Fuzzy Systems and Data Mining VIII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220376

A New Student Behavior Analysis Method Based on K-Means Algorithm and Consumption Data of Campus Smart Card

Jinglong ZUO[a, 1] and Marifel Grace C. KUMMER[b]
[a] *College of Electronic Information Engineer, Guangdong University of Petrochemical Technology, Maoming, China*
[b] *School of Information Technology and Engineering and Graduate School, St.Paul University, Philippines*

Abstract. In order to provide decision support for students' education management, a new student behavior analysis method based on the K-means algorithm and consumption data of campus smart card is proposed in this paper. An optimized Apriori algorithm is used to analyze the relationship between consumption behavior and academic performance. This method effectively provides management decision support for student managers, and improves management efficiency. This should improve the process of intelligent management.

Keywords. Campus smart card data, behavior analysis, correlation analysis, machine learning

1. Introduction

A campus smart card system is a control system that integrates identity recognition, access control, fund settlement, official management, and other services. Especially in colleges and universities, campus smart card can replace all kinds of identity certificates and consumption cards, and are widely used in campus management, student attendance, venue registration, identity recognition, living consumption, and other aspects. The management efficiency has been greatly improved.

The wide application of campus smart card has produced a large number of flow data which record the traces of card use and reflect the users' behavior and habits to a certain extent. However, this massive amount of campus smart card data has not been fully utilized. Most colleges and universities only provide simple query functions for campus smart card applications and rarely discover the value behind the data. How to effectively use this data, mine the hidden patterns and values through big data analysis, and achieve efficient intelligent management and high integration of information has important guiding significance for the educational reform of colleges and universities.

Take a student's daily life as an example. Starting from getting up in the morning and going out of the dormitory through the channel machine, his/her food consumption, class attendance, and going to the library for study will successively generate many

[1] Corresponding Author: Jinglong ZUO, College of Electronic Information Engineer, Guangdong University of Petrochemical Technology, Guangdong, China; E-mail: oklong@gdupt.edu.cn.

different campus smart card data. These data do not seem to be related to each other, but through data mining technology, we can find an appropriate behavior segmentation model to explore students' daily living habits and consumption rules. Combining this with academic performance, we can use machine learning to analyze the relationship between students' behavior and performance, and find important influencing factors that affect students' learning performance, so as to guide students to improve their behavior habits and improve their academic performance.

Therefore, it is necessary to analyze and study the student behavior data to help the university managers understand students more comprehensively and determine the behavior development trends, which are important for the improvement of management efficiency and the quality of education and teaching in colleges and universities.

2. Related work

2.1. Mining Campus Smart Card Data in Colleges and Universities

With the rapid development of educational informatization, many researchers began to study big data approaches and how these apply to campus smart card data records of students on campus.

Xie et al. believe that in combination with the methods and technologies of educational technology management research, and based on campus campus smart card data, they can explore influences to improve the support and service for school education management [1]. Wang used campus card data and analyzed the application of data mining in a campus card system [2]. Yu considered the analysis of students' all-in-one card consumption data [3]. Fu et al. analyzed the characteristics of campus IC (Integrated Circuit) card consumption data and used data mining and statistical analysis methods to discover consumption behavior [4]. Han et al. proposed a variety of data mining methods and developed a new solution for mining using campus data [5]. Chen applied mathematical methods to the statistical analysis of a campus all-in-one card management system. They obtained relevant information such as the dining peak in the canteen, meal times on holidays, average consumption for different cardholders in daily life, etc., and has certain guiding significance for the analysis of all-in-one card data [6]. Ji et al. mined and analyzed the consumption data generated by campus cards, classified the students' consumption habits and characteristics, and presented the results visually so as to provide assistance for the school's catering managers and improve the management and service level [7].

In summary, campus smart card data mining is closely related to users' application scenarios and user group attributes. In this area, there has been much research on consumption data mining. However, different universities have different management requirements for students, and some parameters of the consumption model are difficult to obtain accurately in the actual analysis. Therefore, the accuracy of the model cannot be guaranteed. Because there are many differences in the actual management of dormitories in colleges and universities, there is not much research that analyzes the traffic data of dormitories and mines rules on student living.

2.2. Correlation Analysis of Learning Performance

With the explosive growth of data, data analysis has become a hot topic. People use data

analysis technology to analyze and predict the behaviors of college students, which has changed traditional thinking in this area. Intelligent application systems in colleges and universities have accumulated a large amount of campus smart card data, which is closely related to students' study and life. It has become a trend to mine and analyze these data to understand the influencing factors of learning performance.

