Analysis and prediction of college students’ consumption behavior

Wenming Hou¹, Xiaoqiang Di¹, Xingxu Zhang¹, Yong Ma¹ and Huamin Yang¹

¹School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, Jilin, 130000, China
²Information Center of Changchun University of Science and Technology, Changchun University of Science and Technology, Changchun, Jilin, 130000, China
*Corresponding author’s e-mail: dixiaoqiang@cust.edu.cn

Abstract. Campus data analysis is becoming increasingly important. In this work, based on the students’ consumption behavior in the canteen, we study the correlation between the students’ consumption expenditure and China’s national price level, as well as the relationship between the number of merchants to whom students make payments and students’ own economic conditions. Specifically, the students’ consumption expenditure increases with the price level, and the male consumption is significantly higher than that of the female; in addition, the better the students’ economic condition, the more canteen merchants that students choose. Finally, we propose a model for predicting the student consumption group, which can divide the students into high, moderate and low consumption groups. Experiment results show that the model can be used as an effective tool for predicting the consumption groups of students.

1. Introduction
The vigorous development of digital campus construction promotes the research of university big data. Campus one-card system, as the core of digital campus construction [12], generates a large number of data with regard to consumption, internet usage, campus trajectory, et al. Among them, the consumption data reflects the students’ behavior to a great degree. Therefore, analyzing the students’ consumption data can not only provide reasonable guidance and management to the students, but also help relevant departments make more scientific teaching plans.

Generally, consumption data of colleges is characterized by large number of users, large daily transaction volume and small amount [8]. These data contain abundant knowledge, which can be effectively counted, analyzed and mined to find some patterns and rules of students’ consumption.

In this work, we focused on analyzing the consumption data of a university collected through the campus one-card system [5]. The purpose of this paper is to study the consumption data of college students and find out their consumption behavior and habits. At the same time, we hope to put forward a model that can predict students’ consumption groups, so as to provide convenience for university management. We first analyze the students’ consumption behavior, and use the clustering algorithm to divide the students into high, moderate and low consumption groups according to their average consumption and total consumption; then, based on the analysis of the consumption groups, we find that the consumption expenditure of college students is related to the price level of the country, and at the same time find the positive relationship between the students level of consumption and the number of canteen merchants; finally, a prediction model of student consumption group is proposed.
Experiment results show that our model is proved to be very meaningful for improving university management and students’ satisfaction.

2. Related work
The campus card system stores a huge amount of data and has received extensive attention from researchers. Based on the information collected from the digital campus card, Fan [1] explored the relationship between students’ achievements and their learning partners. Lian [6] studied the book-borrowing history and students’ academic performance history, and found that student’s academic achievement can be predicted by records of book borrowing. Jiang and You [7,10] utilized the consumption data of canteen and supermarket from the smart card system and extracted some related features. They used an improved clustering method to analyze students’ consumption behavior in the canteen and supermarket, and their method provided reference for university management. Wang [9] analyzed the campus behavior of college students using consumption records, and found out the changing trends and gender difference in students’ campus behavior. Guan [2] carried out research on the relationship between the students’ behavior and their financial conditions, and found that the consumption data collected from campus one-card system can help discover students in financial hardship. Through the methods of statistics, clustering and correlation analysis, the students’ behaviors in the university were found out through data mining, and these methods all achieved good results.

This paper firstly obtains the classification of student consumer groups through clustering, and then uses neural network model to predict future student consumption groups.

3. Data sets
The data set used in this paper is provided by the authority of a university located in Jilin province of P.R. China, which includes the consuming records generated by students’ campus cards when making payments to merchants in the campus, as well as students’ basic information. The data set includes 14111 students, of which 8771 are males and 5340 are females. The time range is one semester, from 2016.09 to 2016.12, and there are 2930383 records in total. Ideally, the original data collected can be used directly in the research. However, in real world, there is always noise, along with the problem of data loss and in consistency in the original data. Therefore, data cleaning is especially important.

(1) Data desensitization: Since there is always some sensitive information related to personal privacy in the data, such as: name, ID number, bank account number, email address, mobile phone number, etc., the reliability of these sensitive private data has to be protected during data processing to ensure not to infringe the privacy[3]. Thus, before processing the data, some desensitization rules are used to deform the corresponding sensitive information to protect privacy.

