Applying data mining techniques with data of campus card system

Yufan Li*, Huifu Zhang and Shangli Liu

School of Computer Science and Engineering, Hunan University of Science and Technology, 411201 Xiangtan, China

*E-mail: 494958939@qq.com

Abstract. In recent years, campus card system (CCS) has been widely used in universities. At the same time, the scale of the data stored in the database of the CCS is also increasing rapidly. Therefore, it would be better to adopt data mining techniques to reveal the status, causes and trends of students’ consumption data. In this research, the K-Means clustering algorithm is applied to the students’ consumption behavior. The Apriori algorithm is used to study the students’ consumption behavior and performance. The results not only improve the efficiency of student work but also help to build smart campus for our university.

1. Introduction

Data mining is the process which reveals relationships, trends and patterns by analyzing large amounts of data [1]. It has been widely used in banking, marketing, insurance, education and other domains. And nowadays, college students usually have various consumption types, such as dining, shopping, checking out books, purchasing water, surfing on the net [2], it will generate magnanimous data. Therefore, data mining is an effective approach, which can provide a great help in finding hidden relationships behind the data of CCS.

Jiang et al. [3] analyzed all the consumer records of masters, it is concluded that the boys overall consumption is higher than the girls, but there is no difference between their session and major. Xue et al. [4] used cross tabulations and Kohonen neural network to the data of shop consumption and they found that the types of consumer have a greater influence on consumption, while gender is nearly have no impact on it. By using OLAP(On-line Analytical Processing), Zhang [5] proposed an effective data mining model of the campus card system to classify the students’ consumption data and it helps the university leadership to make appropriate decisions.

In this research, we study the CCS data of Hunan University of Science and Technology in China. The paper is organized as follows: First, data mining process is briefly described in Section 2. Then the results and discussion are shown in Section 3. Finally, conclusions and future work are outlined.

2. Data mining process

2.1. Data set

The original data of CCS is stored in Oracle database. The users include undergraduates, graduate students, teachers, etc. Since freshmen came to school for a short time, they have not yet formed a stable habit of consumption, while senior students often have their internships for a long time out of school. Consequently, we collected the consumption data from session 2015 to 2016. Initially size of
the data is 380723 and it contains 960 students. Table 1 shows the structure of table stored in transaction database:

| SNO   | MERCNAME | TRANAMT | EFFECTDATA |
|-------|----------|---------|------------|
| 15xxxxxx01 | 4   | 1900    | 2018-3-27 11:03:34 |
| 15xxxxxx01 | 2   | 550     | 2018-6-5 11:55:28  |
| ...   | ...     | ...     | ...         |
| 16xxxxxx30 | 1   | 200     | 2018-4-10 7:13:42 |
| 16xxxxxx30 | 6   | 900     | 2018-4-11 12:08:33 |

In table 1, SNO (student number) is the primary key, MERCHAME is the ID of canteen, TRANAMT means the expense of every record, and EFFECTDATA is the transaction time. For example, the first record means a student with a student number is 15xxxxxx01 and expenses ¥19 at No.4 canteen on March 27th, 2018, at 11:03:34.

2.1.1. Analysis of student consumption records
We created a new field named DT which means the results that we divided transaction time into time section by an hour. The time before 7am is value T1 in DT, the time between 7:00 and 8:00 is T2 and the time after 10:00 is T17. Figure 1 shows the student records of consumption based on time.

![Figure 1. Based on time.](image)

In figure 1, we can see that the amount of student consumption records during 7:00 to 8:00, 11:00 to 13:00, 17:00 to 19:00 are lager than others, and T6 (11:00 to 12:00) has the biggest amount of records while T10 (15:00 to 16:00) is the smallest in all time sections. And we argue that if a student who has breakfast before 8:00 or dinner between 17:00 to 19:00, then he or she is consider to have a regular consumption behavior.
Figure 2 shows the relationship between amount of records and canteen’s location. We can see that No.3, No.4, No.5 have more records than other canteen. Those canteens which have low amount of records may far away from residence halls, but they can attract students by offering discount or handing out leaflets.

2.2. Data selection and transformation
In this step, only those fields were selected which were necessary for applying data mining algorithms. We chose a few derived variables, while some variables were extracted from the database. All the response variables which were derived from the Oracle database will be elaborated in section 3.

2.3. Data mining methodologies
A. K-Means Clustering Algorithm:
Clustering is a method which can divide data into classes of similar objects. K-Means clustering algorithm is based on a simple iterative method which is partitioning a dataset into a user-specified number of clusters, k [6]. The following figure 3 shows the steps of K-Means clustering algorithm.

\[
\text{Input:} \text{Number of desired clusters } K \\
D = \{x|1, 2...N\}
\]

Methods:
1. Select k data objects from dataset D as initial centers randomly.
2. Repeat;
3. Assign each data object to its closest centroid after calculate the distance between each data object x.
4. Recalculate new mean for each cluster.
5. Until no change in the center of cluster.

\[
\text{Output:} \text{A set of k clusters}
\]
B. Apriori Algorithm:

Apriori algorithm is one of the most popular association rule mining algorithm to find frequent item set from large data sets and hence to get association rules \( A \rightarrow B \) contains which meet the following conditions [7]:

\[
\text{Support}(A \rightarrow B) = \frac{P(AB)}{P(A)} \geq \text{Min} \_ \text{sup} \tag{1}
\]

\[
\text{Confidence}(A \rightarrow B) = \frac{P(B|A)}{P(A)} \geq \text{Min} \_ \text{conf} \tag{2}
\]

The Min_Sup and Min_Conf are threshold values defined by users.

