Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 2071025, 18 pages
https://doi.org/10.1155/2020/2071025

Research Article

Analysis and Identification of Students with Financial Difficulties: A Behavioural Feature Perspective

Yong Ma,[1] **Xingxu Zhang**,[1] **Xiaoqiang Di**,[2,3] **Tao Ren**,[1] **Huamin Yang**,[1] **and Binbin Cai**[3]

[1]*School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China*
[2]*Jilin Province Key Laboratory of Network and Information Security, Changchun, China*
[3]*Information Center, Changchun University of Science and Technology, Changchun, China*

Correspondence should be addressed to Xiaoqiang Di; dixiaoqiang@cust.edu.cn

Received 3 April 2020; Revised 28 May 2020; Accepted 2 June 2020; Published 28 June 2020

Academic Editor: Rigoberto Medina

The identification of students with financial difficulties is one of the main problems in campus data research. Effective and timely identification not only provides convenience to campus administrators but also helps students who are really in financial hardship. The popular using of smart cards makes it possible to identify students with financial difficulties through big data. In this paper, we collect behavioural records from undergraduate students' smart cards and propose five features by which to associate with students' poverty level. Based on these features, we proposed the Apriori Balanced Algorithm (ABA) to mine the relationship of poverty level with students' daily behaviour. Association rules show that students' poverty level is most closely related to their academic performance, followed by consumption level, diligence level, and life regularity. Finally, we adopted the semisupervised *K*-means algorithm to more accurately find out students with financial difficulties. Tested by classical classification algorithms, our method has a higher identification rate, which is helpful for university administrators discover students in real financial hardship effectively.

1. Introduction

Nowadays, college students gradually become the main labour force in the society and have an important impact on the country's economic and social development [1]. In recent years, thanks to the rapid development of digital campus, college students' daily behaviours can be recorded in the campus smart-card system, so researchers are increasingly paying attention to the study of campus big data [2–5]. As a branch of campus behaviour research, finding students with financial difficulties can not only effectively help those who really need it but also provide school administrators with a solution to find them and to give some financial support. Therefore, it is imperative for educators to discover students with financial difficulties.

In China, the selection of students with financial difficulties mostly adopts the method of "students proclaiming + advisers assessing" [6], where the process of evaluating students' qualification is usually manually conducted concerning students' family background, daily expenditure, and academic performance. When there are a large number of students, this scheme can be time-consuming and tend to involve some subjective judgments as well. Fortunately, thanks to the rapid construction of smart campus, the student one-card system, also known as the smart-card system, has been designed to record students' behaviours of daily life. These behaviours include the consumption in the canteen, the Internet login records [7], the book-borrowing records [8], checkin records, and so on. The increasing amount of these data has provided opportunities for us to analyse students' behaviour through novel information technologies.

Several previous research studies have focused on the topic of students' behaviour analysis. Some studies pay attention to records from the smart-card system, using them to explore students' academic performance with daily behaviour [7]. These records also give a multiaspect display of their campus daily life, revealing the changing trend in their

learning career and showing different living habits of different genders [9]. Moreover, there has been growing interest underlining the importance of online education systems and online learning platforms [10, 11]. Learning records generated by these tools have revealed the dependencies among learning time, subject, activity type, activity complexity, and performance, which gives suggestions for behavioural changes to optimize learning experience. Besides, judging from students' learning modality, the trends and deficiencies in the use of LMS by students can easily be detected, which is beneficial to quickly grasp their learning status.

The above works prove the feasibility of data mining on students' behaviour to identify their behaviour patterns through the daily records generated from the smart cards. However, in terms of students' economic status, few studies have been conducted using campus behavioural data. Present studies can mainly be divided into two branches, namely, the prediction of students' financial hardship and the discovery of students' financial difficulties. The former one has been considered as a multilabel classification problem using features such as smart-card usage, Internet usage, and trajectories in campus [12]. However, only the pairwise correlation of students is studied, but not the correlation of poverty level and behaviour characteristics. The BP neural network was also utilized to construct a nonlinear mapping between the economic conditions of college students and the needy students identifying [13], but lacked fine-grained analysis of how different behaviour influenced students' economic status. The latter one has been studied through active learning [6], but such method requires the intervention of human knowledge. Although accuracy has been improved, human intervention can involve too much personal will. In addition, the correlation of different behaviours with financial hardship has not been analysed.

Hence, in this paper, we proposed the Apriori Balanced Algorithm (ABA) to explore the relationship between students with financial difficulties and their behaviours. In addition, a semisupervised *K*-means algorithm is established to identify students with financial difficulties to decrease human intervention. To be specific, we extract "consumption level," "GPA," "GPA_percentage," "life regularity," and "diligence" from the smart card to describe students' behaviours. Then, we applied the ABA to obtain the relationship of students' poverty level and their behaviour features. Finally, we adopted the semisupervised *K*-means algorithm to identify the financially difficult students.

Overall, the contributions of this paper are summarized as follows:

(1) Faced with complicated data exported from the smart-card system, we proposed five behavioural features, GPA, GPA_percentage, consumption level, life regularity, and diligence, which can better reflect the behavioural characteristics of students in financial hardship.

(2) Secondly, we proposed the Apriori Balanced Algorithm (ABA) based on the original Apriori algorithm by modifying the *Support* to *Balanced_support* with a balance factor C. After such modification, for items with small proportion in the dataset, the association rules containing such items will be more accurately found out. Therefore, it is useful for our task of mining the association between students' economic status and behavioural features, since financially difficult students only account for a small scale of the whole students. Test results on the Groceries dataset prove the adaptability of the ABA, and the relationship between the proportion of poor students and different behavioural features shows that the association rules we obtained are consistent with the ground truth.

(3) Thirdly, we proposed a method based on semisupervised *K*-means to identify students in financial hardship. Previous works have used methods such as active learning to discover financially difficult students, but this may involve too much personal will. Our method will effectively decrease human intervention without losing identification performance. Experiments on the dataset processed by our method through four classical classification models indicate a higher prediction performance.

2. Materials and Methods

2.1. Motivation. In this section, we will describe the motivation of our research in detail. The first is the motivation of proposing the Apriori Balanced Algorithm. In the traditional campus big data research, personal will may be involved in the experiment. In this paper, we expect to find the relationship between students with financial difficulties and other behavioral characteristics through data mining, so as to reduce such disadvantages. However, traditional algorithms for mining association rules, such as Apriori and FP-growth, are based on support and confidence. When used for mining rules containing small-scale items, the result may not reflect the truth hidden in the dataset. This phenomenon is found in our previous data mining that other papers may not pay attention to, so for this reason, we proposed the Apriori Balanced Algorithm (ABA).

The second is the motivation of identifying students in financial hardship. It has been found in previous data mining for the original poor student list provided by the university that for students labelled with financial difficulties, some of them have different behavioural features from others, while some of the students without financial difficulties have the same behavioral features as most students with economic difficulties. Based on such phenomenon, we think that in terms of behavioural characteristics, some students in the original poor student list provided by the university are not real financially hard students. Meanwhile, a small proportion of students are not labelled as "Poor" by the university but have the same behavioural characteristics as "Poor" students. These students are not accurately identified in the poor student list. Therefore, to solve the above problems, we proposed a method based on semisupervised *K*-means to relabel students in the poor student list according to their

behavioural characteristics. In this way, university administrators can more accurately identify students in financial hardship and provide targeted funding.

Figure 1 shows the basic work flow of our framework, which includes four major parts. The entire framework mainly focuses on identifying financially difficult students and finding out the hidden poor students. Firstly, behavioural characteristics of students including consumption level, academic performance, diligence, and life regularity are extracted from records of the smart-card system. Secondly, the Apriori Balanced Algorithm (ABA) is proposed and used to correlate the poverty level with other behavioural features, by which 2-item set and 3-item set consisting of students' behavioural feature labels are obtained. Thirdly, labelled data and unlabelled data are selected based on the predefined rule and are then input to the semisupervised learning algorithm to label the unlabelled data and build the new datasets. Finally, new datasets are used to train different models for prediction to verify the effectiveness of the framework.

2.2. Dataset. There are two datasets used in this paper. The first dataset is exported from the database management system provided by the Information Center of our university, which consists of three parts, with the time range from Sep.1st, 2013, to Jun.30th, 2014, including students' consumption records in the canteen, the GPA for the spring and autumn semesters, and the records of poor students' list. The students selected were enrolled in 2012 and 2013. Not all students have all of these three kinds of records, so after combining different data tables and removing error data, there are records of 6224 students remained for experiment. The data statistics are illustrated in Table 1.

The second dataset used in this paper is the Groceries dataset. This dataset is often used for association analysis by Apriori, FP-growth, and Eclat algorithms. The dataset is the real transaction records of a grocery store within a month. There are 9835 consumption records and 169 products. The data format of the Groceries dataset is shown in Table 2.

2.3. Experiment Tool. All the experiments were conducted by Python 3.6 on a 64-bit Windows 8 with 16 GB memory and 2.3 GHz CPU.

2.4. Feature Extraction. Traditional research usually obtains information about students' family situation by means of elaborate rules and regulations of the funding system [12]. However, there are also shortcomings. For example, to obtain financial support from the school, students may deliberately describe their family as financially difficult ones. In addition, dealing with case-study assessments manually put a lot of pressure on the staff [12]. Thanks to big data technology, researchers have been provided a fast, efficient, and accurate way to students' behaviour. In this paper, combining with the campus data, we identify students with financial difficulties by their behaviours. Initially, we propose four assumptions.

Assumption 1. Financially difficult students tend to consume less.

