*International Journal of Engineering Education* Vol. 35, No. 2, pp. 563–571, 2019
Printed in Great Britain

0949-149X/91 $3.00+0.00


# Predicting At-Risk Students using Campus Meal Consumption Records*

WENJUN QUAN
Chongqing University of Technology, Chongqing, 400054, China. E-mail: 79545989@qq.com

QING ZHOU, YU ZHONG and PING WANG
Chongqing University, Chongqing, 400044, China. E-mail: tzhou@cqu.edu.cn, zhongyu529@126.com, bingyingping@163.com

Predicting student performance (PSP) is of great use from an educational perspective, especially for the at-risk students who need timely support to complete their study. Previous PSP studies have been mainly based on data from questionnaires and specific learning systems. Such data sources have some innate shortcomings. Instead, we used a novel data source, the massive students' campus card usage records, to predict at-risk students. This method has two advantages: convenience in data collection and ability to predict students' overall academic performance. However, as the original data are complex, large in scale and with a lot of noise, it is challenging to extract proper features from them. We adopted a four-step procedure for data preprocessing and the Naive Bayes model for performance prediction. Experiments showed that the proposed prediction model could identify about 70% of the at-risk students. Some features of the at-risk students were also discovered, which might help student counselors and educational researchers better understand the relationship between students' consumption behaviors and their academic performance.

**Keywords:** educational data mining; predicting student performance; campus card; feature selection; classification

## 1. Introduction

Predicting students' performance (PSP) is the oldest and most popular application of data mining in education. PSP's goal is to estimate the unknown value of a student's performance, knowledge, score or mark [1]. It is of particular importance to predict at-risk students who need temporary or constant assistance to complete their study. If they could be identified as early as possible, they might be provided with support in time.

In many PSP studies, the data used mainly come from questionnaire survey and certain specific learning system. Questionnaire survey is quite time-consuming and troublesome. What's more, participants' responses to these question items maybe to some extent inaccurate and subjective rather than objective. As for learning systems, the records on these systems could provide insights into students' performance in certain subjects while it should be noted that student counselors and educational administers usually care more about students' overall academic performance instead of specific courses.

Information systems are widely used in higher education institutes like university, college, etc. to facilitate student management. With these management systems being used, lots of records are generated and can be captured and used in educational research. Among those data, students' records in campus card can be a valuable source of data for research. In China, campus cards are essential for students' daily life. They use their card to dine in the school canteen, purchase in campus shops and do laundry, etc. All these comprehensive and large-volume data are well captured by the campus cards. Such a massive data provides us an opportunity to automatically analyze students' behaviors via data mining techniques. Previous studies have demonstrated the correlation between students' lifestyle (e.g., dietary and sleeping habits) and their academic performance [2–5]. For example, some studies suggest that students who have regular eating habits generally achieve a better academic performance [2, 3], some other studies suggest that students with good sleeping habits (e.g., go to bed early, get up early) tend to have a better academic performance [4, 5]. As consumption records in campus card could partially reflect the students' lifestyles, an interesting question arises: can students' campus card usage records predict their academic performance?

Motivated by this question, we are about to predict students' academic performance using the consumption records in campus cards. This data set has advantages over questionnaires. Campus cards record students' behaviors which are more objective than questionnaires. Besides, these existing data are more accessible. However, there are two difficulties in using this method. First, it is not easy to extract proper features reflecting students' lifestyles from the original campus card usage records. As these records contain the details of every consumption via campus card for each student, the data are complex, large in scale and mixed with a lot of noise. Second, Although the correlation between students' life-

* Accepted 25 November 2018.

styles and their academic performance has been reported in some studies, still it is unclear whether students' lifestyle could accurately predict their academic performance. To be specific, we are interested in the following two questions:

1. Can at-risk students be identified at an early stage from their consumption records via campus card?
2. Which features can best identify at-risk students?

To answer them, the campus card usage records and the academic performance records of 167 students were investigated. Features regarding to students' lifestyles were extracted from their consumption records in campus card. Then, the Naive Bayes model was adopted to identify at-risk students. Experiments showed that the proposed model can identify about 70% of the at-risk students. The relationship between students' lifestyles and their academic performance was also discussed.

The rest of this paper is organized as follows. Section 2 introduces previous related research. Section 3 describes the data set and preprocessing methods. Section 4 includes data analysis and discussion. Finally, Section 5 summarizes this research.

