A Self-Control Analysis Method of University Students Based on Campus Big Data

Liang Ge*, Jia Yu, Qing Zhou, Yanyun She and Zhongwei Yuan
College of Computer Science, Chongqing University, Chongqing, 400044, China
*Corresponding author’s e-mail: Geliang@cqu.edu.cn

Abstract. Nowadays, the analysis of university students is the main concern of educational institutions. In this paper, we investigate the problem of analysing students’ self-control ability. Our model focuses on the process of feature extraction, and then classification algorithms are applied. We propose the concept of behaviour chain to model the behavior of student. On this basis, we propose a spatial periodic feature extraction method based on the longest common sub-sequence(LCS). And we extract some statistics features from consumption data, access control data and others. Our model SCA(Self-Control Analysis) performs with the precision of 0.72, the recall of 0.78 and the F1-score of 0.74. It means that it’s possible to analyse student’ self-control ability with campus big data.

1. Introduction

With the rapid development of Educational Data Mining (EDM)[1], the analysis of student behavior data combined with data mining methods has become a trend. Studies in many fields such as education and sociology have found that students’ behavioral patterns have objective reflections on their performance, emotional state and mental health.

Universities have established the campus-card system, and have accumulated a large amount of structured and unstructured data. Most of the behavioral information of students is basically recorded in the campus-card system. Information related to students’ life is collected in the campus-card system, which mainly includes students’ consumption data, students’ access control data and so on.

Many scholars use campus big data to predict the development trend of students achievement. Mohan et al. [2] used 26 students’ campus Wi-Fi data for nearly 16 weeks to reveal students’ life patterns and lifestyles through Wi-Fi mobile data, and used clustering method to study the relationship between Wi-Fi data and student achievement.

Big data on campus can also be used to analyze and summarize habits and preferences of students. Dong Ding et al. [3] proposed a K-means clustering analysis method based on density partitioning, and they proposed a set of student behavior description index system and student behavior subdivision model.

Big data on campus can also be used to evaluate the social life status of college students. Yu Liu et al. [4] proposed a method to describe social relationships, and use campus-card data based on students’ daily behaviour to predict students’ social frequency.

At present, the traditional questionnaire is often used to measure students’ self-control ability [5]. However, questionnaires have some obvious disadvantages. First of all, it cannot ensure the honesty of participants. Beside, compared with an automatic approach, the questionnaire is much less efficiency. Last but not least, the questionnaire assesses students’ self-control ability from a specific point in time,
so that its difficult to follow the changes in students’ self-control ability over time. Therefore, we propose an automatic approach based on classification to find how students’ self-control ability is.

In this paper, we investigate the problem of analysing students’ self-control ability base on campus big data. We assume that each students self-control ability is associated with multiple facets of his/her daily life, including consumption behavior, internet usage behavior and access control of dormitory and so on. We propose the concept of behaviour chain to model the behavior of student. On this basis, we propose a spatial periodic feature extraction method based on the longest common subsequence(LCS). Periodic similarity measure indicates the degree of regularity of students’ periodic behavior characteristics in a certain period of time.

2. Statement and work overview
Most of the behavioral information of students is basically recorded in the campus-card system. Information related to students’ life is collected in the campus-card system, which mainly includes students’ consumption data, students’ access control data, students Internet service usage data and so on. Therefore, analysing university students self-control ability requires to disaggregate those records into different features(detailed in Section 3).

Formally, given a set of university students \( U = \{u_1, u_2, ..., u_n\} \) and two types of students self-control ability, the goal of our problem is to classify students self-control ability according to Self-Control-Scale(SCS), i.e. \( Y = (y_i)_{i=1}^{N} = f(X) \), \( y_i \in \{0,1\} \) denotes the self-control ability of student \( i \), where \( y_i = 0 \) denotes student is self-control ability is bad. \( X_n \subset X(= (x_i)_{n=1}^{N}) \) is the feature vector extracted from students’ behavior.

![Figure 1. The framework of our model.](image)

Figure 1 shows the framework of our system. we investigate the students behavior from three perspectives: consumption data, internet usage within campus, and access control data. And we also concerned with performance data and student basic information. We will detail these steps in the following sections.

3. Feature Extraction

3.1. Participants and Materials
In total 161 students from Chongqing University took part in the experiment. These students are asked to complete the Self-Control Scale(SCS) [5]. With a higher the self-control score, student is better in self-control ability. We conduct experiments on the datasets of these students.

The datasets are collected from March to July in 2018. Our data sets include the records of consumption data, internet usage and access control data. Besides, the performance and basic information of students were collected. To protect students privacy, the student ID were encrypted.
3.2. Definitions
Along with each swipe, a record is generated and stored in the database, which contains the trigger
time, location, and related attributes. We treat a record as a unit data, defined as an event.