The most direct intuition of students in financial difficulties is that they are lack of money. In their normal campus life, they may consume less than others in lunch and in dinner, reflected by generally smaller consumption amount in the smart-card records. Therefore, we proposed the consumption level to describe students' consumption behaviour.

Assumption 2. Financially difficult students tend to perform better in academic activities.

Students with financial difficulties may have a deep understanding of their own situation, so they cherish learning opportunities more than others, and perform better in academic performance. In the smart-card records, the grades are generally high in all subjects, so we propose the academic performance level to describe students' learning behaviour.

Assumption 3. Financially difficult students are more irregular in life.

Students with financial difficulties may less self-disciplined in life. Lacking of money, they may not eat breakfast on time. Also, they may not attend classes on time every day due to part-time jobs. So, we put forward the life regularity to describe life behaviour of students.

Assumption 4. Financially difficult students may skip breakfast to save money.

A major challenge facing students in financial hardship is their limited available money. Consequently, they may skip breakfast to save as much money as possible. In the smart-card system of our university, a record is generated once a student swipes his/her student card on the card reader device, so each consumption is recorded with a timestamp every time he/she comes to the canteen for meals. Therefore, we regard the time of students' first meal as a rough reflection of their diligence level.

2.5. Consumption Level. The consumption data selected for this research are 2108250 records in total. The data format of consumption is shown in Table 3, in which *Stu_ID* shows the ID of each student (similarly hereinafter) and *Location* shows the place he/she buys food. In our university, food is served from different buttery hatches of four canteens, and *Canteen1, BH1* means the first buttery hatch in Canteen1. *Time* is the time he/she buys food, *Consum_amount* is the amount of money spent during this consumption, and *Card_balance* is the balance of his/her student card after this consumption. After the exploratory data analysis, we found that the transaction amount of each breakfast is extremely lower compared to lunch and dinner because porridge and pancakes are served at a low price during breakfast time. If breakfast is included in the consumption level statistics, some students cannot guarantee to eat it every day, so there will be mistakes in the classification. Therefore, only lunch

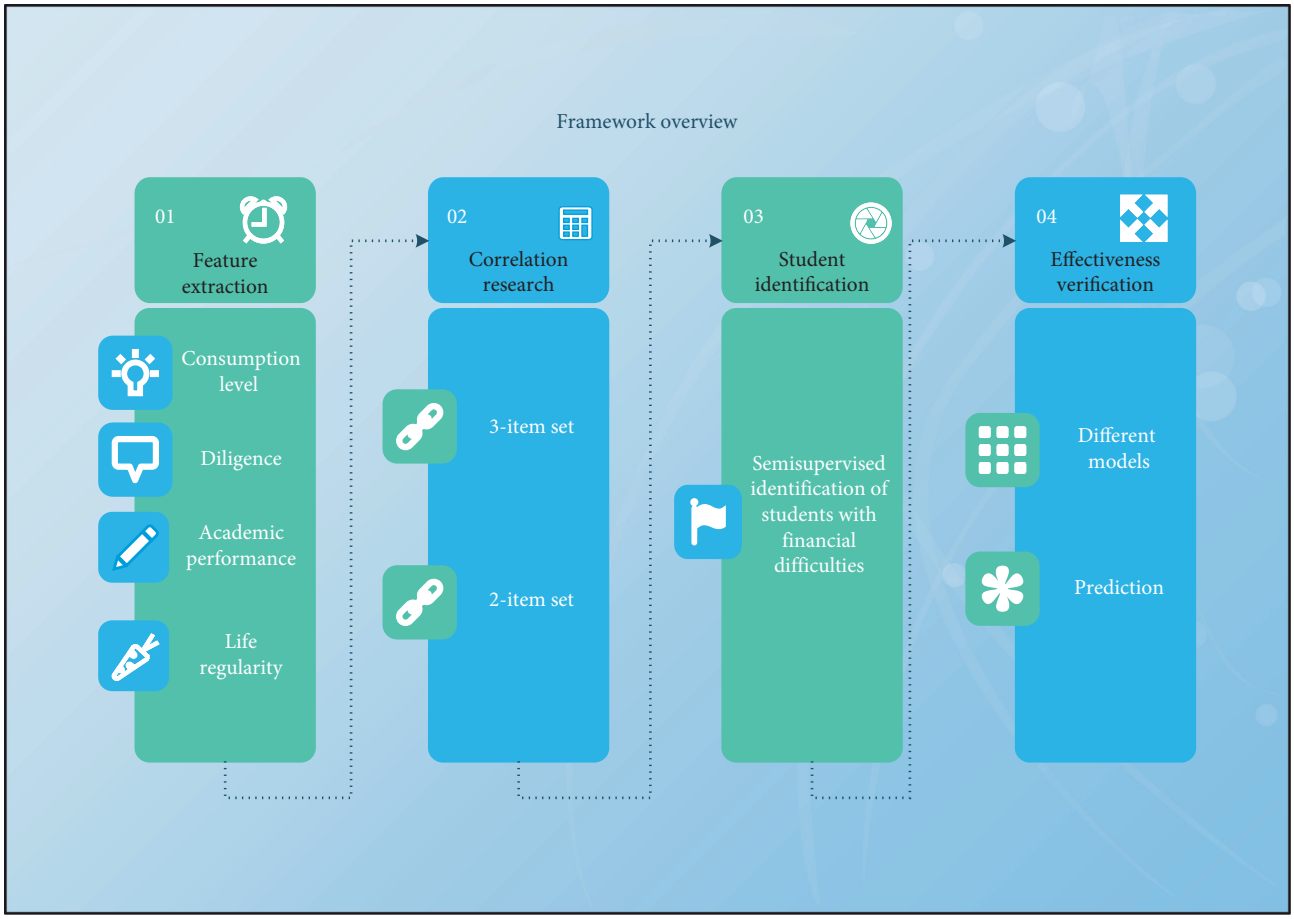

Figure 1: Framework overview.

Table 1: Basic statistics of our dataset.

Data type	Semester	
	Autumn	Spring
Number of students	3192	3032
Number of poor students	725	786
Number of consumptions	1,048,576	1,059,674
Number of grades	125,273	125,782

Table 2: The data format of the Groceries dataset.

ID	Records
1	{citrus fruit, semifinished bread, margarine, ready soups}
2	{tropical fruit, yogurt, coffee}
3	{whole milk}
4	{other vegetables, whole milk, condensed milk, long-life bakery product}

Table 3: Data format of consumption data.

Stu_ID	Location	Time	Consum_amount	Card_balance
2012001	Canteen1, BH1	2013/9/2 7:41	1.7	52.1
2012002	Canteen3, BH5	2013/9/2 8:45	4.5	80.1
2012003	Canteen4, BH4	2013/9/2 10:57	1.6	71.2

and dinner consumption is considered for mining consumption behaviour.

Along this line, we need to ensure that every student only eats lunch and dinner one time every day. First of all, we define the time intervals for different types of meals. According to the dining rules of our university, we set 11:00–13:00 as lunch time and 16:00–18:00 as dinner time. It is important to note that a student may swipe the card for food more than one time during each meal, for example, buying some snacks during lunch. So, we proposed lunch-time consumption (LTC) and dinner-time consumption (DTC), which, respectively, represent how much money a student spends during lunch time or dinner time. LTC is defined as formula 1, and DTC is defined similarly:

$$\mathrm{LTC} = \sum_{i=1}^{n} \mathrm{flow}_i, \tag{1}$$

where flow_i represents the i-th consumption record of the total n records between 11:00 and 13:00 of a day, so that LTC represents the total consumption amount during 11:00–13:00.

Next, it is a key problem to convert consumption records into indicators of consumption. Previous work [9] simply calculated the number of consumption records in the canteen in different time durations during one hour. However, in this way, students with more consumption records are more likely to be considered to spend more, while the ones with fewer records tend to be considered to spend less. To avoid this situation, we propose the average consumption and the consumption speed. Average consumption is defined as the average consumption amount during one LTC or DTC, denoted by *Avg_consum* in formula (2). Consumption speed is defined as the number of consumption times to spend up per 100 yuan, denoted by *Spd_consum* in formula (3):

$$\mathrm{Avg\_consum}_t = \frac{\mathrm{Total}_t}{\mathrm{Sum}_t}, \tag{2}$$

where Total_t represents the total consumption of student t within a semester, while Sum_t represents the total number of consumptions during one LTC or DTC:

$$\mathrm{Spd\_consum}_t = \frac{\mathrm{Num}_t * 100}{\mathrm{Total}_t}, \tag{3}$$

where Num_t represents the total record number of consumptions of student t in a semester and Total_t has the same definition as formula (2).

2.6. Poverty Level. We exported data of financially difficult students of the 2013–2014 academic year from the database, totalling 3400 items. The data format is given in Table 4, in which the *Semester* column indicates the corresponding semester of the record ("2013-20141" means the first semester, and "2013-20142" means the second semester). Besides, for each of the 3400 records, the *Financial_status* column indicates whether the corresponding student in this entry is in financial hardship. If so, the value will be "Poor," otherwise it will be "Not_poor." Such labels can be convenient for the subsequent processing.