Grivokostopoulou et al. proposed a method to analyze student learning, extract semantic rules that can be used to predict final performance in the course, analyze test performance in the semester, predict final performance in the course, provide everyone with their own learning methods, and improve the learning performance [8]. Asif et al. used data mining methods to study and predict student academic performance in the four years of university, and gave timely early warnings to students with poor performance by studying their behavior [9]. Sweeney et al. identified students with academic problems using K-means clustering and established a model to predict student performance by using multiple linear regression [10]. Mujkic et al. proposed to use the correlation between students' disciplines to predict their academic performance. They selected 361 students and the scores for 33 courses of these students, set the learning status of each student through labels, and quantified the scores of each subject in the range 6-10. The research results show that students with good academic performance in some specific related subjects are more likely to achieve good results in other subjects [11]. Taking the campus card transaction data as the research focus, Li adopted an improved association rule algorithm based on clustering and found that water consumption was related to the gender of students [12].

In summary, the influencing factors of student learning performance have a very important relationship with student behavior habits. There is a certain relationship between course performance, work and rest behavior, and academic performance. At present, there are few studies on the relationship between the rules that govern student living, consumption, and learning performance through campus smart card data. Exploring the influencing factors of academic performance is still a useful research direction.

3. Research Design

3.1. Architectural Design

The purpose of this study is to use a machine learning algorithm to analyze and study the behavior patterns of students based on campus smart card data, and through multidimensional analysis and aggregation analysis of data, we can accurately understand the situation of students and provide strong data support for colleges and universities. The architectural design is divided into four steps, as shown in Figure 1.

Figure 1. Architectural design.

The data sources of the system mainly come from two data sets: one is the campus smart card consumption data from the smart card management center, and the other is the definition standard of the consumption quota of poor students from the student work department. By analyzing students' consumption records and consumption times, we can analyze the implicit rules of students' consumption behavior.

3.2. Key Technical Details

3.2.1. Student Behavior Analysis Based on Data Mining

First, the student behavior is subdivided and reasonably classified. Since the consumption data is multidimensional, it is necessary to reduce the dimensionality of the data. Assume that the data set is $X = \{X_1, X_2, \cdots, X_n\}$, where $X_i = \{X_{i1}, X_{i2}, \cdots, X_{im}\}$, and the data set X constitutes an $n \times m$ order matrix. The dimensionality is reduced according to the following steps:

Step 1: Normalize the data and map the data to the interval $[-1, 1]$;

Step 2: Calculate the correlation coefficient matrix R of X;

Step 3: Find the characteristic root of R and the corresponding characteristic vector;

Step 4: Calculate the principal component. The calculation of the principal component is achieved as follows:

$$y_{ij} = a_{1j}x_{i1} + a_{1j}x_{i1} + \cdots + a_{1j}x_{i1} \ (j = 1,2,\cdots,m) \quad (1)$$

Step 5: Calculate the principal component cumulative contribution rate SCR.

Second, use a Tyson polygon to select the initial number of clusters and cluster center points; these steps are followed to determine all the core objects:

Step 1: Create a simple rectangle and stack the clustered data sets in the two-dimensional closed rectangular area;

Step 2: Use the incremental method to construct the Tyson polygon of all spatial plane areas;

Step 3: Find out the intersection of all cells u at the edge of the Tyson polygon and adjacent cells V with the rectangular edge. The initial unit is W;

Step 4: Observe the intersection of the cell and the rectangle. If one of the intersections is on a different edge, add an intersection in the vertex set.

Third, calculate the density coefficient. The target Tyson polygon V is obtained by the incremental method, the area of the unit area VI is calculated, and the density coefficient is calculated. Filter out the data points that do not meet the conditions until all the density coefficients are greater than the density threshold.

Finally, the centroid of each cell is calculated and the dataset is divided again until the convergence condition is satisfied.

The dimension reduction pseudocode of principal component analysis is given in Algorithm 1 and the pseudocode of the K-means algorithm is presented in Algorithm 2.