## 2. Related work

### *2.1 Predicting students' academic performance*

Academic performance can be regarded as a significant outcome of education and could reflect the extent to which a student has achieved his/her educational goals. Therefore, predicting students' performance (PSP) is an important research topic. However, PSP is a challenging work because many factors may influence students' performance, such as culture, society, family, socio-economic status, psychological profile, previous schooling, historical academic performance and the interaction between students and the faculty [6].

Many PSP studies utilized questionnaire survey data and learning systems records. Questionnaire surveys were widely used in early PSP studies [7, 8]. However, questionnaire survey generally costs extra money and human resources and the participants might not actively cooperate with it. Recently, the applications of Learning Management Systems (LMS), such as Moodle, Blackboard and WebCT, became popular. LMS can create powerful, flexible and engaging online learning courses and experiences and accumulate rich information about students' behavior. They have been employed in predicting students' academic performance. Romero et al. [9] utilized students' usage data from on-line discussion forums provided by Moodle to predict their final performance. Delgado et al. [10] employed neural network models to predict students' marks by using Moodle logs. It should be noted that while LMS data can be used to predict students' performance on the specific courses in the LMS system, they might not be so effective in predicting the student overall performance or on other courses not included in the system. In addition, many PSP studies predicted students' performance using their historical scores or scores combined with other features. Tarek et al. [11] used students' high school GPA and SAT scores to predict their college level performance. Huang et al. [12] predicted student's academic performance on the dynamics final comprehensive exam by using their historical scores.

The methods to predict students' performance could be summarized into two categories: classification and regression [13]. Classification methods are used to divide the objects into groups. The widely-used classification methods in PSP include Naive Bayes, Decision Tree and Neural Network. Kotsiantis et al. [14] proposed an online ensemble of classifiers that combined an incremental version of Naive Bayes, the 1-NN and the WINNOW algorithms to predict whether the students could pass the *Informatics* course. Romero et al. [15] compared the prediction performance of several different classification algorithms, such as decision tree and neural network, to predict students' grades on a number of Moodle courses. Regression methods are used to estimate students' test scores. Grudnitski [16] employed multiple regression to develop a performance prediction model for an introductory course of financial accounting based on the data collected from questionnaires. Gansemer-Topf et al. [17] used multiple linear regression techniques to predict students at risk of earning below a 2.0 grade point average (GPA) in their first semester of college by using institutional data.

### *2.2 The correlation between students' lifestyle and their academic performance*

Lifestyle indicates many aspects, such as doing sports, eating habits, sleeping orderliness, communication medium like using mobile phone, instant messages, etc. And many previous studies have suggested a correlation between students' lifestyle and their academic performance [18–23]. For instance, Marina studied the relation between the lifestyle (i.e. drinking, smoking, sexual life, mobile phone using and parental strictness) of students in Eastern Samar State University and their academic performance. It was found that students' smoking habits correlated significantly with their academic performance [19]. Faught et al. [20] examined the associations between health behaviors and aca-

demic achievement and found that all health behaviors (e.g., diet, physical activity, sleep duration, recreational screen time usage, height, weight, and socioeconomic status) exhibited associations with academic achievement.

The present study mainly takes into consideration students' dietary behaviors and sleeping habits while investigating the association of students' lifestyle and their academic performance. Some studies have suggested that students' eating habits have an impact on their academic performance [2, 3, 24–27]. Kim et al. studied the association between Korean teenagers' dietary behaviors and their academic performance using questionnaire data obtained from grade 5, 8 and 11 students. They found that the academic performance of these students was strongly associated with dietary behaviors. To be specific, regular breakfast and lunch were more important in grades 5 and 8, while regular dinner was more related with academic performance in grade 11. What's more, the importance of regularity of meals was greater than that of social-economic status [2]. Stea and Torstveit examined the relation of the lifestyle and academic performance of 2,432 Norwegian adolescents. It was suggested from questionnaire data analysis that a regular meal pattern and intake of healthy food (e.g., fruits and vegetables) contribute to higher academic results [3].

Some studies suggested that students' academic performance was associated with their sleeping habits [4, 5, 28–32]. Liang et al. [5] studied the association of the lifestyle and academic performance of students in a Chinese college and found those early to bed and early to rise generally had a better academic performance than those who did not. In a study examining the relation between sleeping behaviors and academic performance of Japanese high school students, Tanaka found that high-ranking school students with higher grades of academic performance of English and mathematics reported more total sleep, earlier bed times on school nights. In addition, improved sleeping quality increased their concentration during school classes and their motivation of study [4].