2.7. Academic Performance

2.7.1. GPA. The academic performance data we selected in this research are composed of 251,055 records, which contain the scores of each course of each undergraduate student in two semesters. The data format of grades is shown in Table 5. In this table, all the students are grouped by *Stu_ID*, and for each student, each course he/she attended during the two semesters is recorded as one entry. The *Score* column is the score he/she obtained for that course, *Credit* is the credit assigned to that course, and *Course_type* indicates whether the course is compulsory or elective. Generally, in the student management system, a student's academic performance is measured by the GPA (grade points on average). In this research, we proposed a metric similar to the original GPA to measure students' academic performance. This metric is defined in the following formula:

$$\mathrm{grade}_{\mathrm{sum}} = \frac{\sum_{i=1}^{m} \mathrm{grade}_i * \mathrm{credit}_i}{\sum_{i=1}^{m} \mathrm{credit}_i}. \tag{4}$$

For each student s, grade_i denotes the score of a single course, credit_i is the credit for that course_i, and m is the number of courses in a specific semester. Through this formula, the score of per credit for each student is obtained, which can later be used to divide all students into two groups. Concretely, after obtaining the $\mathrm{grade}_{\mathrm{sum}}$ and sorting them descendingly, the top 50 percent of students are labelled as GPA_high, and remaining 50 percent are labelled as GPA_low. These two labels can be used as the features of students' academic performance.

2.7.2. GPA_Percentage. Although GPA is generally considered as a metric for evaluating students' academic performance, it is rather a coarse-grained measure. This is because in China, the difficulty of different subjects varies with majors. Courses of liberal arts majors tend to be given higher marks due to flexible answers of certain exam questions, while those of science and engineering majors are much harder to get A due to complex calculation, analysis, deduction, and reasoning. Despite the difference existing in different courses, students of the same major will face the exams of same subjects. Therefore, it is required to figure out the ranking of students within their respective majors. To this end, we propose GPA_percentage here, which is defined in the following formula:

$$\mathrm{GPA\_percentage}_i = \frac{n_i}{m_i}, \tag{5}$$

where n_i represents the GPA ranking of student i within the range of his/her major and m_i represents the total number of students in his/her major.

According to the criteria for evaluating personal scholarship in our university, students who rank top 20% in his/her major will win the "first-level scholarship" and "second-level scholarship," while those who rank between

Table 4: Data format of poor student list.

Stu_ID	Semester	Financial_status
2012001	2013-20141	Poor
2013004	2013-20141	Not_poor
2013060	2013-20142	Not_poor

Table 5: Data format of grade data.

Stu_ID	Semester	Course	Score	Credit	Course_type
2012004	2013-20141	English	79	4	Compulsory
2012004	2013-20142	Modern history of China	72	2	Compulsory
2013008	2013-20142	Intellectual property law	85	1	Elective

top 30% and top 50% will be awarded the "third-level scholarship." Therefore, students whose GPA_percentage is between 0 and 0.2 are labelled as *Gper_A*; those with GPA_percentage of 0.2 to 0.5 are labelled as *Gper_B*, and the rest are labelled as *Gper_C*.

2.8. Life Regularity. The regularity of students' behaviour can be expressed with the regularity of eating breakfast [7]. In order to describe the regularity of different students as much as possible, we regard a student's first record in the smart card as his/her first activity every day. So, we select 5: 00–11:00 as the time interval for the regularity of students' behaviour. Therefore, our processing steps are as follows.

Firstly, we divide the time intervals from 5:00 to 11:00 in the morning into 12 bins, each of which spans 30 minutes and is encoded from 1 to 12, respectively. Then, inspired by the concept of information entropy [14], we define a life entropy here to express the life regularity of students, which is calculated by the following formula:

$$\mathrm{LE}_{(x)} = -\sum_{i=0}^{12} p(x) * \log(p(x)), \tag{6}$$

where $p(x)$ represents the probability of each time interval. We know from the definition of entropy that LE is de facto, the distribution of an arbitrary student X's eating time in a semester. Therefore, the larger the LE is, the more scattered and irregular the breakfast eating period is, while the smaller the LE is, the more concentrated and regular the period is.

Next, a threshold value needs to be determined to label the regularity for different students according to LE. This can be considered as a problem of one-dimensional data clustering. Therefore, we sort LE first and then use the K-means clustering to obtain a threshold H. According to the threshold H, students can be divided into *Regular* or *Irregular*.

2.9. Diligence. It has been said that the first smart-card record in each day can be regarded as surrogates of students' bedtime [15]. Inspired by this, we use students' first smart-card record in each day as their first daily activity. Since the meal consumption in canteen accounts for a large majority of all consumption records, we calculated the time of first meal for each student. This is then used as a measure for students' diligence level. Specifically, we transformed the raw date-time format into Unix timestamps. After obtaining the time of first meal consumption for each student in each day, the diligence level can be calculated as follows:

$$\mathrm{Diligence} = \frac{\sum_{i=1}^{t} \mathrm{time}_i}{t}, \tag{7}$$

where t is the total number of days that student i has in consumption records and time_i is the time of his/her first meal. In this way, we obtained the average time of each student's first meal within one semester. Subsequently, we clustered all the students into two groups according to the diligence value, labelling those with smaller diligence as "Early" and larger diligence as "Late".

2.10. Apriori Balanced Algorithm. The Apriori algorithm is one of the most popular and widely used algorithms in both data mining and educational data [16]. Previous research has studied the utilization of Apriori on user behaviour prediction, for example, using Apriori for mining rules related with the study to provide a basis for optimizing educational decision [17] and mine the association rules of enrollment information to explore the factors affecting college enrollment [18]. However, the traditional Apriori algorithm may not be able to mine out the rules of items with small proportion. This problem is mainly caused by the different proportion of various labels in the datasets. People tend to accept rules with high support and high confidence. However, low proportion labels generate rules with low support, which may easily be ignored. In our datasets, various labels have different proportions. Therefore, Apriori is not suitable for mining the association rules hidden in students' poverty level and daily behaviour.

Based on the above problems, the Apriori Balanced Algorithm (ABA) is proposed.

Given a dataset D, N is the number of data in D, $L = \{l_1, l_2, \ldots, l_n\}$ is the set of different items in D, $P = \{p_1, p_2, \ldots, p_n\}$ is the proportion of different data items.

Let $U = \{l_i, l_j, \ldots, l_t\}$ be a rule of expectation. The support of U is defined as follows:

$$\mathrm{Support}_{(U)} = \frac{\mathrm{len}(B(U))}{N}, \tag{8}$$

where B is a subset of D, $B(U)$ represents the data items in D that contain U, and $\mathrm{len}(B(U))$ represents the number of data items in B. The confidence of U is defined as follows:

$$C_i = \mathrm{Confidence}_{(l_i \longrightarrow (U - l_i))} = \frac{\mathrm{len}(B(U))}{\mathrm{len}(B(l_i))} = \mathrm{Support}_{(U)} * \frac{1}{p_i}. \tag{9}$$

To make sure that the items in U are closely related to each other, compute $C_i * C_j, \cdots, C_t$ in the following formula:

$$C_i * C_j, \cdots, C_t = \left(\text{Support}_{(U)}\right)^m * \prod_{x=i}^{t}\left(\frac{1}{p_x}\right), \quad (10)$$

where x is the index of the terms in U and m is the number of items of U.

The reasonable range of p_x is [0, 1], so the range of $1/p_x$ is [1, +∞]. If normalized, its range will be [0, 1], which has the same distribution as $1 - p_x$. Therefore, $1/p_x$ is replaced with $1 - p_x$. Since we focus on how to deal with the imbalanced proportions of different behavioural labels, it has nothing to do with the calculation of *Support*, so here we just set $m = 1$. To adapt the support to different numbers of item sets, a balance factor C is defined as follows:

$$C_x = \frac{T}{2} * \frac{\text{Inum}(\max)}{\text{Inum}(\min)}, \quad (11)$$

where T is the number of the type of labels that belong to the same behaviour category as the item x. (For example, x is the item "Consumption Low," since the labels of consumption behaviour are "High," "Medium," and "Low," in such situation, $T = 3$.) $\text{Inum}(\max)$ represents the maximum number of samples of the labels in each behaviour category. Similarly, $\text{Inum}(\min)$ is the minimum number of samples of the labels in each behaviour category.

With the balance factor C, $Balanced\_support_{(U)}$ is defined as follows:

$$\text{Balanced\_support}_{(U)} = \text{Support}_{(U)} * \prod_{x=i}^{t}\left(1 - p_x\right)^{c_x}. \quad (12)$$

Table 6 shows the different values of C for each label in each behaviour category. The labels are listed in the first column, and the "Proportion" column shows the number of samples for each label. Different labels are delimited by a colon. Some behaviour categories have 2 labels, while some have 3, so the last column for behaviours with 2 labels is left blank.

Algorithm 1 consists of two parameters: the Dataset D and the Balanced_support threshold value S, which is set by the experimental operator. Different S corresponds to different numbers of frequent item sets. The algorithm firstly scans the whole dataset and regards the generated set as the frequent 1-item set. Next, calculate the Balanced_support of the frequent 1-item set. Then, remove the items whose Balanced_support is lower than S to obtain frequent 2-item set. Next, calculate the Balanced_support of frequent 2-item set. The above procedures are repeated until there is no item in the frequent k item set or only one item left, and the program ends at this point.

Table 7 is an example of the input data for Algorithm 1, which has 7 columns. The first column shows the *Stu_ID*, the second column gives the financial status of that student, and the rest columns record the labels of different behaviours, as defined previously. Since D is the parameter, different behavioural labels can influence the algorithm. The influence will be analysed in the later sections.

2.11. Semisupervised K-Means. Semisupervised learning is an important method in the field of pattern recognition and machine learning, which carries out pattern recognition using a large amount of unlabelled data and a small number of labelled ones. Therefore, this method receives increasing attentions from various areas of research, including predicting dropout rate based on behavioural features [19] and predicting students' academic performance by constructing students' social relationship based on their campus behaviour [20].

The basic idea of semisupervised learning is to label the unlabelled samples by creating a learner using the model hypothesis of data distribution. Its basic setting is as follows.