3. Results and discussion

To determine the best number of desired K for K-Means algorithm we use the measure called Silhouette Coefficient. The silhouette coefficient, relative to mean intra-cluster distance \( a \) and mean nearest-cluster distance \( b \) for each sample, is defined as:

\[
\frac{b - a}{\max(a,b)} \tag{3}
\]

Table 2. Student related variables.

| Variables | Description | Possible values |
|-----------|-------------|-----------------|
| AC        | Amount of Consumption Per Month | {High:,Medium,Low} |
| SC        | Number of Times for Swiping Cards Per Month | {High:,Medium,Low} |
| RB        | Number of Times for Regular Breakfast Per Month | {High:,Medium,Low} |
| RD        | Number of Times for Regular Dinner Per Month | {High:,Medium,Low} |
| GCS       | Grades of Compulsory Subject | {Poor:,Average,Good,Outstanding} |
| GES       | Grades of Elective Subject | {Poor:,Average,Good,Outstanding} |
| LC        | Library Consumption | {Yes,No} |
| RC        | Retaking Course | {Yes,No} |

Table 2 explains all the related variables in this research.

In table 3, the comparison of silhouette coefficient with K is illustrated. It shows that clustering effect is the best when K=2. Table 4 shows the results after the application of K-Means algorithm.

Table 3. Comparison of silhouette coefficient with K.

| K | Silhouette coefficient |
|---|------------------------|
| 2 | 0.549                  |
| 3 | 0.511                  |
| 4 | 0.505                  |
| 5 | 0.508                  |
| 6 | 0.490                  |
| 7 | 0.479                  |

Table 4. Clustering results.

| Cluster center | AC    | SC    | RB    | RD    | Students |
|----------------|-------|-------|-------|-------|----------|
| I              | 369.002 | 141.241 | 14.148 | 47.634 | 434      |
| II             | 160.852 | 64.414 | 6.835  | 22.105 | 526      |

Students in cluster I consume more in canteen and appear to have a regular eating time. While in cluster II, students exhibit a low interest in canteen, which means they may have a take-away for their meal. Since the students in cluster 2 have a higher proportion of complete 960 students, the canteen should update their menu or offer discount to attract students to consume.
Table 5. Results of Apriori algorithm.

| Rule | Conf |
|------|------|
| 1. GCS=Outstanding, LC=No, RD=Medium, AC=Medium ==> SC=Medium | 1 |
| 2. AC=Low, RB=Low, RC=Yes ==> RD=Low | 0.915 |
| 3. SC=High, RD=High, GES=Outstanding ==> AC=High | 0.843 |
| 4. SC=High, RC=No ==> GES=Outstanding | 0.841 |
| 5. RD=Medium, AC=Medium, SC=Medium, GES=Outstanding ==> GCS=Good | 0.800 |

Table 5 shows the results after adopting Apriori algorithm. From the results of the Apriori algorithm as shown in row numbered four of table 5, we can clearly see that students who have high SC and without retaking courses appear to have outstanding GES. Moreover, from rows numbered one, two, three and five, we find that students with good performance usually have a regular consumption behavior, habit of checking out books and daily routine.

4. Conclusion

This paper aims to study the data of CCS and analyzes it with data mining techniques like K-Means clustering algorithm and Apriori association rule algorithm. The results we got can help relevant departments in decision-making and improve the efficiency of education management. Data is the core part in CCS and also has great potentialities. In the future, we hope that data mining techniques such as classification algorithms can be widespread used in CCS and thus we could build an effective model to support the construction and application on smart campus.

References

[1] J. W. Han, M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc. (2005)
[2] W. C. Lv. Design of Campus Smart Card System. Applied Mechanics and Materials, 347-50, 3915-3918 (2013)
[3] T. J. Jiang, J. M. Cao, S. Dan, X. L. Yang. Analysis and Data Mining of Students Consumption Behavior Based on a Campus Card System. International Conference on Smart City and Systems Engineering (2017)
[4] L. M. Xue, W. X. Luan. Data Analysis and Consumption Model Research of E-card system. International Industrial Informatics and Computer Engineering Conference. (2015)
[5] Y. Zhang. The Research of Data Mining Technology in Campus Card Consumption System. International Conference on Electronics & Optoelectronics. (2011)
[6] X. D. Wu, V. Kumar, JR. Quinlan, et al. Top 10 algorithms in data mining. Knowledge & Information Systems, 14(1),1-37(2008)
[7] S. Rathee, A. Kashyap. Adaptive-Miner: an efficient distributed association rule mining algorithm on Spark. Journal of Big Data, 5(1), 1-17(2018)