## 3. Pre-processing

### *3.1 Data Description*

We investigated 167 students (129 males, 38 females) from The Department of Computer Science and Technologies, Chongqing University, China. All of them started their university education in the year of 2012. There were two data sets. The first contained the final exam scores of all courses in the third and fourth semester for each student. They were obtained from the management information system of school academic affairs office. The second included the students' campus card usage records in the two semesters, which were collected in the school campus card service center.

As for the academic performance, about 37% of the students would fail some courses in one semester. If these students could be identified early, they could get timely support from student counselors and course teachers. In this study, at-risk students were defined as those who failed one or more courses in one semester.

The second data set recorded students' consumptions on campus. Since most students live on campus and the campus card was a convenient payment passport, they would use the campus card for daily consumptions in the canteens, campus stores and student laundries, etc. The campus card information system stored students' consumptions at places mentioned above on the campus. Table 1 lists a few samples of a student's consumption records, each including the student ID and the location, time and expenditure for the consumptions. The original campus card usage records showed that most students had consumption records in canteens but only a few of them had records in supermarkets or laundries. Therefore, the focus of this study is to identify at-risk students from their consumption records in the university canteens.

### *3.2 Data preprocessing*

The purpose of data preprocessing is to extract a set of features from the original data set. This step is important as the feature set establishes the foundation of prediction model construction. According to previous studies and student counselors' experience, students' eating habits and sleeping habits are associated with their academic performance. These concepts cannot be directly measured from students' consumption records. Yet they might be partially reflected in the students' consumption time, expenditure and regularity of daily meals, including breakfast, lunch, supper and

**Table 1.** Examples of a student's consumption records

| Student ID | Consumption location | Consumption time | Consumption expenditure |
|---|---|---|---|
| 2012**** | Rainbow Supermarket | 2014/6/23 13:44:04 | 5 |
| 2012**** | No.3 Canteen | 2014/5/10 11:39:17 | 11.9 |
| 2012**** | Orchid Garden Laundry | 2013/5/11 21:00:56 | 3 |

night snack. In order to make accurate analysis, the following problems needed to be considered.

1. How to judge whether a record is a meal consumption record? The original dataset includes consumption records at canteens, campus stores and student laundries, etc., only a few of them belong to canteens. What's more, even a consumption record occurs in the canteen, the student may just buy a drink or snack but not have a regular meal.
2. Should all the meal consumption records be taken into account? Most of the students do not live on campus in summer and winter vacations; many students have few campus consumptions in weekends and holidays. Students' consumption records indicate that there might be a huge gap in meal consumption behavior between these days and regular days.
3. How to determine the type of a meal consumption record? There is no strict definition for different types of meals. A meal at 10:30 a.m. could be considered as either a breakfast or a lunch, depending on individual conceptions. Besides, one meal might involve a couple of consumption records in one canteen or even in several adjacent canteens. These factors make the classification of a record complicated.

In order to solve these problems, the following four-step data processing procedure was employed based on the interviews with students and canteen faculties and on the statistics of the original records.

Step 1. Data filtering. Three types of consumption records were filtered. The first is the records of sellers that are not canteens in this university. The second is the records produced at weekends or in holidays as they cannot reflect the stable characteristics of a student. The third is the records after midterm of each semester, because the student counsellors hope to identify at-risk students by the midterm.

Step 2. Meal labeling. A day was divided into four time intervals according to the feedback from students and canteen faculties. The type of a meal consumption record is determined according to the interval in which the consumption time lies, as follows: breakfast (0:00–11:00), lunch (11:00–16:00), supper (16:00–20:00) and night snacks (20:00–24:00).

Step 3. Noise removal. If the record is not a breakfast, the consumption amount is less than 5 Yuan (around 0.7 U.S. dollar), and the time span between two neighbor consumptions is more than one hour, then the student was more likely to buy a drink or snack other than to have a regular meal. Such record was removed as a noise.

Step 4. Record combination. All records with the same type in the same day were combined as one record. The consumption time of the new record was set as the earliest time of old records, and its consumption expenditure was set as the sum of old records.

After the above steps, a set of cleaner consumption records were generated. Then we extracted 21 initial features from these records, which could be categorized into three types: meal time, meal expenditure, and meal regularity (Table 2).