Given a labelled sample set $L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_1, y_1)\}$ with unknown distribution and an unlabelled sample set $U = \{x_1, x_2, \ldots, x_n\}$, it is expected to learn a function f: $X \longrightarrow Y$, which can set the label of unlabelled set U. Here, x_i and x_j are d-dimensional vectors, and $y_i \in Y$ is the label of sample x. $|L|$ and $|U|$ are the sizes of set L and set U, respectively.

The semisupervised learning includes two kinds of hypotheses, and clustering is one of them. It derives from the intuition that two samples are more likely to have the same label when they are in the same cluster. Based on this idea, the semisupervised K-means algorithm calculates the centroid of each cluster using the labelled data. Then, for each cluster, the Euclidean distance between each unlabelled sample and the centroid is calculated according to formula 13. These unlabelled samples are gradually incorporated into the labelled ones. The above process is iterated until each cluster of unlabelled data is stable.

The following formula shows how the Euclidean distance is calculated:

$$\text{Dis}_{(X,C)} = \sum_{i=1}^{k}\left(x_i - c_i\right)^2, \quad (13)$$

where X and C are both vectors and $X = \{x_1, x_2, \ldots, x_m\}$ and $C = \{c_1, c_2, \ldots, c_m\}$, in which k is the number of samples in each vector.

Based on its core idea, here is the label propagation process of semisupervised K-means, as shown in Algorithm 2.

The input of Algorithm 2 consists of two parts of data, including the array of labelled data L and that of unlabelled data U. Lines 1 to 3 show the initialization part of the algorithm. Firstly, L and U are combined into a new array LU, and at the same time, the centroids of each cluster are calculated. These centroids are then put into a set C, $C = \{C_1, C_2, \ldots, C_s\}$, where s is the number of centroids. Then, a flag is set denoting whether the data label is stable. After that, an iterative loop is entered to judge whether each element in LU has greater distance to the centroids of other clusters than that to the centroid of its belonging cluster. The cluster of data is updated based on the above process, until all the clusters are in a stable state.

The specific format of array L and array U is shown in Table 8, where the top half is the labelled data and the bottom half is the unlabelled data. For both L and U, the first column (*Stu_ID*) shows the ID of students, and the next six columns successively show the value of each behaviour according to

Table 6: The result of C.

Label	Proportion	C of label 1	C of label 2	C of label 3
High: Middle: Low	959 : 1730 : 504	0.4784	0.1997	0.7014
Gper_A: Gper_B: Gper_C	663 : 1063 : 1467	0.6794	0.5105	0.3600
Regular: Irregular	974 : 2218	0.6647	0.2588	
Early: Late	2273 : 919	0.2141	0.6570	
GPA_high: GPA_low	1597 : 1595	0.5	0.5	

Input: The Dataset D, Balanced_support threshold value S
Output: Maximum frequent k item set
(1) Scan all the datasets and get all the data that have appeared, as a candidate frequent 1-item set.
(2) $k = 1$, the frequent 0-item set is considered an empty set.
(3) **While** 1 **do**:
(4) Scan data to calculate the Balanced_support of candidate frequent k item set
(5) Remove the datasets whose Balanced_support of candidate frequent k item set is lower than the threshold value S. Get frequent k items.
(6) **If** The frequent k item set is Empty **Then**:
(7) return frequent $k-1$ item sets as result, and ABA over.
End if
(8) **If** the number of items in frequent k dataset is equal 1 **Then:**
return frequent k item set as result, and ABA over.
End if
(9) $k = k+1$
(10) **End while**

Algorithm 1 The Apriori Balanced Algorithm (ABA).

Table 7: The input data format of Algorithm 1.

Stu_ID	Financial status	Life regularity	Diligence	GPA	GPA_per	Consumption level
2012003	Poor	Irregular	Late	GPA_high	Gper_A	Low
2013014	Not_poor	Irregular	Early	GPA_low	Gper_B	High
2012006	Poor	Regular	Early	GPA_high	Gper_A	Low
2013405	Poor	Regular	Late	GPA_high	Gper_C	Medium

Input: Labelled data array L and unlabelled data array U
Output: Label array LS
(1) Combine L and U into a new array LU
(2) Calculate the centroid of each cluster, appending them into a set C.
(3) Set the loop $Flag \leftarrow Changed$
(4) **While** $Flag \equiv Changed$ **do:**
(5) $Flag \leftarrow Unchanged.$
(6) **For** $lu \in LU$:
(7) Calculate the distance of lu and C_i as D_i.
(8) Put D_i in the array D.
(9) Get the minimum of D, record the label as Lc.
(10) **If** $lu \equiv Lc$ **Then:**
(11) Change the lu label.
(12) $Flag \leftarrow Changed.$
(13) **End if**
(14) **End for**
(15) **End while**

Algorithm 2 Semisupervised K-means label propagation.

Table 8: The input data format of Algorithm 2.

	Stu_ID	Mean	Speed	GPA	GPA_per	Regular	Diligence	Label
Labelled data	2012003	6.82	14.65	82.16	0.30	1.32	25,315	1
	2013005	7.2	13.7	76.6	0.73	1.6	27,392	0
	2012017	10.5	9.44	66.3	0.97	1.4	31,090	0
Unlabelled data	2012052	8.1	12.2	78.4	0.256	1.45	29,610	−1
	2013065	6.6	15.0	82.7	0.11	1.3	26,080	−1
	2013101	7.6	13.0	68.6	0.86	0.68	28,751	−1
	2012115	9.6	10.35	76.2	0.51	0.88	34,947	−1

our previous definition. Finally, the *Label* column indicates the financial status of students, which is converted from the *Financial_status* column of Table 6, with *Poor* denoted as 1 and *Not_poor* denoted as 0. The only difference for L and U is that the *Label* column is initially set to −1 for U, meaning there is no label assigned at the beginning.

3. Results and Discussion

After data preprocessing, behavioural features including students' GPA, GPA_percentage, life regularity, diligence, and consumption level are obtained. Combining these features with the financial status, the processed dataset is produced. As is shown in Table 9, except for *Student_ID* (the first column), each column shows different behaviour labels of students.

Figure 2 illustrates the proportion of different labels for different behaviours. It is obvious that the data to be processed are significantly imbalanced. Therefore, the problems involved in this paper are suitable for the ABA.

3.1. Application of ABA and Results. The results of the ABA are shown in Tables 10 and 11, which are a 2-item set and a 3-item set showing the correlation between behavioural characteristics and poverty level. From the 2-item set table, we first observe from the last column that the Balanced_supports of rules "Poor, Regular" and "Poor, Irregular" are almost the same, which indicates that students' financial hardship has no obvious correlation with their life regularity. Difference emerges when exploring the diligence level. It is clear that the Balanced_support of "Poor, Late" (0.2000) is higher than that of "Poor, Early" (0.1489). This is because that the diligence level is measured by the time of students' first meal every day, and students may skip or seldom eat their breakfast for the sake of saving money, leading to a later time of first meal. When it comes to consumption level, we find that the Balanced_supports of "Poor, Medium" and "Poor, Low" are both higher than "Poor, High," indicating that students in financial hardship spend lower money on average, which is in accordance with the reality. As for the Academic Level, we have found that the Balanced_support of "Poor, GPA_high" is higher than "Poor, GPA_low." This suggests that financially difficult students generally score higher. They may cherish every opportunity to study hard, resulting in better grades. Besides, for the comparison of GPA_percentage, the rule "Poor, Gper_A" has higher Balanced_support than "Poor, Gper_B" and "Poor, Gper_C." This further proves that financially hard students generally study better.

Similar conclusion can be drawn from the 3-item set table. For instance, it can be seen that the Balanced_support of "Poor, Low, Good" (0.05756) is undoubtedly higher than that of "Poor, Medium, GPA_high" (0.032088), "Poor, Medium, GPA_low" (0.022446), and "Poor, High, GPA_high" (0.032985). This further suggests that students in financial hardship tend to spend fewer money and get higher grades. Moreover, financially hard students have relatively lower diligence level, because the Balanced_support of those items containing "Late" is generally higher than the ones containing "Early." For similar reasons as the 2-item set, there is no obvious difference in terms of life regularity.

Compared with the original Apriori, it is worth noticing that if using *Support* (the metric of the original Apriori algorithm) for association rule mining, the *support* of "Poor, Medium" is larger than that of "Poor, Low." However, results from last step show that financially hard students have low consumption level rather than medium. Therefore, the traditional support cannot reflect the patterns hidden in the original data distribution, but the Balanced_support will solve such a problem.

To prove the validity of the Balanced_support, we did the following steps.

Firstly, for each behavioural feature, we figured out two proportions. One is the proportion of students labelled with each specific behavioural feature in all of the students and the other is the similar proportion in those students in financial hardship.

Secondly, the changing trend of the obtained two proportions is compared. As is shown in Figure 3, the percentage of poor students on some behavioural labels have increased, for example, GPA_high, Gper_A, Irregular, and Late, with the increasing rate being 8%, 6.6%, 2.9%, and 1.6%, respectively. Such results indicate that the group of financially hard students has different distributions of behavioural labels compared with that of all the students. That is to say, students in financial hardship show different behaviours. Specifically, they tend to be more hard-working, with better academic level and lower consumption. Besides, based on our definitions for life regularity and diligence, students in financial hardship live a little bit more irregular life and tend to be less diligent. Generally, such difference in the two proportions suggests that our proposed Balanced_support is reasonable, because an increasing proportion on a certain label of certain behaviour indicates an

Table 9: Format of the processed dataset.

