{y}_i, \hat{y}_i^{(t-1)}) \), \( h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(\hat{y}_i, \hat{y}_i^{(t-1)}) \). Then the target function becomes:

\[
\tilde{L}(\alpha) \approx \sum_{i} [g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_i)
\]

(3)

Define \( I_j = \{ i | q(x_i) = j \} \). According to Equation (3),

\[
\tilde{L}^{(\alpha)} = \sum_{i} [g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j} \omega_j^2
\]

(4)

\[= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) \omega_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) \omega_j^2] + \lambda T\]

Because \( q(x) \) has a fixed structure, the weight of leaf \( j \) can be calculated as \( \omega_j^*: \)

\[
\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}
\]

(5)

Substituted into the objective function:

\[
\tilde{L}^{(\alpha)}(q) = -\frac{1}{2} \sum_{j=1}^{T} (\sum_{i \in I_j} g_i)^2 + \frac{1}{2} \sum_{i \in I_j} h_i + \lambda + \gamma T
\]

(6)

Equation (6) can be used to calculate the score of tree structure \( q \). Generally, it is impossible to enumerate the structure \( q \) of all possible trees. A greedy algorithm is used to add a segmentation to the existing leaves every attempt. Suppose that \( I_L \) and \( I_R \) are the nodes after splitting the left and right sub trees. For \( I = I_L \cup I_R \), the loss function after segmentation is like Equation (7):

\[
L_{split} = \frac{1}{2} \left[ \left( \sum_{i \in I_L} g_i \right)^2 + \left( \sum_{i \in I_R} g_i \right)^2 \right] + \left( \sum_{i \in I_L} g_i \right)^2 - \gamma
\]

(7)

When searching for the best segmentation points, considering the inefficiency of the traditional greedy method of enumerating all possible segmentation points of each feature, XGBoost implements
an approximate algorithm. According to the percentile method, several candidates who may become segmentation points are enumerated, and then the best segmentation point is calculated from the above formula. [6]

XGBoost takes into account that the training data is sparse. It can specify the default direction of the branch for missing or specified values, which can greatly improve the efficiency of the algorithm, up to 50 times [7,8].

The sorted feature columns are stored in memory in block form and can be reused in iteration. Although algorithm iteration must be serial, parallelism can be achieved when processing each feature column. Storage by feature column can optimize the search for the best segmentation points. But when gradient data is computed in row mode, it will lead to discontinuous access to memory [9,10].

5. A discussion on the relevance between students’ network behavior and achievement

If the decision tree is too complex, it is difficult to understand. Therefore, the decision tree should be pruned. The root node of each decision tree is the criterion, and the leaf node is the output value. The importance of the XGBoost model is represented by the average information gain. The calculation formulas are as follows:

\[ Gain(D, a) = Ent(D) - \frac{1}{V} \sum_{i=1}^{V} \left[ \frac{|D_i|}{|D|} \cdot Ent(D_i) \right] \]

\[ (8) \]

\( D \) is the data set. \( a \) is the selected attribute, and \( a \) has \( V \) values. The data set \( D \) is divided by \( V \) values, and the data sets \( D_1 \) to \( D_V \) are obtained. The information entropy of \( V \) data sets is calculated and the weighted average value is obtained.

By analyzing the network behavior, the most important 5 characteristic parameters affecting students’ performance and their importance ranking can be obtained, as shown in Table 2.

| Ranking | Characteristic Design of Consumption Data | Importance |
|---------|------------------------------------------|------------|
| 1       | Average Internet bytes from Monday to Friday | 173.0      |
| 2       | Total Internet consumption                | 163.0      |
| 3       | Number of Internet stays up late          | 91.4       |
| 4       | Average daily earliest logon time          | 96.0       |
| 5       | Average bytes per day                     | 84.8       |

Staying up late to surf the Internet disturbs students' normal work and rest. According to statistics, as many as 80% of students stay up late and surf the Internet at different levels, and only a small number of students are for study, most of them are for entertainment, social needs and so on. The use of smart phones has become quite common among college students. The majority of college students use mobile phones to surf the Internet in class. Of course, there are many students who inquire about the content of classroom teaching. But if they spend too much time on the internet, they will miss the key points and details of classroom content, and will inevitably affect the learning effect, which is reflected by their academic achievements.

Analyzing the correlation between students' online behavior characteristics and their academic performance, it can be found that the more bytes they use from Monday to Friday, the lower their scores rank; the more the total amount of online fees consumed, the lower their grades rank; the earlier they log on online, the lower their grades rank; the more they stay up late, the lower their grades rank. The more bytes used on the Internet, the lower the ranking. The analysis results are shown in Figures 4, 5, 6, 7 and 8, respectively. It can be concluded that students' online behavior has a profound impact on their academic performance. In order to maintain a good learning state and ensure a better learning effect, students must strictly limit the free time of Internet access, effectively control the time of Internet access, and ensure that their daily sleep is not affected by the Internet behavior.
6. Conclusion
(1) Based on the decision tree algorithm, this paper proposes a model for student network behavior analysis, which combines 32 characteristics of network behavior; trains and predicts the average academic performance of 3199 students of Beihang Class 2013 in four years, and constantly revises the model. Based on the model, the important behavioral characteristics of students are found and studied, and some universal conclusions are obtained.

Figure 4. Relevance analysis of the number of bytes on the Internet from Monday to Friday and the ranking of scores

Figure 5. Relevance analysis results of total net fee consumption and achievement ranking
Figure 6. Relevance analysis of the daily average earliest online logon time and performance ranking

Figure 7. Relevance analysis of the number of stay-up nights and achievement ranking

Figure 8. Relevance analysis of average online bytes and performance ranking

(2) The study found that the number of bytes on the Internet from Monday to Friday, the total consumption of Internet fees, the number of staying up late, the earliest daily login time and the
average number of bytes on the Internet every day are 5 most important factor affecting students' academic performance. Therefore, this paper analyses the correlation between these 5 characteristics and academic performance one by one.

(3) Analysis shows that students' online behavior has a profound impact on students' academic performance. In order to maintain a good learning state and ensure better learning results, students must strictly limit the time spent on the Internet in idle time, effectively control the time spent on the Internet, and ensure that their daily sleep is not affected by the online behavior.

References

[1] DENG H (2017). The analytics of college students’ behavior based on machine learning and big data technology. Beijing: Beijing University of Posts and Telecommunications. pp. 1-2.

[2] DONG Z N (2018). Improving the accuracy of ideological and political work in colleges by using big data of students. Practical Research -- Ideological and Theoretical Education, 4:108-111.

[3] ZHOU Z H (2016). In: Machine learning. Tsinghua University Press, Beijing. pp. 73-78.

[4] Trevor H, Robert T, Jerome F (2016). In: The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Second Edition. Springer, Berlin. pp. 24-26.

[5] Friedman J H (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics. pp. 1189-1232.

[6] Friedman J H (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4): 367-378.

[7] He X, Pan J, Jin O, et al (2014). Practical Lessons from Predicting Clicks on Ads at Facebook. In: Eighth International Workshop on Data Mining for Online Advertising. ACM. pp. 1-9.

[8] Yuan T T, Chen Z, Mathieson M (2011). Predicting eBay listing conversion. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM. pp. 1335-1336.

[9] LI Z J, MO J L (2013). An improved concept lattice-based data mining algorithm. Journal of Chongqing Normal University (Natural Science), 2: 92-95.

[10] Didrik N (2016). Tree Boosting With XGBoost. Why Does XGBoost Win "Every" Machine Learning Competition?. Trondheim: Norwegian University of Science and Technology. pp. 11-12.