Algorithm 1: Dimension reduction by principal component analysis

Input: Training sample set $D = x^{(1)}, x^{(2)}, \cdots, x^{(m)}$, Dimension d of low dimensional space

Output: Dimension-reduced dataset X'

1. Centralize all samples (de mean operation): $x_j^{(i)} \leftarrow x_j^{(i)} - \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$
2. Calculate the covariance matrix of the sample XX^T

3. Eigenvalue decomposition of covariance matrix XX^T
4. Take the eigenvectors $w_1, w_2, \cdots, w_d$ corresponding to the largest d eigenvalues
5. Multiply the original sample matrix by the projection matrix: $X \cdot W$ is the data set X' after dimension reduction

Algorithm 2: Pseudocode of the K-means algorithm

Input: Training sample set $D = x^{(1)}, x^{(2)}, \cdots, x^{(m)}$, Dimension d of low dimensional space

Output: Clustering results

1. Dimension reduction by principal component analysis and output of dimension-reduced dataset X'
2. Incremental method to generate Tyson polygon satisfying that all density coefficients are greater than the density system threshold
3. Calculate the centroid of each cell and divide the dataset again
4. Calculated value is used to update centroid
5. Judge whether the convergence conditions are met. If yes, it ends. If not, it returns to step 3

3.2.2. Using The Apriori Algorithm to Analyze the Correlation Between Consumption Behavior and Academic Performance

Students' consumption behavior is closely related to their living habits and learning attitudes, and thus to their academic performance. The Apriori algorithm is used to analyze the correlations. The algorithmic flow is as follows:

First, preprocess the data set, and divide the transaction database $\mathrm{D} = \{D_1, D_2, \cdots, D_n\}$ based on the division criteria, where $D = D_1 \cup D_2 \cup \cdots \cup D_n$ and $D_1 \cap D_2 \cap \cdots \cap D_n = \emptyset$.

Second, scan the transaction database D (with m transactions and n transaction items) to construct the Boolean transaction matrix M:

$$M = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{2}$$

$$a_{ij} = \begin{cases} 1, a_{ij} \in T_i \\ 0, a_{ij} \notin T_i \end{cases} \quad i = 1,2,\cdots,m; j = 1,2,\cdots,n \tag{3}$$

where the ith record in the transaction dataset T is represented by T_i. m transactions and n transaction items correspond to rows and columns, respectively, in the Boolean transaction matrix M.

Third, calculate the transaction data centralized weights $\mathrm{W}(I_j)$ and $\overline{\mathrm{W}}(T_i)$:

$$\mathrm{W}(I_j) = \frac{m}{l} \tag{4}$$

$$\overline{\mathrm{W}}(T_i) = \sum_{j=1}^{n \& I_j \in T_i} \frac{\mathrm{W}(I_j)}{|T_i|} \tag{5}$$

where m is the number of transactions, and l is the number of times I_j occurs in transaction set T.

Fourth, generate candidate item sets.

Fifth, calculate support (S).

Finally, generate frequent itemsets to judge whether the generated frequent itemsets are empty. If it is empty, all P_i will be merged to generate the final frequent itemset; if it is not empty, connect automatically and return to step 5.

The pseudocode of the Apriori algorithm is given in Algorithm 3.

Algorithm 3: Pseudocode of the Apriori algorithm

Input: Transaction database D

Output: Frequent itemset P

1. Define the minimum support and divide the transaction database D into n partitions D_i $(i = 1, 2, \dots, n)$ based on the division criteria
2. Scan D_i to generate the Boolean transaction matrix M_i $(i = 1, 2, \dots, n)$
3. Calculate the transaction data centralized weights $\mathrm{W}(I_j)$ and $\overline{\mathrm{W}}(T_i)$
4. Generate candidate item sets
5. Calculate support (S)
6. Generate frequent itemsets and judge to generate the final frequent itemsets

4. Experiment Results

Data set Description: multi-source data such as student consumption, classroom attendance, channel records and academic achievements generated by smart card with student information as a data set, which provides an important component for student behavior analysis and prediction of life consumption law.

4.1. Study Effect Mining and Influence Factor Analysis

In this paper, students who visited the library 5 days or more a week were regarded as data samples with normal learning positive behavior habits, and students who visited the library 2 to 4 times a week were regarded as regular learning habits; Students who visit the library occasionally or never are regarded as inactive students' learning habits. According to the K-means clustering analysis method (learning habits "positive", "general" and "not positive") on the above basis, cluster the visit data of the campus smart card library of our students. The results are shown in Figure 2.