Stu_ID	Academic1	Academic2	Regularity	Consumption	Poverty	Diligence
2012005	Gper_A	GPA_high	Regular	High	Poor	Early
2013008	Gper_B	GPA_high	Irregular	Medium	Not_poor	Late
2012074	Gper_C	GPA_low	Regular	Low	Not_poor	Late

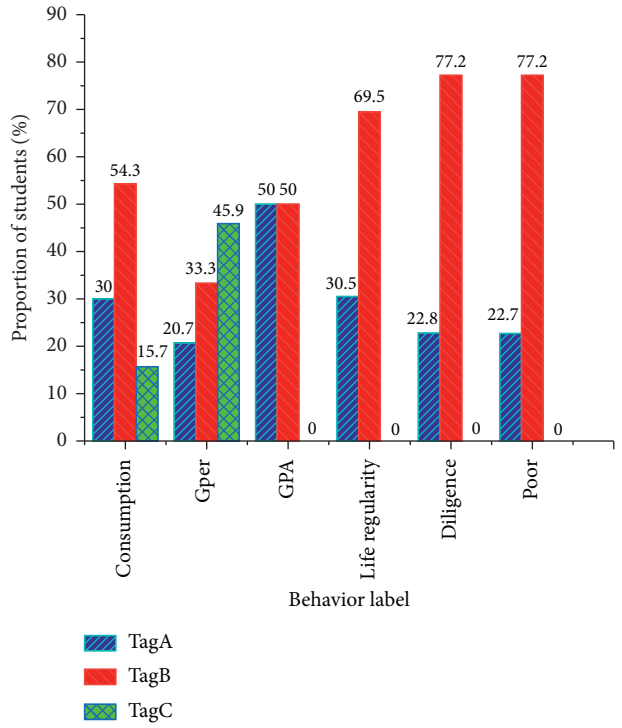

Figure 2: Label distribution.

Table 10: 2-item set.

Item content	Support	Balanced_support ($C = T/2*$ (Inum(max)/Inum(min)))
Poor, Regular	0.2758	0.1833
Poor, Irregular	0.7241	0.1874
Poor, Early	0.6965	0.1489
Poor, Late	0.3034	0.2000
Poor, High	0.2165	0.1035
Poor, Medium	0.5462	0.1090
Poor, Low	0.2372	0.1663
Poor, GPA_high	0.5641	0.2820
Poor, GPA_low	0.4358	0.2179
Poor, Gper_A	0.2731	0.1855
Poor, Gper_B	0.3296	0.1682
Poor, Gper_C	0.3972	0.1429

increasing tendency that poor students are more likely to have this kind of behaviour.

We have introduced that the input parameter of the Apriori Balanced Algorithm (ABA) contains a dataset D, as shown in Table 7, where six behaviour features are listed as columns. Next, we dive deeper into the fine-grained relationship between the proportion of poor students and different types of behaviour features to see whether the previous conclusions drawn through Balanced_support are in accordance with the behaviour characteristic of poor students in our dataset.

Firstly, we explored the distribution of score among financially hard students. In our university, the maximum score for each subject is 100 points. Dividing all the students into 5 categories according to their scores, we figured out the proportion of poor students in each category. Seen from

Table 11: 3-item set.

Item content	Support	Balanced_support ($C = T/2*$ (Inum(max)/Inum(min)))
Poor, Low, GPA_high	0.164137	0.05756
Poor, Medium, GPA_high	0.32137	0.032088
Poor, Medium, GPA_low	0.2248	0.022446
Poor, High, GPA_low	0.1379	0.032985
Poor, Medium, Gper_A	0.1627	0.0220745
Poor, Medium, Gper_B	0.1724	0.017575
Poor, High, Gper_C	0.1158	0.019943
Poor, Good, Gper_A	0.2717	0.092296
Poor Good, Gper_B	0.2517	0.06424
Poor Not_Good, Gper_C	0.36	0.0648
Poor, Early, Irregular	0.5075	0.028120
Poor, Early, Regular	0.1889	0.026882
Poor, Early, Gper_A	0.1848	0.026880
Poor, Early, Gper_C	0.2744	0.0211496
Poor, Early, Gper_B	0.2372	0.025925
Poor, Late, GPA_high	0.16137	0.05301
Poor, Early, GPA_high	0.4027	0.043109
Poor, Early, GPA_low	0.2937	0.031440

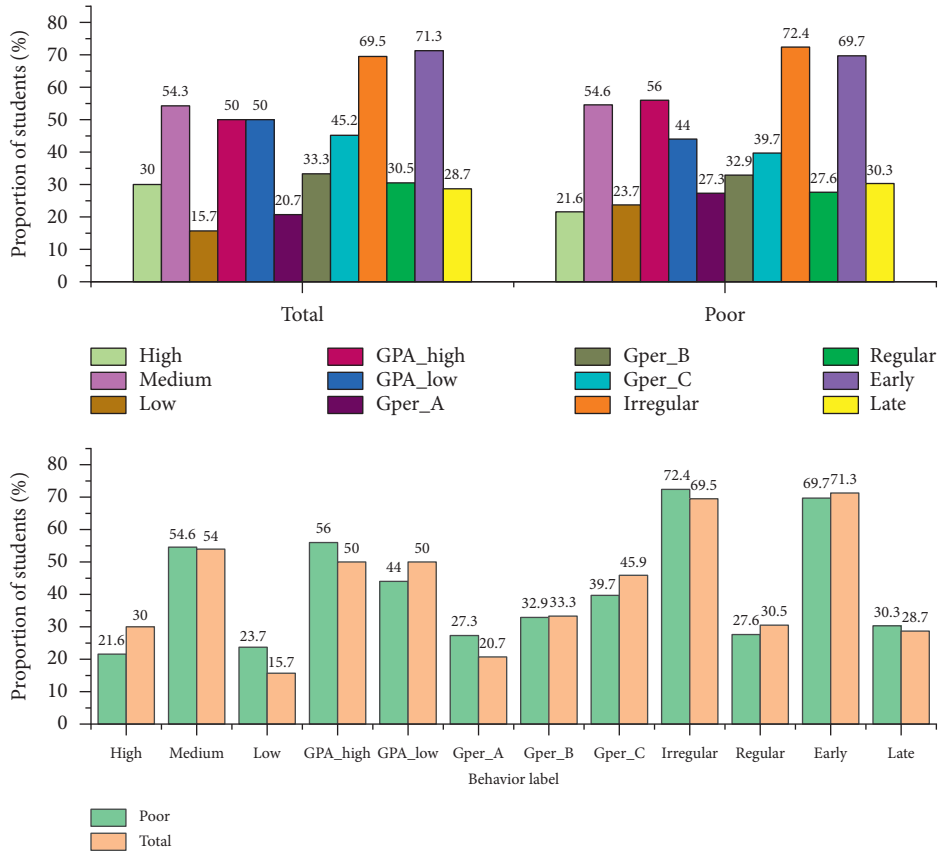

Figure 3: Label proportion and changing trend.

Figure 4, for students score higher than 90 points, 28.57% of them are poor ones. For lower-score categories, such proportion generally follows a downward trend. Such phenomenon indicates that students in financial hardship basically have a higher academic performance.

Besides, we also carried out similar experiments on GPA_percentage. In Figure 5, students are divided into 10 categories to calculate the proportion of poor students in different GPA Ranking. For example, 0–0.1 means the ranking of a student within his/her own major is top 10%

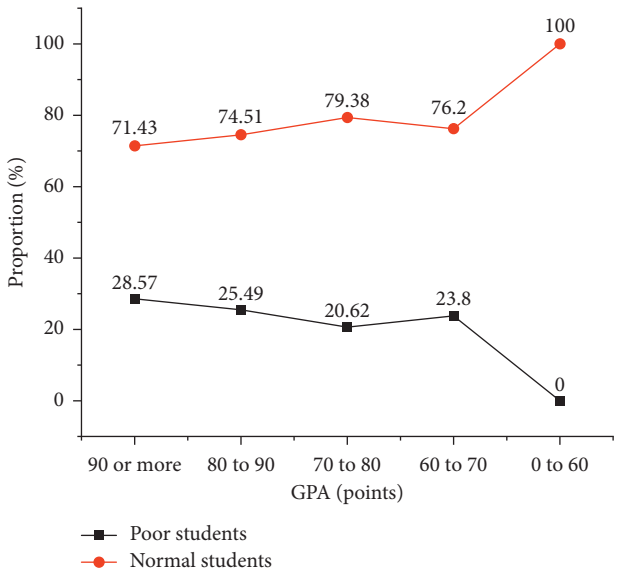

Figure 4: Proportion of poor students in different score ranges.

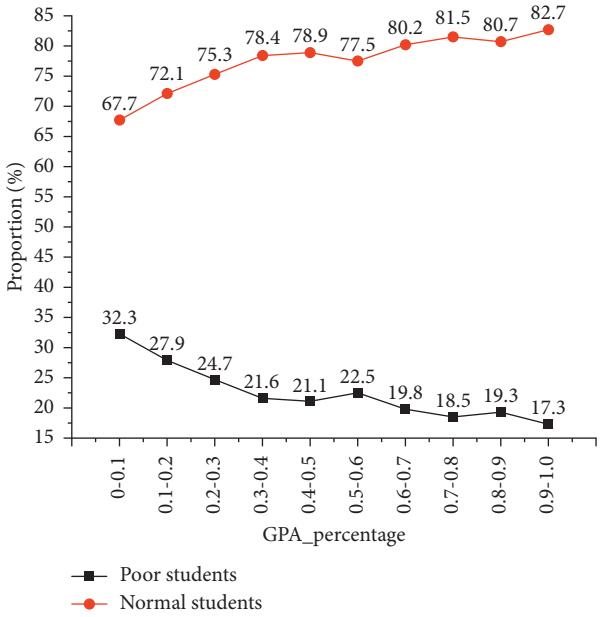

Figure 5: Proportion of poor students in different GPA rankings.

and 0.1–0.2 corresponds to the ranking of top 10% to top 20%. It can be seen that for students ranking top 10%, 32.3% of them are in financial hardship, but with their GPA ranking increasing (meaning a decreasing academic performance), the proportion of poor students in respective ranking declines. When it comes to the last 10%, poor students only account for 17.3%. Such results act as another proof that financially hard students generally perform better in study.