**Table 2.** The original features of Students' meal consumptions

| Types | Features |
|---|---|
| Meal time | Average starting time of breakfast |
| | Average starting time of lunch |
| | Average starting time of supper |
| | Average starting time of the three meals (breakfast, lunch and supper) |
| Meal expenditure | Average expenditure of breakfast |
| | Average expenditure of lunch |
| | Average expenditure of supper |
| | Average expenditure of the three meals |
| Meal regularity | Standard deviation of starting time of breakfast |
| | Standard deviation of starting time of lunch |
| | Standard deviation of starting time of supper |
| | Standard deviation of starting time of the three meals |
| | Standard deviation of breakfast expenditure |
| | Standard deviation of lunch expenditure |
| | Standard deviation of supper expenditure |
| | Standard deviation of expenditure of the three meals |
| | Proportion of night snacks |
| | Number of breakfasts |
| | Number of lunches |
| | Number of suppers |
| | Number of all meals |

## 4. Data analysis and discussion

### *4.1 Data analysis*

Since the main purpose of this study is to predict at-risk students based on the campus card usage records, we need to establish a prediction model. The input of the model is the preprocessed data introduced in section 3, which includes 21 features (see Table 2). Feature selection is a procedure to select a subset of useful features from the original feature sets [33], which can drastically cut the training time by reducing the number of features. Moreover, it can make the model simpler and easier to understand [34]. In this study, a classic feature selection method called sequential forward selection (SFS) is adopted, which is a search method that starts with an empty set of features and adds a single feature at each step with a view to improving the overall performance of the prediction. After the process, 15 features were selected (see Table 4).

We adopt the Naive Bayes (NB) model in our study because it's simple in computation while highly efficient for classification tasks [35]. NB is based on the Bayes theory (Eq. (1)).

$$P(C/X) = \frac{P(X/C)P(C)}{P(X)}, \tag{1}$$

where $C$ denotes the class label of a sample, e.g., an at-risk student or a low-risk student, and $X$ denotes characteristics of a sample, e.g., the average starting time of breakfast. $P(C/X)$ denotes the probability of a sample belonging to c when it presents the characteristics of x. As NB assumes that all features are conditionally independent of the class, $P(X/C)$ can be calculated using Eq. (2):

$$P(X/C) = \prod_i P(x_i/C), \tag{2}$$

where $x_i$ denotes the $i$-th feature.

Given all features of a sample, NB outputs the class that maximizes $P(C/X)$. As indicated by Eq. (1) and Eq. (2), $P(C/X)$ is proportional the products of all $P(x_i/C)$, which provides an interpretation about the relationship between features and prediction results. For the current study,$C$ represents at-risk students, $X$ for the features extracted from the original dataset,e.g., average starting time of the three meals (breakfast, lunch and supper). The classification process was done with the NB classifier in the sklearn library of python, see Table 3 for results of the prediction.

Table 3 lists the prediction performance by using NB model. The performance measures adopted included recall, precision and F1. Assume that the category to be predicted is positive class. Precision is the fraction of the number of samples correctly labeled as positive class and the number of all the samples labeled as positive class by the classifier, while recall is the fraction of the number of samples correctly labeled as positive class and the total number of samples that actually belong to positive class. F1, the harmonic mean of recall and precision, suggests the overall performance of the prediction. In the experiments, recall denotes the proportion that at-risk students are correctly labeled as at-risk students; Precision denotes the proportion that the students labeled as at-risk students indeed belong to at-risk students; Generally, there is a trade-off between recall and precision. As stated by student counsellors in the department of computer science, more importance is attached to recall than precision. More specifically, a precision higher than 50% is satisfactory, while the recall should be as high as possible. The overall performance of the prediction model is approximately 65%, indicating about 2/3 of the at-risk students could be identified (see Table 3).

Table 4 lists the correlation between class labels and 15 selected features. Correlation coefficients and the P value are introduced in the table, where the former indicates the positive/negative correlation between two variables, the later indicates the significance. Generally, correlation is highly significant when $p < 0.05$. It's worth mentioning that when the p value is 0 in the table, it actually means the value is very small but not zero (<<0.0001). Based on correlation coefficients and the P value, we can see that generally the students who have breakfast earlier and have less night snacks are less likely to fail some courses. This finding is strengthened by the difference in the mean values of these features between at-risk students and low-risk students, as illustrated in Table 5.