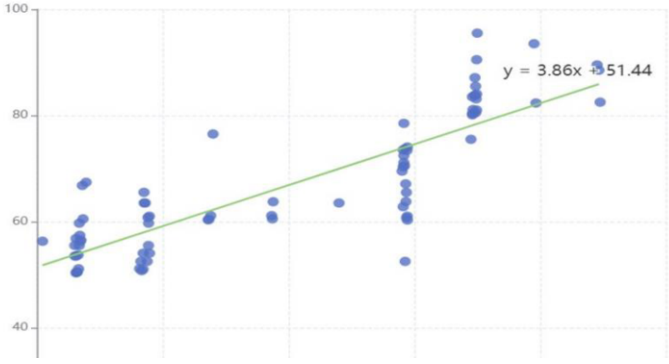

Figure 2. System Regression Diagram of Library Factor Learning Effect.

Figure 2 shows the student behavior analysis, association rule mining and cluster analysis have been widely used. There is a certain correlation between academic achievement and going in and out of the library. Students' eating on time also implies good work and rest habits, which can also have a certain correlation with the learning effect. The implicit rules and knowledge between data are reflected through some relationship between them. Association rule mining is to extract the implicit rules and knowledge by mining the potential unobvious relationship network between objects.

In this paper, the students who eat before 7:50 a.m. in the school are regarded as the consumption data samples with normal consumption behavior habits (because the first class in the school is at 8:00 a.m., and it is easy to be late for eating after 7:50). The students who eat later than 7:50 may not have the habit of getting up early, which is regarded as irregular work, rest and life; The students who go to class without breakfast may often skip classes and have meals, so they are also regarded as irregular data samples of life consumption. Considering that most students have the habit of not eating breakfast on Saturday and Sunday, this study excluded the breakfast consumption behavior and habit data on weekends every week. According to the K-means cluster analysis method (consumption behavior habits "regular" and "irregular") on the basis of the above, cluster the campus smart card consumption data of our students.

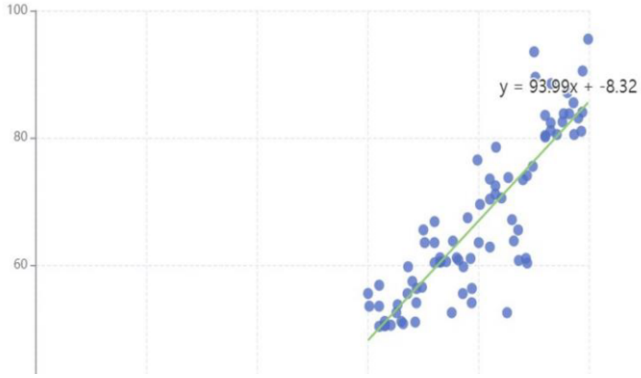

Figure 3. System Regression Diagram of Consumption Factor Learning Effect.

Based on the Figure 3, data mining analysis method can also be used in the correlation analysis between smart card consumption log, classroom check-in log and grades. The data sources are respectively from the smart card classroom check-in log record, consumption log record, library access log record and academic achievement record of the educational administration system. Firstly, through data preprocessing, the higher quality raw data are selected from the massive raw data, and then data cleaning, data integration and data specification are carried out respectively to realize data deep mining and correlation analysis.

The data shows that students are used to getting up early every day, eating and going to class on time. If they can avoid skipping classes, absenteeism or even being late or leaving early in their regular work and rest and life, correct their learning attitude and cultivate good self-consciousness, initiative and positive consciousness will help to improve their academic performance to a certain extent.

4.2. Life Rules Mining and Influence Factor Analysis

The performance of students in the dormitory can be excavated by the time and times

they pass through the dormitory channel. For example, students with regular work and rest can go out to study and eat according to the normal work and rest time, and the time to return to the dormitory is also within the scope required by the school. In China's colleges and universities, special attention is paid to understanding whether students return to the dormitory at the specified time in the evening. We hope to master which students have the habit of returning late (or even not), and try to find out which colleges have more students returning late and which periods of time students return late. Therefore, this kind of life law is the focus of this study.