(2) Data deletion: It is well known that whenever a student swipes the smart card in the canteen, it will be recorded in the system, but some students seldom use the card during the selected time range in this research. With the popularity of electronic payment means, many of them choose to pay meals by their smartphones, or order take-outs rather than go to the canteen. To carry out better research on their consumption behavior, we set up the upper threshold and lower threshold of the times that students swipe the card in the canteen per month. Concretely, the lower threshold filters out the too small times, since these numbers are not useful to analyse the consumption behavior. To our knowledge, we believe that each student should swipe the card for meal in the canteen for at least 60 times a month, so that their consumption behavior can be better represented. On the other hand, the upper threshold, which is set to be 120 times a month, filters out the too large times, since these large numbers may influence the experimental results, and may be caused by such situations as students’ friends helping them swipe the card for meal. For this reason, we only keep those students who consume 60 to 120 times a month in the canteen. The pre-analysis of the distribution of our data suggests that this part of students make up for the majority, accounting for nearly 85 percent of the entire dataset.
4. Experiment

After the data processing, we can get the number of students who swipe the card 60 to 120 times a month. The number of students in these four months is 3818, 3023, 3093 and 3867 respectively, which constitutes the Data Set 1. However, not all students swipe the card between 60 and 120 times a month for all of the four months. For example, some students may have consumption records in month 1, but not in month 2 to 4, or some may have records in month 1 and 2, but not in month 3 and 4. Therefore, after analyzing the original data, we got 1147 students who swiped the card between 60 and 120 times a month for all the considered four months. That is to say, from Month 1 to Month 4, these 1147 students have consumption records in each month. These students constitute the Data Set 2. Data set 2 constitutes less students than data set 1. Two reasons may account for this: Firstly, this university does not demand that students must pay meals through swiping the card. Secondly, some students choose to pay meals by their smartphones, or order take-outs rather than go to the canteen. This indicates that the management departments of this university should further standardize the use of campus smart card, and only in this way can they study students' behavior better to provide improved service for them. This research is mainly based on data set 2, because these students have four months' data, we can study the changing trend of students' consumption.

| Table 1. The data set. |
|------------------------|
| Month | 1 | 2 | 3 | 4 |
| Data Set 1 | 3818 | 3023 | 3093 | 3867 |
| Data Set 2 | 1147 | 1147 | 1147 | 1147 |

4.1. Student consumption and price level

Campus consumption is one of the activities in students' daily life, and contains information about the activities of individual students and groups. Information transferred between individuals can affect the behavior of student groups. Based on the consumption data of canteen, it is more direct and practical to study the characteristics of student group consumption from the macroscopic view. Specifically, the level of consumption of college students can be measured by total consumption amount and average consumption amount [12]. Take one month as an example, the former represents the student's comprehensive consumption level, i.e., the total amount of money that a student spends in a month, and the latter represents the average daily expenditure of students, i.e. the average amount of money that a student spend in a month. In general, the higher the total consumption amount and average consumption amount, the stronger the students’ consumption capacity.

For Data Set 2, the gender is distinguished. Concretely, 927 students are male, and 220 are female. We calculated the average amount and total amount of money that each student spends on meals per month. Note that male students are far more than female ones. This is because the university selected for this research mainly consists of majors related to science and technology, and in China, males account for a larger proportion in these majors. Then, the K-means algorithm is utilized to analyze students' consumption groups, with the silhouette coefficient [13] as the evaluation standard for the algorithm. The silhouette coefficient can be calculated as follows:

$$ S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (1) $$

For each object in a cluster, the average distance to all other subjects in this cluster is calculated, which is denoted by $a(i)$. In addition, for each object and any cluster not including the object, the average distance to all of the objects in that cluster is calculated, which is denoted by $b(i)$. The reasonable range of the coefficient is [-1,1], and the larger value indicates a better clustering performance. Through calculation, when $K = 3$, the silhouette coefficient is largest. Therefore, we classify students' consumption groups into three categories: low consumption, moderate consumption and high consumption. Figure 1(a) shows an example, which is the clustering result of the first month for male. Here, green represents the low class, red is the moderate class and blue is the high class.

In summary, we got 8 clustering results for male and female of four months, and then statistically analyzed the trends of the three consumption classes obtained above. As shown in Figure 2, the trend
of average consumption and total consumption for four months is illustrated, where the first month of the X-axis starts from 2016.09. Two interesting patterns are observed herein: (1) the average consumption and total consumption of the four months generally shows an increasing trend, indicating that the economic cost of students has increased slowly in the four months; (2) the average consumption and total consumption of males in the three categories of consumer groups are significantly higher than those of females, and even the average consumption of the moderate consumption classes of male is higher than the females of the high consumption classes. Also, there is a clear stratification for both male and female in total consumption.