For consumption level, using the similar method as above, we first divided the average consumption amount into seven categories. In Figure 6, for all students whose average consumption amount is in 0–5 yuan and 5–6 yuan, those in financial hardship take up for 63.6% and 40.6%,

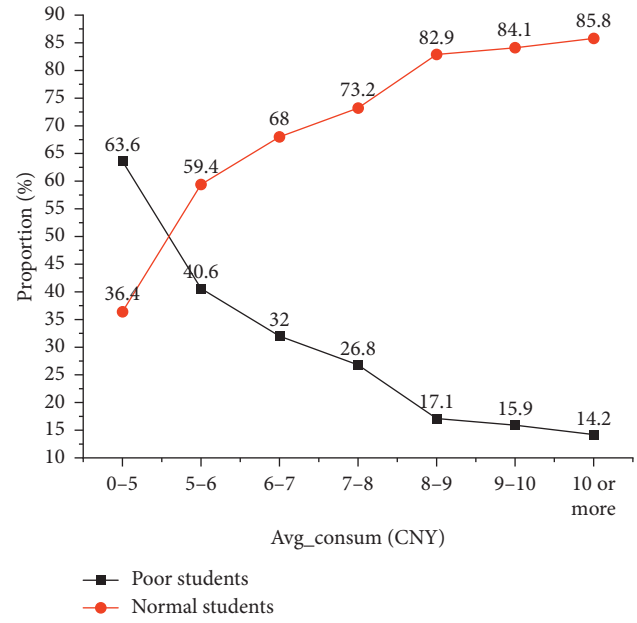

Figure 6: Proportion of poor students in different average consumption amounts.

respectively. Thus, it is obvious that most students have a rather low consumption level, which is also evidenced by the 14.2% of poor students in the group of "10 or more."

Since consumption speed is also a component of our definition for consumption level, we also find out the relationship of distribution of poor students and their average times to spend up 100 yuan. Students are grouped into seven categories, and we lay particular emphasis on those whose times are over 10. According to Figure 7, for those students who use 20 or more times to spend up 100 yuan (meaning a low consumption amount each time), 63.6% are financially hard students. Also, the last three groups all show a high percentage of poor students. This further proves the low consumption level of students in financial hardship.

Previously, we defined life entropy (LE) to represent students' life regularity. Here, we spotlight poor students and explore their distribution with different values of LE. Judging from Figure 8, grouping students into seven categories by LE, the proportion of poor students increases when LE becomes larger. According to our definition, a larger LE corresponds to a lower regularity. This can be considered as an extra evidence for our previous conclusion about the life regularity for poor students.

Finally, we studied the relationship between students' average time of first meal and the proportion of poor students, which shows the pattern of diligence level. As shown in Figure 9, after dividing first meal time into 6 categories, we can find that poor students take up the most in the 8:00–9:00 group, followed by 7:00–8:00 and 9:00–10:00. This means that poor students tend to eat their first meal later, showing a relatively lower diligence level according to our definition.

The above analysis further explained the detailed distribution of poor students relating to different behavioural features, and the obtained results basically conform to the

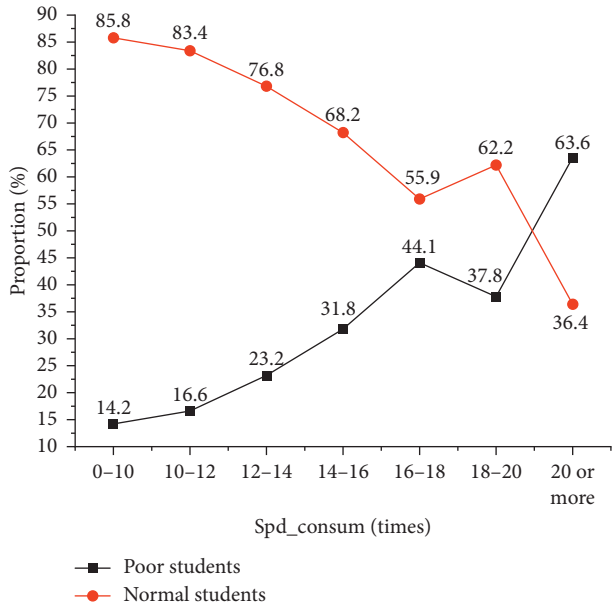

Figure 7: Proportion of poor students in different consumption speeds.

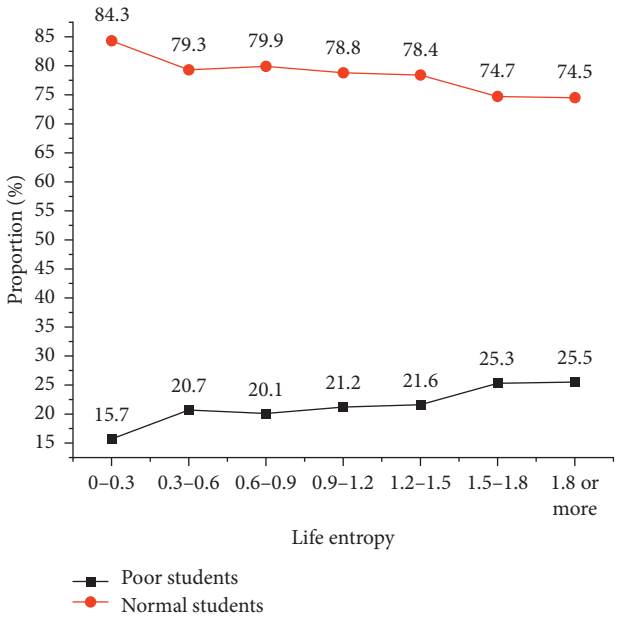

Figure 8: Proportion of poor students in different life entropies.

association rules we have found using the proposed Balanced_support. Therefore, the Apriori Balanced Algorithm (ABA) can be used for mining the relationship between students' poverty level and their daily behaviour.

3.2. Validation on the Groceries. To further verify the effectiveness and adaptability of the ABA algorithm, using Apriori as the comparison algorithm, we tested our algorithm on the public dataset Groceries. The data format of Groceries is shown in Table 2. The comparison results of the

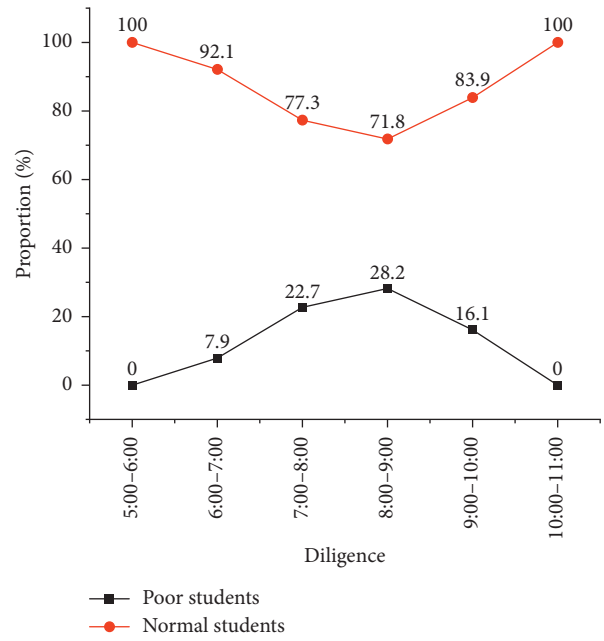

Figure 9: Proportion of poor students in different times of first meal.

Apriori and ABA algorithm are shown in Table 12. Besides, the number of different products in the dataset is shown in Table 13.

Seen from Table 12, for most item sets, a larger *Support* can represent the association strength among the items. In the comparison of Group 1 and Group 2, item sets with higher *Support* also have a higher *Balanced_support*. For example, the *Support* of "soda, sausage" is higher than "soda, pastry," so as the *Balanced_support*. We also know from Table 13 that the number of pastry is 875, which is close to that of sausage (924). That is to say, when comparing item sets with items of similar quantities, the proportion of items in the whole dataset will not influence the association rules obtained. In this situation, we can get the same association rules using either *Support* or *Balanced_support*, i.e., the association between soda and sausage is stronger than soda and pastry.

On the contrary, when the quantities of the items of the same item set are really different, the *Support* of the item with larger quantity is obviously higher than that with lower quantity. For instance, in Group 3, we can find that the *Support* of "whole milk, soda" (0.040061) is obviously higher than "whole milk, shopping bags" (0.024504), since the quantity of soda is 1715, while that of shopping bags is 969. The quantity of these two items are obviously different, so the proportions of them in the whole dataset are also quite different. However, the *Balanced_support* of "whole milk, shopping bags" is indeed higher than "whole milk, soda." That is to say, regardless of the difference in proportions, the association of "whole milk, shopping bags" is actually higher. Such conclusion also has significance in practical use. For example, when stores intend to increase the sales volume of whole milk by increasing the sales volume of "shopping bags" and "soda," conclusion from the ABA tells that "shopping bag" will be a better choice.

Table 12: Results of ABA on Groceries.