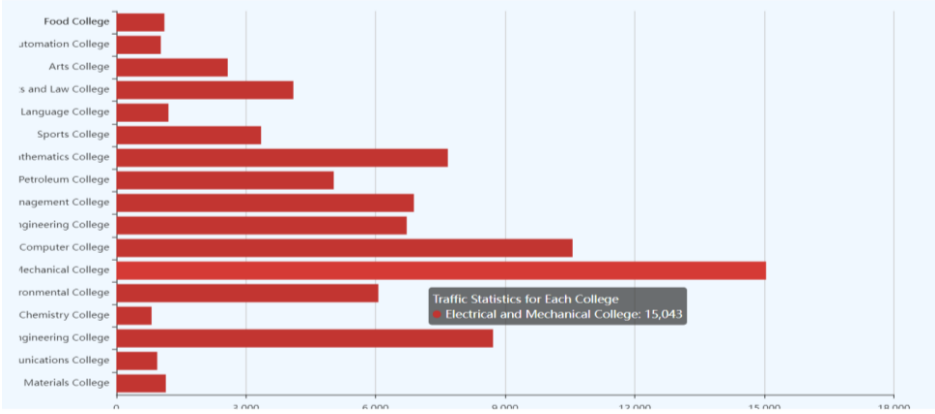

Figure 4. Statistics of Students' Life Habits.

In the analysis of student behavior model, as shown in figure 4, the mining objects of life rules mainly come from the data sources of students' access to dormitories, libraries and dining consumption in canteens. They contain the habits and trends of students in life, learning and consumption. Of course, the consumption rules is also a kind of life rules, which has been discussed above. Therefore, in the data mining of life rules, it mainly focuses on the behavior patterns of students entering and leaving the dormitory.

The data also shows that excellent students and students with learning difficulties also have significant differences in library entry behavior in each semester. The number of library entry of excellent students is significantly higher than that of poor students. Therefore, the number of card use, regular breakfast and library entry are significant factors to predict students' learning effect.

5. Conclusions

On the basis of campus smart card data, this study uses the K-means algorithm and the Apriori algorithm to mine student consumption and life behavior, analyzes the correlation between student behavior and academic performance, designs and implements a student behavior characteristic analysis system, visually presents the statistical analysis results in the form of charts, effectively provides management decision support for student managers, and improves management efficiency. This should improve the process of intelligent management.

Acknowledgements

Key Realm R&D Program of Guangdong Province (2021B0707010003). Guangdong Basic and Applied Basic Research Foundation (2021A1515012252, 2020A1515010727, 2022A1515012022). Key Field Special Project of Department of Education of Guangdong Province (2020ZDZX3053). Key Realm R&D Program of Guangdong Province (2021B0707010003). Guangdong Province Ordinary Universities Characteristic Innovation Project (2019KTSCX108). Maoming Science and Technology Project (210429094551175, mmkj2020008, mmkj2020033). Intelligent Exploration and Production Team for Complex Oil and Gas Reservoirs (20190315).

References

[1] Xie YR, Li KD. Fundamentals of educational technology research methods (2nd Edition). Higher Education Press, 2017.
[2] Wang Ya. Application of data mining technology in Colleges and Universities under the background of big data -- Taking campus card system as an example [J]. Journal of central China Normal University, 2017: 9-12.
[3] Yu XX, Wang Y, Wang B, Zhao XM, Yao HL. Research and application of cluster analysis optimization of students' consumption behavior. Appl Comp Syst, 2017, 26(006), 232-237.
[4] Fu Y, Jing M, Cheng D. Analysis on data of the Campus IC Card based on Hadoop. Wireless Internet Technology, 2016(15): 77-79.
[5] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques Third Edition. The Morgan Kaufmann Series in Data Management Systems, 2011, 5(4), 83-124.
[6] Chen F. Analysis and data mining of dining consumption behavior of college users based on campus one card system. China Education Informatization: Higher Vocational Education, 2014(5): 47-49.
[7] Ji Z, Su B, Li J, et al. Analysis of college students' consumption characteristics based on campus card data. 2015 (the fourth) National Undergraduate statistical modeling competition paper, 2015.
[8] Grivokostopoulou F, Perikos I, Hatzilygeroudis I. Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance [C]//2014 IEEE international conference on teaching, assessment and learning for engineering (TALE). IEEE, 2014: 488-494.
[9] Asif R, Merceron A, Ali S A, et al. Analyzing undergraduate students' performance using educational data mining. Computers & Education, 2017, 113: 177-194.
[10] Sweeney M, Lester J, Rangwala H. Next-term student grade prediction //2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015: 970-975.
[11] Mujkic A, Boban I, Dugandzic I, et al. Decision tree-based students' grades analysis //2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2014: 133-136.
[12] Li N. Research and implementation of campus all in one card decision support system. Tongji University, 2008.