![Figure 1](image1)

**Figure 1.** (a)The clustering result of the first month for male, and (b)the growth rate of CPI.

In addition, Consumer Price Index (CPI) is often considered as an important reference indicator of market economic activities, reflecting the overall price level [4]. We obtained the growth rate of the CPI index from China’s official statistical office, of which the time range is aligned with the data set. As shown in Figure 1(b), in the first three months, the CPI index gradually increased, corresponding to the increase in the overall price level. However, in the fourth month, the growth rate of CPI was lower than before, but still large, which means that the price level is still growing. The increasing price level has led to the increasing economic expenditure of students, which is shown in Figure 2, where the total consumption and average consumption of students have increased gradually in four months. Based on the above analysis, we think that students’ consumption expenditure is related to the national price level, i.e., the higher the price level, the more money students may spend.

![Figure 2](image2)

**Figure 2.** (a)The average consumption, and (b)the total consumption. Males and females are divided into three consumption groups, high, moderate and low, respectively.

4.2. **Analysis of the number of canteen merchants**

The university in this study has three canteens, of which the total number of merchants is 73. Here, the merchants refer to the vendors from whom students can buy the food they like when having breakfast, lunch or dinner. From Data Set 2, the number of distinct merchants that made payments by male and female each month is analyzed and counted. Then they are clustered using the K-means method. After that, we are able to get the clustering results of the number of canteen merchants that students consume each month. The results are shown in Table 2.
We divide the number of merchants from whom the students buy meals per month into three categories, namely low, moderate and high, finding that the results of the four-month classification of male and female students are roughly the same. As shown in Table 2, the low category has a merchant number of less than 15 or 16, while the high category has a number of more than 20 or 21. If the number of merchants is between low and high category, the corresponding students belong to the moderate category. For all the consumption groups clustered in the previous section, we analyze the trends of the number of merchants from whom students buy meals in the three types of consumption groups each month for males and females. The results are shown in Figure 3, where blue represents the number of low merchants, red is the number of moderate merchants, and gray the number of high merchants.

![Figure 3](image)

Figure 3. Trends in the number of merchants made payments by (a) males and (b) females. Each month, the males and females are divided into three categories: high, moderate and low consumption.

In the low consumption group in each month, low merchants and moderate merchants account for the vast majority, but the number of high merchants is small. For the moderate consumption group, the number of low merchants is reduced, while the number of moderate and high merchants are increased. Besides, in the high consumption group, the amount of low merchant group is less, while moderate and high merchant groups occupy a large part. Meanwhile, the proportion of high merchant group is the highest in the high consumer group. Therefore, there is a positive relationship between the number of merchants consumed by students and students’ own economic condition. The higher the students’ economic level, the more canteen merchants that students choose. This provides a supporting basis for schools to assess students’ economic status, thereby further provides humanistic help, which can help students in financial hardship to complete their studies smoothly.

4.3. Student consumption group prediction model

This study selects BP neural network as a prediction model for dividing student consumption groups. BP neural network is a kind of one-way and multi-layer feedforward network trained through error back propagation. In our study, students are divided into high, moderate and low groups according to their total consumption and average consumption. Therefore, the input is composed of total consumption and average consumption. The output is the consumption group classification.
consumption and average consumption, and the output is the corresponding consumption groups of students. The number of neurons in the input layer is 2, and in the output layer is 1. Choosing the appropriate number of hidden layer neurons has important implications for the accuracy of learning and the speed of training. Based on empirical equation [14], we can determine the number of hidden layer neurons.

$$h = \sqrt{m+n} + a \quad (2)$$

Among them, $h$ is the number of hidden layer nodes, $m$ is the number of input layer nodes, $n$ is the number of output layer nodes, and $a$ is the adjustment constant between 1 and 10. Therefore, we choose $h = 3$ to $12$ to carry out experiments. Next, we use the first three months of Data Set 2 as the training set for the input of the model, and the fourth month as the test set to test the model accuracy. In addition, since the optimal number of hidden neurons will affect the generalization ability of the neural network model [11], it is important to select the appropriate structure of the hidden layer. Therefore, we carried out experiments for several times, in which the hidden layer was set to be composed of 3 to 12 neurons respectively. The result is shown in Figure 4.