Group	Item content	Support	Balanced_support
1	Whole milk, frankfurter	0.020538	0.016902
	Whole milk, other vegetables	0.07483	0.031346
	Whole milk, bottled beer	0.02043	0.015487
2	Soda, pastry	0.021047	0.01081058
	Soda, sausage	0.0243009	0.01181901
3	Whole milk, shopping bags	0.024504	0.015633
	Whole milk, soda	0.040061	0.014462
4	Other vegetables, rolls/buns	0.04260294	0.013889
	Other vegetables, pastry	0.0257244	0.0152135
5	Yogurt, root vegetables	0.02582613	0.010246
	Yogurt, whipped/sour cream	0.020742247	0.0125130

Table 13: Number of products in the Groceries dataset.

Products	Number of product
Whole milk	2513
Frankfurter	580
Other vegetables	1903
Bottled beer	1087
Soda	1715
Pastry	875
Sausage	924
Shopping bags	969
Rolls/buns	1809
Yogurt	1372
Root vegetables	1072
Whipped/sour cream	705

3.3. Semisupervised K-Means Application and Results

3.3.1. Data Preparation. Based on the association rules obtained in the previous sections, i.e., poor students tend to study better and spend less money, we constructed the labelled dataset and the unlabelled dataset, which are the input parameters of Algorithm 2. Labelled data refer to the data with a label that indicates a student's poverty level, so the key problem is how to select poor students from the whole dataset. Results from Figures 4–9 have shown that financially hard students have higher academic performance, lower consumption, and irregular life. Based on such principle, we defined 4 rules for choosing financially difficult students, as shown in Table 14.

Different rules correspond to different amounts of labelled data and unlabelled data, since the criteria for different behavioural labels vary. As is shown in Table 15, if we set R1 as rule, for example, the labelled data will contain 9 Poor students and 30 Not_poor students, and the amount of unlabelled data will be 3151. Similarly, rules R2, R3, and R4 correspond to different amounts of labelled data and unlabelled data, respectively.

Table 14: Rules for choosing financially difficult students.

Rule	GPA	GPA_percentage	Avg_consum	Regular
R1	≥ 85	≤ 0.25	≤ 7	≥ 1.5
R2	≥ 80	≤ 0.3	≤ 7.5	≥ 1.5
R3	≥ 75	≤ 0.35	≤ 7.5	≥ 1.5
R4	≥ 70	≤ 0.4	≤ 8	≥ 1.5

Table 15: The amount of labelled and unlabelled data under different rules.

Rule	Labelled data		Unlabelled data
	Poor	Not_poor	
R1	9	30	3151
R2	60	200	2928
R3	75	255	2860
R4	114	384	2992

According to the experiment of semisupervised *K*-means, we finally realized the identification of poor students, including identifying Not_real financially hard students and finding out the hidden poor students from the poor student list provided by the university.

3.3.2. Evaluation Metric. Predicting students with financial difficulties is extracted as a binary classification problem in this paper. To validate the effectiveness, four commonly used metrics are selected:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FP} + \text{FN}}, \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TF} + \text{FN}}, \\ \text{F1} &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \end{aligned} \tag{14}$$

where TP means the number of students with financial difficulties that are classified correctly, TN is the number of students without financial difficulties that are classified correctly, and FN and FP mean the number of students with financial difficulties and normal students that are incorrectly classified.

3.3.3. The Rules' Influence on Model. In the previous sections, we have defined different rules for choosing financially hard students, and the amount of labelled data and unlabelled data varies with rules, and such amount will influence the prediction performance accordingly. In order to explore the impact of different rules on the prediction performance, we use different rules to generate different datasets and conduct comparison experiments in the logistic regression model. The results are compared in Table 16.

From Table 16, we find that R2 has better performance than other rules, so in the next experiment, we use R2 as the

Table 16: Performance comparison of different rules on logistic regression.

Rule	Accuracy	Precision	Recall	F_1
R1	0.9529	0.9472	0.8965	0.9212
R2	**0.9682**	**0.9177**	**0.9661**	**0.9413**
R3	0.9616	0.8969	0.9694	0.9317
R4	0.9586	0.8735	0.9888	0.9276

main rule for choosing the amount of labelled data and unlabelled data in Algorithm 2.

3.3.4. The Process of Label Propagation. The data processed by the semisupervised *K*-means algorithm contain six dimensions. We selected GPA as *X*-axis and $\text{Avg}_{\text{consum}}$ as *Y*-axis. The process of label propagation is displayed in Figure 10.

Semisupervised learning requires both unlabelled data and labelled data. Here, the initialized data contain 20% labelled data and 80% unlabelled data. As shown in Figure 10(a), in labelled data, blue points represent students with financial difficulties, orange points represent students without financial difficulties, and gray points represent data without labels. In Figure 10(b), the proportion of labelled data increased from 20% to 40%, and that of the unlabelled data decreased from 80% to 60%. This is because after the process of the semisupervised *K*-means algorithm, the 20% unlabelled data were divided into different categories according to the Euclidean distance of the centre point of the two categories. Figures 10(c)–10(e) successively show that the labelled data propagate labels to 60%, 80%, and 100% of total data, respectively. Figure 10(f) shows the classification of all data by means of SVC, and it is found that SVC can fit and work out a classification curve well, indicating good propagation effect.

During the propagation, a number of points changed from blue to orange, representing the process of identifying Not_real financially hard students. Some of the dots change from orange to blue, representing the process of finding students with financial difficulties. As shown in Figure 10(a), the blue point X_1 represents a student with financial difficulty. After identification by the model, it was found that X_1 did not conform to the behavioural characteristics of students with financial difficulties, so it was remarked as orange in the process of propagation. The orange point X_2 in Figure 10(a) represents a student without economic difficulties. After model identification, it was found that the behavioural characteristics of X_2 accord with the characteristics of students with economic difficulties, so it was remarked as blue in the process of propagation. This process represents the identification of students with hidden financial difficulties. Therefore, this model can be used to identify the students without financial difficulties in the poor student list and discover the hidden students with financial difficulties from all of the students.

3.3.5. Label Propagation's Influence on Prediction. In this section, four classical classification algorithms are used on our new dataset processed through the proposed method. The input format of all algorithms is shown in Table 17. Their performances are compared in Table 18. Compared with the model trained by the original data with an old label, the performance of the model trained by the new dataset with a new label has been significantly improved. This means that label propagation has greatly improved the prediction effect of the model. When tested on logistic regression, it achieves an accuracy of 0.96, much higher than other algorithms. Besides, it achieves a relatively higher F_1 score of 0.94 and a highest recall of 0.96 despite the lowest precision. This suggests that our method is more suitable for logistic regression when used for the identification of financially hard students.

3.3.6. The Influence of Different Behavioural Features on the Model. Although we have found that the logistic regression achieves the best result on our new dataset, how its performance changes with different features is still under exploration. In this section, we test different behavioural features on logistic regression, and the performance is shown in Table 19. Among all the behavioural features, GPA_percentage is the most outstanding, with an accuracy of 0.88, a precision of 0.61, a recall of 0.95, and a F_1 score of 0.74. In addition, GPA also achieves a high accuracy and high precision. This suggests that GPA_percentage and GPA are more distinguishing features for the identification of financially hard students. On the contrary, *Avg_consum*, *Regular*, and *Diligence* achieve relatively lower accuracy and very low precision, especially for *Regular* and *Diligence*, whose precision, recall, and F_1 score are all 0. Such phenomenon indicates that these behavioural features, if used independently, cannot determine if a student is in financial hardship. As for *Spd_consum*, though it achieves fairly low precision and F_1, it still contributes to the identification of financially hard students. Therefore, it can be safely concluded that GPA_percentage, GPA, and *Spd_consum* contribute a lot in the identification of students in financial hardship, while the rest features have smaller contributions.

Finally, the model is trained using all of these six behavioural features, and the result is shown in the *Total* row in Table 16 with the highest accuracy, recall, and F_1 score. That is to say, identifying financially hard students is a comprehensive process determined by multiple behaviour features, and our proposed features are effective for such a process.