Definition 1. An event. \( e = (t, l, m) \) of student \( u \) includes a time stamp \( t \) of the deal, the location \( l \),
and corresponding deal money \( m \). We use \( E_u \) as the event set of student \( u \).

For a particular scenario, there is a mapping function \( f(.) \) that defines a uniquely deterministic
behavior label for the event based on the activity time, location category, and transaction type.

Definition 2. Behavior Chain. The set of student behaviors sorted by time over a period of time, \( B = \{(t_1, l_1, b_1), (t_2, l_2, b_2), ..., (t_n, l_n, b_n)\} \). Where: \( l_i \) represents the ith position; \( T_i \) represents the time
of occurrence, \( t_{i,1} \leq t_i; N \) represents the number of events in the sequence, \( 1 \leq i \leq n \).

3.3. Periodic Similarity measure
We propose to apply the Longest Common Subsequence (LCS) method[6] to the sequence of
behaviors in each period of the students, fully extract the spatio-temporal information in the data, and
analyze the behavior patterns of the students by combining the spatio-temporal characteristics of the
behavior data. We propose an effective behavioral sequence matching method. All the students' behavior
sequences during the period time are matched by two pairs, and the average similarity of the
longest common subsequences between all the behavior sequences is calculated.

Given \( A = \{ a_{i1}, a_{i2}, ..., a_{in} \} \) and \( B = \{ b_{i1}, b_{i2}, ..., b_{jm} \} \) which represent for behaviour
sequences of two weeks, where \( n(1 \leq i \leq n) \) represents the length of the behavior sequence \( A \), \( m(1 \leq j \leq m) \)
represents the length of the behavior sequence \( B \), and \( D[i, j] \) represents the length of the longest common subsequence between sequence \( A \) and \( B \). We can obtain \( D[i,j] \) defined as in equation (1).

\[
D[i, j] = \begin{cases} 
0 & (i = 0 \ or \ 1) \\
D[i-1, j-1] & (i > 0, j > 0, l_{ai} = 1_{bj}) \\
\max\{D[i, j-1], D[i-1, j]\} & (i > 0, j > 0, l_{ai} \neq 1_{bj}) 
\end{cases}
\]

(1)

After obtaining the longest common sub-sequence length between the behavior sequences in any
two weeks, we tried to calculate the periodic similarity \( PSim(A, B) \). The frequency of the common
subsequence is used as the measure. The larger the value is, the more the same behaviors in the same
time and place between the behavior sequences, the higher the similarity is.

Assuming that the number of weeks is \( N \), the similarity of the student's behavioral pattern during the
\( N \) weeks is Score, and the formula is as follows:

\[
PSim(A, B) = \frac{D[n, m]}{n + m} \quad (2)
\]

\[
Score = \frac{1}{N} \sum_{A} \sum_{B=A} PSim(A, B) \quad (3)
\]

First, we classified and summarized the campus card data according to students' student id, and
extracted behavior sequences during this period. Then traverse behavior sequences in a week of each
student. If there are only one week’s sequence for a students, we assigned the periodic similarity
measure of the student to 0. For other occasions, weekly behavior sequences are paired to calculate the
periodic similarity based on LCS. The average of these value serves as the user's final periodic
similarity score.

3.4. Statistics Features
The experimental data include the records of consumption data, internet usage data and access control
data. Through observation of these data, it is found that these data are messy and difficult to
understand. In order to better understand the data in preparation for further analysis, the data was preprocessed separately.

1) Consumption Features: we first filter the consumption records of non-canteens. The standard meal time is defined according to the canteen opening hours and class schedule: breakfast time(6:30-09:30); lunch time(11:00-13:00); dinner time(17:00-19:00). The consumption data in the cafeteria is divided into breakfast, lunch and dinner consumption records according to the meal time. After pre-processing and statistical analysis of the data, we obtained the following features:

Table 1. Consumption features.

| Consumption Type       | Features               |
|------------------------|------------------------|
| Breakfast              | Average time           |
| Lunch                  | Amount                 |
| Dinner                 | Frequency              |
|                        | Entropy                |
|                        | Number of locations    |
| Eating at a non-standard time | Probability          |
| Late-night consumption | Frequency              |

2) Access Control Features: Dormitory access control data is generated when students enter and leave the dormitory. We analyze the access control data of the dormitory and understands the rules of students entering and leaving the dormitory. The analysis of the access control data of the dormitory is mainly to analyze the time points of the students entering and leaving the dormitory, focusing on the two time intervals of early departure and late return. The record before 8 o’clock is defined as the early departure time, and the record after 23 o’clock is defined as the late return time which is the dormitory requirement. After pre-processing and statistical analysis, we obtained the following features in table 2:

Table 2. Access control features.