To further explore the relationship between the selected features and at-risk students, we conducted a cluster analysis with the k-means algorithm. We chose five representative features: average starting time of breakfast, average expenditure of breakfast, number of breakfasts, proportion of night snakes and the standard deviation of starting time of lunch, and $k$ was set as 2. Then ANOVA analysis was made to check whether there was a significant difference between the clusters. It turned out that the two clusters had significant difference ($p < 0.001$) in the five features (Table 6) and in proportion of students who failed at least one course (Table 7). It can be inferred that the students in Cluster 1 are more likely to be at-risk students, while the students

**Table 3.** The prediction performance of NB model

| Recall | Precision | F1 |
|---|---|---|
| 0.6788 | 0.6509 | 0.6529 |

**Table 4.** Correlation between selected features and academic performance

| Features | Correlation coefficient | p |
|---|---|---|
| Average starting time of breakfast | 0.3156 | 0 |
| Standard deviation of starting time of breakfast | 0.1506 | 0.006 |
| Average expenditure of breakfast | 0.255 | 0 |
| Standard deviation of breakfast expenditure | 0.1733 | 0.0015 |
| Number of breakfasts | –0.3426 | 0 |
| Standard deviation of starting time of lunch | 0.2117 | 0.0001 |
| Number of lunches | –0.2158 | 0.0001 |
| Average starting time of supper | 0.201 | 0.0002 |
| Standard deviation of starting time of supper | 0.1587 | 0.0037 |
| Number of suppers | –0.0864 | 0.116 |
| Number of all meals | –0.2616 | 0 |
| Average starting time of the three meals | 0.3308 | 0 |
| Standard deviation of starting time of the three meals | 0.0462 | 0.4014 |
| Standard deviation of expenditure of the three meals | 0.212 | 0.0001 |
| Proportion of night snacks | 0.2807 | 0 |

**Table 5.** Mean values of at-risk students and low-risk students in the selected features

| Features | Mean of low-risk | Mean of at-risk student |
|---|---|---|
| Average starting time of breakfast | 8:44:03 | 9:03:31 |
| Standard deviation of starting time of breakfast | 0:48:01 | 0:52:03 |
| Average expenditure of breakfast | 3.88 | 4.65 |
| Standard deviation of breakfast expenditure | 1.69 | 2.02 |
| Number of breakfasts | 25.73 | 18.19 |
| Standard deviation of starting time of lunch | 0:40:23 | 0:50:13 |
| Number of lunches | 25.33 | 21.21 |
| Average starting time of supper | 18:15:32 | 18:28:14 |
| Standard deviation of starting time of supper | 0:57:41 | 1:05:28 |
| Number of suppers | 25.51 | 23.70 |
| Number of all meals | 76.58 | 63.11 |
| Average starting time of the three meals | 13:05:35 | 13:48:06 |
| Standard deviation of starting time of the three meals | 3:55:21 | 3:57:30 |
| Standard deviation of expenditure of the three meals | 3.68 | 4.34 |
| Proportion of night snacks | 0.04 | 0.07 |

**Table 6.** Two clusters (at-risk and low-risk) of students' eating patterns

| | Average starting time of breakfast | Average expenditure of breakfast | Number of breakfasts | Proportion of night snacks | Standard deviation of average starting time of lunch |
|---|---|---|---|---|---|
| (Cluster1:at-risk) n=150 (M/SD) | 9:15:31/0:21:30 | 4.70/1.61 | 17.67/9.47 | 0.07/0.06 | 3256.20/1396.52 |
| (Cluster2:low-risk) n=182 (M/SD) | 8:31:11/0:18:40 | 3.72/1.17 | 27.32/9.51 | 0.04/0.04 | 2132.54/1066.90 |
| F(ANOVA) | 403.67*** | 40.23*** | 85.08*** | 31.31*** | 68.99*** |

**p < 0.01, ***p < 0.001.

**Table 7.** Comparison of the academic performance in different clusters (at-risk and low-risk)

| | Proportion of students who failed at least one course |
|---|---|
| (cluster1:at risk) n=150 (M/SD) | 0.53/0.50 |
| (cluster2:low risk) n=182 (M/SD) | 0.24/0.43 |
| F(ANOVA) | 32.58*** |

** p < 0.01, *** p < 0.001.

in Cluster 2 tend to be low-risk. As shown in Table 6 and Table 7, for at-risk students, the value of number of breakfasts is significantly lower than low-risk students but the values of other features are significantly higher.

### *4.2 Discussion*

We were interested in finding indicators of at-risk students, i.e., which features can best predict at-risk students. Two impressive features were the average starting time of breakfast and the proportion of

night snacks in total meals, both significantly positively correlated to at-risk students. As these features indicate the time of getting up and going to bed, we see that students getting up late and going to bed late were more likely to fail the courses. These findings corroborated previous research findings [4, 5].