In Figure 4, the red line represents the Data Set 1, while the blue one stands for the Data Set 2. For data set 2, the accuracy of the model prediction fluctuates with the increase of number of neurons in the hidden layer, but the overall prediction accuracy is relatively high, with the average accuracy rate reaching 86.65%, under which circumstances the optimal number of hidden neurons is 6. At the same time, Data Set 2 is a four-month data collection of 1,147 students. To expand our data set, we use the data of a total of 9929 students in the first three months of Data Set 1 as the input of the training set, and use the 3867 students in the fourth month as the test set to verify the validity of the prediction model. The red line in Figure 4 characterizes the trend of the accuracy of Data Set 1. It can be seen that with the increase of the data amount, the accuracy of model prediction has been significantly improved, with the average accuracy rate reaching 89.43%, and at this time the optimal number of hidden neurons is 7.

According to the above analysis, the accuracy of the student consumption group prediction model based on the BP neural network is desirable, and the larger the data volume of the training set, the higher the accuracy. Therefore, this model can be applied to the prediction of the groups of students’ consumptions, which provides assistance for the management department to better serve students, and also provides certain reference value for the student management of colleges and universities.

5. Discussion and conclusion
This research conforms to the development of university informatization, and analyzes the canteen consumption data of students using clustering and neural network technology. From our work we can clearly realize that the consumption in campus are obviously stratified, the consumption of males at all levels is higher than that of females, and students’ consumption expenditure is related to China’s national price level. At the same time, there are more canteen merchants selected by high consumption
groups when eating in the canteen. Finally, experiment results have shown that we can accurately predict the students’ consumption groups through BP Neural networks.

However, there is still room for improvement in this study. By mining student online records and achievement records, the prediction results can be much better. This will be the focus of the next step of our work.

Acknowledgments
This work was supported by the Higher Education Research Project of Jilin Province (Grant No. ZD18027).

References
[1] Fan S, Li P, Liu T, et al. (2015) Population behavior analysis of Chinese university students via digital campus cards[C]/2015 IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City. pp. 72-77.
[2] Guan, C., Lu, X., Li, X., Chen, E., Zhou, W., & Xiong, H. (2015) Discovery of college students in financial hardship[C]/2015 IEEE International Conference on Data Mining. Atlantic City. pp. 141-150.
[3] Heo J, Lim H, Yun S B, et al. (2019) Descriptive and Predictive Modeling of Student Achievement, Satisfaction, and Mental Health for Data-Driven Smart Connected Campus Life Service[C]/Proceedings of the 9th International Conference on Learning Analytics & Knowledge. Tempe. pp. 531-538.
[4] Hongyan, Y., Deyin, C., et al. (2011) Dynamic correlation analysis between china’s consumer price index and food price index. Price: Theory & Practice 2, 52–55.
[5] LI S, SUN M, WANG H, et al. (2007) Practice and consideration of campus ic card system's construction[J]. Journal of Xiamen University (Natural Science), 2: 162-165.
[6] Lian D, Ye Y, Zhu W, et al. (2016) Mutual reinforcement of academic performance prediction and library book recommendation[C]/2016 IEEE 16th International Conference on Data Mining (ICDM). Barcelona. pp. 1023-1028.
[7] Nan J, Weisheng X. (2015) Student consumption and study behavior analysis based on the data of the campus card system[J]. Microcomputer Applications, 31(2): 35-38.
[8] Rushan C, Jie Y, Education S O, et al. (2014) The Influences of MOOC on Higher Education Teaching Mode[J]. Meitan Higher Education. 2014,0(5).
[9] Wang B, Deng K, Wei W, et al. (2018) Full Cycle Campus Life of College Students: A Big Data Case in China[C]/ IEEE International Conference on Big Data & Smart Computing. Shanghai. pp. 507-512.
[10] Xiang-Ru, Y., Ye, W., Shu, Y., Bin, W., Xin-Miao, Z., Hai-Lun, Y. (2017) Research and application of clustering analysis optimization for students’ consumer behaviors. Computer Systems & Applications 26(6), 232–237.
[11] Yan Z, Rong-Zhao Z. (2007) Security Research on IC System in Digital Campus[J]. Information Security and Communications Privacy. 2007(9).
[12] Yanqiang L. (2006) On College Students' Consumption Capacity, Sports Consciousness and Sports Consumption [J]. Journal of Guangdong Institute of Business Administration,4: 021.
[13] H. Shen, Z. Wang, C. Gao, J. Qin, F. Yso, and W. Xu, (2008) "Determining the number of BP neural network hidden layer units," Journal of Tianjin University of Technology, vol. 24, pp. 13-15.