| Consumption Type       | Features               |
|------------------------|------------------------|
| Late return            | Frequency of weekdays  |
| Early departures       | Frequency of weekend   |
| Entering dormitory     | Average time           |
| Leaving dormitory      | Standard deviation     |

4. Analyse self-control ability

We apply XGBoost[7] to analyse students’ self-control ability. XGBoost is one of the boosting algorithms. The idea of Boosting algorithm is to integrate many weak classifiers to form a strong classifier. We optimally adapt the parameters in a XGBoost model based on ten-fold cross validation to obtain a robust and accurate model. The objective function of XGBoost is:

\[
\text{Obj}(\phi) = L(y, f(x)) + \sum \Omega(f_m)
\]

\(L(y, f(x))\) represents the loss function and \(\Omega(f_m)\) is the regularization item indicating the complexity of the model.

One of the advantage of XGBoost is its support for parallelism. In the learning process of a tree, features need to be sorted by the loss function to determine the optimal splitting point. To achieve an optimal performance of XGBoost for the self-control ability analysis, we need to correctly set the parameters in XGBoost.

In general, XGBoost has the following parameters required to be optimized by cross-validation. 1) Learning rate, 2) Number of subtrees, 3) Gamma, 4) L1 regularization and L2 regularization weights, 5) Maximum depth of a tree, 6) Minimum weight sum of leaf node sample.
To show the effectiveness of our model, we compare our model against the following algorithms:

- **Support Vector Machine (SVM):** SVM is a generalized linear classifier that classifies data in a supervised learning manner [8].
- **k-Nearest Neighbor (KNN):** If a sample has most of the k most similar samples in the feature space belongs to a certain category, then the sample also belongs to this category [9].
- **Random Forests (RF):** Random Forests is a classifier that contains multiple decision trees [10].

| Method | Precision | Recall | F1-Score |
|--------|-----------|--------|----------|
| SVM    | 0.71      | 0.62   | 0.64     |
| KNN    | 0.66      | 0.61   | 0.63     |
| RF     | 0.71      | 0.70   | 0.70     |
| SCA    | 0.72      | 0.78   | 0.74     |

We use 10-fold crossvalidation based on grid search to select parameters. We divide the datasets into a training set(70%) and a testing set(30%) randomly. Then we evaluate with several general metrics like precision, recall and F1-Score. Table 3 shows the results of performance comparison, in which we compare our model SCA(Self-Control Analysis) with other three methods. We can see that SCA provided the best values in these three metrics. It means that it’s possible to analyse student’s self-control ability with our model.

5. Conclusion

In this paper, we investigate the problem of analysing students’ self-control ability. Our model focus on the process of feature extraction, and then classification algorithms are applied. We propose the concept of behaviour chain to model the behavior of student. On this basis, we propose a spatial periodic feature extraction method based on the longest common sub-sequence(LCS). Periodic similarity measure indicates the degree of regularity of students’ periodic behavior characteristics in a certain period of time. And we extract some statistics features from consumption data, access control data and others. As we can see in Table III, SCA performed with the precision of 0.72, the recall of 0.78 and the F1-score of 0.74. It means that it’s possible to analyse student’s self-control ability with campus big data, which include consumption data, Internet usage data and so on.

To our knowledge, no existing research was carried out in the field of self-control ability analysis base on campus big data. So it’s a new research direction with promising future.

References

[1] Romero, C., & Ventura, S. (2010) Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

[2] Mohan, M. G. M., Augustin, S. K., Roshni, V. S. K. (2015) A BigData approach for classification and prediction of student result using MapReduce. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS). IEEE.

[3] Ding, D., Li, J., Wang, H., Liang, Z. (2017) Student Behavior Clustering Method Based on Campus Big Data. 2017 13th International Conference on Computational Intelligence and Security (CIS). IEEE Computer Society.

[4] Liu, Y., Hu, M., Lu, X. (2016) Social Frequency Analysis of University Students via Digital Campus Cards. International Conference on Intelligent Human-machine Systems & Cybernetics. IEEE.

[5] Tangney, J. P., Baumeister, R. F., Boone, A. L. (2004) High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. Journal of Personality, 72(2), 271-324.

[6] Hirschberg, Daniel, S. (1977) Algorithms for the longest common subsequence problem. Journal of the ACM, 24(4), 664-675.
[7] Chen, T and Guestrin, C. (2016) XGBoost: A scalable tree boosting system, in Proc. ACM 22nd SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 785-794.
[8] Vapnik, Vn. (2003) Statistical Learning Theory. Annals of the Institute of Statistical Mathematics 55.2:371-389.
[9] Jain, A. K., Dubes, R. C. (1988) Algorithms for clustering data. Technometrics 32.2:227-229.
[10] Breiman, L. (2001) Random forests. Machine Learning, 45(1), 5-32.