There were other four interesting features reflecting whether a student had a regular eating habit: standard deviation of starting time of breakfast, standard deviation of starting time of lunch, standard deviation of starting time of supper and number of breakfasts. The experiment showed a positive correlation between these features and at-risk students, indicating that a student with more regular eating habit is less likely to fail courses. Some studies about college students suggested that a regular lifestyle eating habit was positively correlated with their academic performance [2, 3]. This was further confirmed by the current study.

Cluster experiments, from a different perspective, also suggested that the selected features are good indicators to distinguish low-risk students and at-risk students. It showed that the proportion of at-risk students in Cluster 1 was significantly higher than that in Cluster 2. And the students in Cluster 1 have higher average starting time of breakfast, average expenditure of breakfast, proportion of night snakes and the standard deviation of starting time of lunch and lower number of breakfasts than the students in Cluster 2.

Students' lifestyle like eating and sleeping habits could to some extent indicate their self-discipline. Ample studies suggested that the more self-disciplined students are, the more likely that they perform academically well and vice versa [36–40]. For instance, Tangney et al. [39] found a positive correlation between self-discipline and self-reported grades. Hogan and Weiss [38] found that high self-discipline identified Phi Beta Kappa undergraduates from non-Phi Beta Kappa students with equal intelligence. Simba et al. [40] found that students with high level of self-discipline have a better academic performance compared to those with low to moderate level of self-discipline. Our research further confirmed the correlation between self-discipline and academic performance.

Identifying at-risk students as early as possible is of great importance. The present study used students' consumption records in campus card to predict their academic performance and achieved high predicting accuracy. It can be learned from this study that educational counselors could carry out tailored interventions to this relatively small group of the at-risk students early, thus improving students' academic performance and counselors' management efficiency as well. In addition, our study confirmed that lifestyle correlated with academic performance, which suggested that giving instruction and advice about students' lifestyle may be a potential approach for student's academic performance improvement.

Higher institutions are suggested to make full use of existing data like campus card records instead of time-consuming large scale surveys. As suggested from this study, campus card records could be a valuable source of data for educational research. These records are comprehensive and high-dimensional, revealing many facets of student campus life, like eating, doing laundry, buying things in campus shops, etc., sketching a fuller picture of students.

For future applications, this study has set an example to demonstrate that in order to develop an effective model for PSP, historical records should be utilized to provide evidence for students' academic performance predication. Therefore, higher institutions are suggested to develop and maintain a robust data management mechanism which can keep these on-campus behavioral records.

## 5. Conclusion

This study showed that at-risk students can be identified at early stage of a term only from their campus card usage records. The prediction model presented in this study has confirmed the correlation between students' features and their academic performance. In particular,the students who get up early in the morning, go to bed early at night and have a regular lifestyle are less likely to fail the courses. This is consistent with previous studies. These findings may help student counselors and enlighten educational researchers for a better understanding of at-risk students.

It has to be acknowledged that this study has the following limitations. The first is the generalizability of model application. There is a chance that the performance or selected features of the prediction model may change if different data set is investigated. Second, the number of the samples used in this study is not large and may have overfitting risk. A larger data set is suggested for further research. We note that the prediction model constructed in this study may not be suitable for other departments or universities, but the method proposed in this paper can be applied in more educational settings. The data of campus card usage can provide more detailed information than questionnaires and could be more objective in many cases. Besides, it takes no extra time of the investigated students. Therefore, the application of data mining techniques on the data of campus card usage can be seen as an important supplement to questionnaire-based research.

*Acknowledgements*—This research is supported by the National Natural Science Foundation of China (Grant No. 61472464), National Natural Science Foundation Project of CQ CSTC (No. cstc2016jcyjA0276) and the Fundamental Research Funds for the Central Universities (No. 106112015CDJSK04JD02, No. 106112016CDJSK04XK09 and 106112016CDJXY180006).