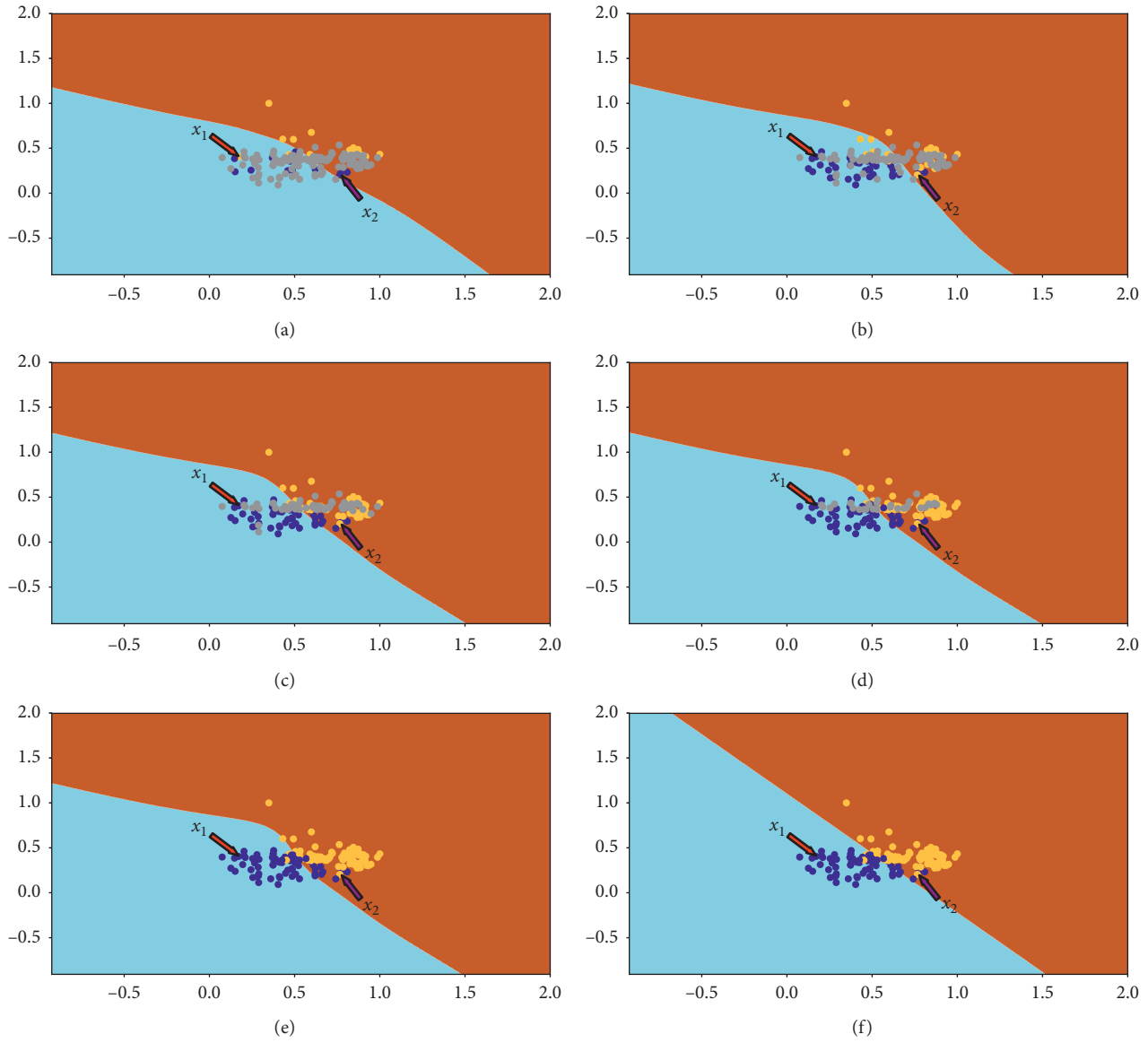

Figure 10: Propagation process. (a) Spreading with 80% unlabelled and 20% labelled, (b) spreading with 60% unlabelled and 40% labelled, (c) spreading with 40% unlabelled and 60% labelled, (d) spreading with 20% unlabelled and 80% labelled, (e) spreading with 100% labelled, and (f) SVC with RBF kernel.

Table 17: The input format of different models with different labels.

Stu_ID	GPA	GPA_per	Avg_consum	Spd_consum	Regular	Diligence	Old label	New label
A	83.68182	0.16	7.98	12.5	1.42	27074.72	0	1
B	73.875	0.79	9.30	10.7	1.63	27978.82	0	0
C	76.83333	0.57	8.51	11.7	1.56	27341.25	0	0
D	81.03704	0.47	8.62	11.5	1.36	27714.55	1	1

Table 18: Performance of different models with different labels.

Data	Model	Accuracy	Precision	Recall	F_1
Old label	SVM	0.74	0.035	0.49	0.07
	KNN	0.72	0.136	0.399	0.203
	Logistic regression	0.74	0.005	0.36	0.01
	Multilayer perceptron	0.736	0.038	0.43	0.07
New label	SVM	0.88	**1.0**	0.69	0.81
	KNN	0.87	0.998	0.68	0.81
	Logistic regression	**0.96**	0.91	**0.96**	**0.94**
	Multilayer perceptron	0.86	**1.0**	0.67	0.80

Table 19: Performance of different features on logistic regression.

Features	Accuracy	Precision	Recall	F_1
GPA	0.75	0.95	0.53	0.68
GPA_percentage	0.88	0.61	0.95	0.74
Avg_consum	0.72	0.001	0.50	0.002
Spd_consum	0.74	0.18	0.63	0.28
Regular	0.72	0	0	0
Diligence	0.72	0	0	0
Total	**0.96**	**0.91**	**0.96**	**0.94**

4. Conclusions

In this work, we proposed the Apriori Balanced Algorithm (ABA) and carried out association rule mining for students in financial hardship through a new measure, *Balanced_support*, which is used to represent correlation strength and better find out how students' poverty level is correlated with their daily behaviour. In addition, through association rule mining, we found that students in financial hardship have better academic performance and lower consumption level with relatively lower life regularity and diligence level. Next, based on the obtained association rules, we noticed that some students selected in the poor student list do not conform to the above rules. Therefore, we used semisupervised *K*-means to identify students in real financial hardship, as well as finding out the students who are not really poor. Tested by classical classification algorithms, the proposed method displays better identification performance compared with the original assessment approach.

In the future, we need to further optimize our framework. For example, data with other dimensions are required to describe student behaviour more comprehensively, such as water consumption records, book-borrowing records, and Internet login records. Also, we will incorporate knowledge from other research areas to explore the behavioural characteristics of financially hard students more deeply, such as combining with psychology to study psychological problems of students in poverty.

Data Availability

The original data of precise behavioural records cannot be released in order to preserve the privacy of individuals.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Higher Education Research Project of Jilin Province (grant no. ZD18027).

References

[1] M. Nie, L. Yang, J. Sun et al., "Advanced forecasting of career choices for college students based on campus big data," *Frontiers of Computer Science*, vol. 12, no. 3, pp. 494–503, 2018.

[2] C. Liu, K. Zhang, W. Huang, and B. Kang, "Big data analysis of campus card based on spark," in *Proceedings of the 2017 9th International Conference on Advanced Infocomm Technology (ICAIT)*, pp. 438–442, IEEE, Chengdu, China, November 2017.

[3] D. Dong, J. Li, H. Wang, and L. Zhu, "Student behavior clustering method based on campus big data," in *Proceedings of the 2017 13th International Conference on Computational Intelligence and Security (CIS)*, pp. 500–503, IEEE, Hong Kong, China, December 2017.

[4] X. Jiang, T. Xu, and X. Dong, "Campus data analysis based on positive and negative sequential patterns," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 5, Article ID 1959016, 2019.

[5] M. Li, C. Huang, D. Wang, Q. Hu, J. Zhu, and Y. Tang, "Improved randomized learning algorithms for imbalanced and noisy educational data classification," *Computing*, vol. 101, no. 6, pp. 571–585, 2019.

[6] Q. Wang, D. Huang, Y. Shen, and Y. Zhang, "The application of active learning in identification of students with financial difficulties," in *Proceedings of the 2017 9th International Conference on Education Technology and Computers*, pp. 136–140, Barcelona, Spain, 2017.

[7] Y. Cao, J. Gao, D. Lian et al., "Orderliness predicts academic performance: behavioural analysis on campus lifestyle," *Journal of The Royal Society Interface*, vol. 15, no. 146, Article ID 20180210, 2018.

[8] D. Lian, Y. Ye, W. Zhu, Qi Liu, X. Xie, and H. Xiong, "Mutual reinforcement of academic performance prediction and library book recommendation," in *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1023–1028, IEEE, Barcelona, Spain, December 2016.

[9] B. Wang, K. Deng, W. Wei, S. Zhang, W. Zhou, and S. Yu, "Full cycle campus life of college students: a big data case in China," in *Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 507–512, IEEE, Shanghai, China, January 2018.

[10] S. Lorenzen, N. Hjuler, and S. Alstrup, "Tracking behavioral patterns among students in an online educational system," 2019, http://arxiv.org/abs/1908.08937.

[11] M. Cantabella, R. Martínez-España, B. Ayuso, J. A. Yáñez, and A. Muñoz, "Analysis of student behavior in learning management systems through a big data framework," *Future Generation Computer Systems*, vol. 90, pp. 262–272, 2019.

[12] C. Guan, X. Lu, X. Li, E. Chen, W. Zhou, and H. Xiong, "Discovery of collegestudents in financial hardship," in *Proceedings of the 2015 IEEE International Conference on Data Mining*, pp. 141–150, IEEE, Atlantic City, NJ, USA, November 2015.

[13] L. Suo and J. Gong, "Identification of university poor students based on data mining," in *Proceedings of the International Conference on Intelligent Computation Technology & Automation*, IEEE, Nanchang, China, June 2015.

[14] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[15] H. Yao, D. Lian, Y. Cao, Y. Wu, and T. Zhou, "Predicting academic performance for college students: a campus behavior perspective," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 3, p. 24, 2019.

[16] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference Very Large Databases, VLDB*, vol. 1215, pp. 487–499, Santiago, Chile, 1994.

[17] J. Yi, S. Li, M. Wu et al., "Cloud- based educational big data application of apriori algorithm and *k*-means clustering algorithm based on students' information," in *Proceedings of the 2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, IEEE, Sydney, NSW, Australia, pp. 151–158, December 2014.

[18] T. Li, "Mining association rules on enrollment information of higher vocational colleges using the apriori algorithm," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 23, no. 4, pp. 775–781, 2019.

[19] W. Li, M. Gao, H. Li, Q. Xiong, J. Wen, and Z. Wu, "Dropout prediction in moocs using behavior features and multi-view semi-supervised learning," in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Vancouver, BC, Canada, pp. 3130–3137, July 2016.

[20] H. Yao, M. Nie, Su Han, X. Hu, and D. Lian, "Predicting academic performance via semi-supervised learning with constructed campus social network," in *Proceedings of the International Conference on Database Systems for Advanced Applications*, pp. 597–609, Chiang Mai, Thailand, April 2017.