## References

1. C. Romero and S. Ventura, Data mining in education, *WIREs DataMining Knowl. Discov.*, **3**, pp. 12–27, 2013.
2. H. Y. Kim, E. A. Frongillo, S. S. Han, S. Y. Oh, W. K. Kim, Y. A. Jang, H. S. Won, H. S. Lee and S. H. Kim, Academic performance of Korean children is associated with dietary behaviours and physical status, *Asia Pacific Journal of Clinical Nutrition*, **12**(2), pp. 186–192, 2003.
3. T. H. Stea and M. K. Torstveit, Association of lifestyle habits and academic achievement in Norwegian adolescents: a cross-sectional study, *BMC Public Health*, **14**(1), pp. 1–8, 2014.
4. H. Tanaka, Sleep, lifestyle and academic performance and sleep education by using cognitive behavioral method, *ICME International Conference on Complex Medical Engineering*, pp. 754–757, 2012.
5. G. Liang, Y. Gao, Z. Wu, W. Zhang, S. Xie, L. Teng and S. Zhang, The association between students' living styles and academic achievement, *Journal of Xinxiang University*, **31**(8), pp. 63–65 (Chinese), 2014.
6. F. Araque, C. Roldán and A. Salguero, Factors influencing university drop out rates, *Computers & Education*, **53**, pp. 563–574, 2009.
7. J. P. Vandamme, N. Meskens and J. F. Superby, Predicting Academic Performance by Data Mining Methods, *Education Economics*, **15**(4), pp. 405–419, 2007.
8. S. R. Ting, Predicting Academic Success of First-Year Engineering Students from Standardized Test Scores and Psychosocial Variables, *International Journal of Engineering Education*, **17**(1), pp. 75–80, 2001.
9. C. Romero, M.-I. López, J.-M. Luna and S. Ventura, Predicting students' final performance from participation in on-line discussion forums, *Computers & Education*, **68**, pp. 458–472, 2013.
10. M. D. Calvo-Flores, E. G. Galindo, M. C. P. Jiménez and O. P. Piñeiro, Predicting students' marks from Moodle logs using neural network models, *Current Developments in Technology-Assisted Education*, pp. 586–590, 2006.
11. T. Abdel-Salam, P. Kaufftnann and K. Williamson, A case study: do high school GPA/SAT scores predict the performance of freshmen engineering students?, *Frontiers in Education Proceedings Conference*, 2005.
12. S. Huang and N. Fang, Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models, *Computers & Education*, **61**(5), pp. 133–145, 2013.
13. W. H¨Am¨AlÄinen and M. Vinni, Classifiers for educational data mining, *London: Chapman & Hall/CRC*, 2011.
14. S. Kotsiantis, K. Patriarcheas and M. Xenos, A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education, *Knowledge-Based Systems*, **23**, pp. 529–535, 2010.
15. C. Romero, P. G. Espejo, A. Zafra, J. R. Romero and S. Ventura, Web usage mining for predicting final marks of students that use Moodle courses, *Computer Applications in Engineering Education*, **21**(1), pp. 135–146, 2013.
16. G. Grudnitski, A forecast of achievement from student profile data, *Journal of Accounting Education*, **15**(4), pp. 549–558, 1997.
17. A. M. Gansemer-Topf, J. Compton, D. Wohlgemuth, G. Forbes and E. Ralston, Modeling Success: Using Preenrollment Data to Identify Academically At-Risk Students, *Strategic Enrollment Management Quarterly*, **3**(2), pp. 109–131, 2015.
18. A. Wald, P. A. Muennig, K. A. O'Connell and C. E. Garber, Associations between healthy lifestyle behaviors and academic performance in U.S. undergraduates: a secondary analysis of the American College Health Association's National College Health Assessment II, *Am. J. Health Promot.*, **28**(5), pp. 298–305, 2014.
19. M. Sabalberino-Apilado, Relationship between parental strictness, lifestyle and academic performance of college students, *International Journal of Current Research*, **7**(5), pp. 15865–15870, 2015.
20. E. L. Faught, D. Gleddie, K. E. Storey, C. M. Davison and P. J. Veugelers, Healthy lifestyle behaviours are positively and independently associated with academic achievement: An analysis of self-reported data from a nationally representative sample of Canadian early adolescents, *Plos One*, **12**(7), pp. e0181938, 2017.
21. M. Heidari, M. B. Borujeni, M. G. Borujeni and M. Shirvani, Relationship of Lifestyle with Academic Achievement in Nursing Students, *J. Clin. Diagn. Res.*, **11**(3), pp. JC01, 2017.
22. E. O. Babatunde, Influence of Health Education and Healthy Lifestyle on Students' Academic Achievement in Biology in Nigeria, *Universal Journal of Educational Research*, **5**(9), pp. 1600–1605, 2017.
23. M. O. A. Elkader and F. A. Mohammad, The Relationship between Lifestyle, General Health & Academic Scores of Nursing Students, *Public Health Research*, **3**(3), pp. 54–70, 2013.
24. M. Valladares, E. Durán, A. Matheus, S. Durán-Agüero, A. M. Obregón and R. Ramírez-Tagle, Association between Eating Behavior and Academic Performance in University Students, *Journal of the American College of Nutrition*, **35**(8), pp. 1–5, 2016.
25. H. Sampasakanyinga and H. A. Hamilton, Eating breakfast regularly is related to higher school connectedness and academic performance in Canadian middle- and high-school students, *Public Health*, **145**, pp. 120–123, 2017.
26. X. Peng, X. Tang and S. Ma, The Relationship between Health-related Behavior and Academic Performance of College Students, *China Journal of Health Psychology*, 2012.
27. B. Blai, Jr, Some Biochemical Correlates of Academic Achievement (College Women–Their Eating Habits and Academic Achievement), *Academic Achievement*, (1), p. 7, 1975.
28. G. Curcio, M. Ferrara and G. L. De, Sleep loss, learning capacity and academic performance, *Sleep Medicine Reviews*, **10**(5), pp. 323–337, 2006.
29. M. T. Trockel, M. D. Barnes and D. L. Egget, Health-related variables and academic performance among first-year college students: implications for sleep and other behaviors, *Journal of American College Health*, **49**(3), pp. 125–131, 2000.
30. H. M. Abdulghani, N. A. Alrowais, N. S. Bin-Saad, N. M. Al-Subaie, A. M. A. Haji and A. I. Alhaqwi, Sleep disorder among medical students: Relationship to their academic performance, *Medical Teacher*, 34 Suppl 1(s1), pp. S37–41, 2012.
31. K. K. Mak, S. L. Lee, S. Y. Ho, W. S. Lo and M. D. Tai-Hing Lam, Sleep and Academic Performance in Hong Kong Adolescents, *Journal of School Health*, **82**(11), pp. 522–527, 2012.
32. M. Hysing, A. G. Harvey, S. J. Linton, K. G. Askeland and B. Sivertsen, Sleep and academic performance in later adolescence: results from a large population-based study, *Journal of Sleep Research*, **25**(3), pp. 318–324, 2016.
33. H. Liu and L. Yu, Toward integrating feature selection algorithms for classification and clustering, *IEEE Transactions on Knowledge and Data Engineering*, **17**(4), pp. 491–502, 2005.
34. I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, **3**, pp. 1157–1182, 2003.
35. U. B. Kjærulff and A. L. Madsen, Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, *Information Science & Statistics*, **22**(487), pp. 1273–1274, 2007.
36. A. L. Duckworth and M. E. P. Seligman, Self-discipline outdoes IQ in predicting academic performance of adolescents, *Psychol Sci*, **16**(12), pp. 939–944, 2005.
37. K. R. Jung, A. Q. Zhou and R. M. Lee, Self-efficacy, self-

discipline and academic performance: Testing a context-specific mediation model, *Learning & Individual Differences*, **60**, pp. 33–39, 2017.
38. R. Hogan and D. S. Weiss, Personality correlates of superior academic achievement, *Journal of Counseling Psychology*, **21**(2), pp. 144–149, 1974.
39. J. P. Tangney, R. F. Baumeister and A. L. Boone, High self-control predicts good adjustment, less pathology, better grades, and interpersonal success, *Journal of Personality*, **72**(2), pp. 271–324, 2004.
40. N. O. Simba, Impact of Discipline on Academic Performance of Pupils in Public Primary Schools in Muhoroni Sub-County, Kenya, *Journal of Education & Practice*, **7**, pp. 164–173, 2016.

**Wenjun Quan** received his MA degree of engineering in 2008, and now is pursuing his doctoral degree in College of Computer Science in Chongqing University, China. His research interests include data mining, machine learning, big data and pattern recognition.

**Qing Zhou** is a professor at college of computer science in Chongqing University. He received the BSc degree of engineering in 2002, the MSc degree in 2005, and his doctor degree of engineering in 2008, all in College of Computer Science in Chongqing University in China. His research interests include: data mining, machine learning, big data, cloud computing and information security.

**Yu Zhong** received her MA degree of foreign linguistics and applied linguistics in 2013, and now is pursuing her doctoral degree of computer science and technologies in Chongqing University, China. Her research interests include educational data mining (particularly in the field of foreign language education) and language assessment.

**Ping Wang** is pursuing her doctoral degree of computer science and technologies in Chongqing University, China. Her research interests include educational data mining and language assessment.