Prediction of Academic Achievement Based on Digital Campus

Zheng Wang
Key Laboratory of Universal Wireless Communications for Ministry of Education
Beijing University of Posts and Telecommunications
Beijing,China
616016323@qq.com

Xinning Zhu
Key Laboratory of Universal Wireless Communications for Ministry of Education
Beijing University of Posts and Telecommunications
Beijing,China
zhuxn@bupt.edu.cn

Junfei Huang
Institute of network technology,Beijing University of Posts and Telecommunications
Beijing University of Posts and Telecommunications
Beijing,China
1465422103@qq.com

Xiang Li
Key Laboratory of Universal Wireless Communications for Ministry of Education
Beijing University of Posts and Telecommunications
Beijing,China
674904403@qq.com

Yang Ji
Key Laboratory of Universal Wireless Communications for Ministry of Education
Beijing University of Posts and Telecommunications
Beijing,China
jiyang@bupt.edu.cn

ABSTRACT

Academic achievement of a student in college always has a far-reaching impact on his further development. With the rise of the ubiquitous sensing technology, students' digital footprints in campus can be collected to gain insights into their daily behaviours and predict their academic achievements. In this paper, we propose a framework named AAP-EDM (Academic Achievement Prediction via Educational Data Mining) to predict students' academic achievements based on the influencing factors we have discovered. Multi-source heterogeneous data including Wi-Fi detection records, usage of smartcards, usage of campus network, is aggregated firstly. Then, instead of the self-reported features or traditional academic assessments like test scores, we extract features reflecting students' behavioural patterns. Specially, we define **DOH** (Degree of Hardworking) to improve the performance of the classifier. Finally, we analyze the features extracted and apply supervised learning methods to predict their academic achievements. Experiments are conducted on real-world data from 528 college students in one faculty, and the classification accuracy can be up to 88%.

Keywords

Digital footprints, academic achievement prediction, multi-source data merging, supervised learning, behavioural pattern

1. INTRODUCTION

Predicting students' academic achievements is one of the most popular applications in Educational Data Mining. One research predicted students' academic achievements by analyzing students' static information such as gender, character, eating habits and place of residence.[2]. Authors used predictive modeling methods to identify at-risk students in a course using standards-based grading.[5]. Authors found that students' achievements were best inferred from their social ties through modified smartphones.[4]. Researchers demonstrated the impact of students' psychology in predicting their academic achievements using examination scores, information processing abilities as features [3]. Under the circumstance of online learning, researchers predicted 145 students' academic achievements utilizing their online learning activities and online discussion forums [7, 8]. There are also authors who used passive sensing data and self-reports from students' smartphones and proposed a model based on linear regression with lasso regularization to predict **GPA** [9].

Our study is conducted to make up for the two shortcomings in the previous studies. On the one hand, compared with standard academic assessments or personal static information, students' daily behaviours which can be monitored anytime can reflect their states of living and learning more sensitively and timely. Past research has shown that students' academic achievements have relationships with their daily behaviours [9]. We inspect students' behaviours by analyzing their trajectories, class schedule, campus network usage and smartcard usage. On the other hand, our study is conducted based on a complete passive detection system with no active participation of students which facilitates continual studies of a larger scale [6, 10]. It is important to mention that we care about the privacy protection very

much and all of students' information involved in the study is anonymous.

In this paper, we propose a framework named **AAP-EDM** (Academic achievement prediction via educational data mining) to analyze data generated from digital campus in order to predict students' academic achievements. The framework contains mainly three main modules. Multi-source heterogeneous data merging is the first. After that, we extract features such as wake-up time, duration of stay in the dormitory, and class attendance. We discovered the potential influencing factors of academic achievements through ANOVA F-test and correlation coefficients analysis. Furthermore, we defined the feature **DOH** (Degree of Hardworking) to consider the features we have extracted comprehensively. Then, we formalized the prediction as a binary classification problem to identify students at risk and choose the best solution from multiple classification algorithms consisting of SVM, Logistic Regression,Naive Bayes and Decision Tree. Finally, we evaluated the proposed framework over a real-world dataset involving 528 undergraduates, and found that the classification accuracy can be up to 88%.

Our main contributions in this paper are listed below:

(1) We predicted students' academic achievements utilizing students' daily life behaviour data rather than using academic assessments such as test scores. The high accuracy rate indicates that students' academic achievements have strong relationships with their daily behaviours.

(2) We extracted abundant features which can describe students daily life in detail and also define the **DOH** which improves the performance of classifiers.

(3) In order to explore students' behaviour patterns extensively, we came up with methods to fuse the multi-source heterogeneous data of college students. Our research can be easily expanded to much larger scale.

2. PROBLEM FORMULATION

Our raw data consists of four components. First, students' usage of campus network is monitored in real time. Then when students use their smartcards on campus such as when having breakfast and going shopping, their behaviours will also be captured. Moreover, through the Wi-Fi monitors we deployed in the entrance of particular places in the campus, Wi-Fi packets from students' smartphones with Wi-Fi enabled can be captured when they pass by the monitors without connecting to the network. Besides the three parts above, we have static data including students' class schedules and academic achievements. We will introduce the data set in detail in the next section. Based on the data, our target is to extract features of students and train models utilizing supervised learning algorithms to predict academic achievements.

Formally, given the input matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ where M represents the total number of students and N is the number of features which will be introduced later and the academic achievements labels matrix $\mathbf{Y} \in \mathbb{R}^{M \times 1}$, our target is to learn the function which satisfies $\mathbf{Y} = f(\mathbf{X})$. Note that the labels in our study are either 0 or 1 where 0 represents good performers and 1 represents students at risk.

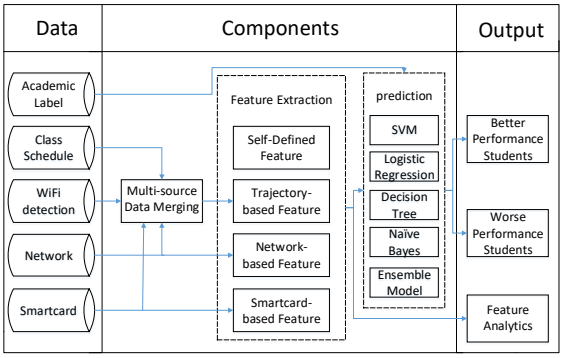


Figure 1: Overview of the framework

Table 1: Data format of Wi-Fi detection records.

MAC Address	Time	RSSI	Location
38BC******91	20160301 12:20:23	-70	Canteen #1

3. METHODS

In this section, we will introduce our framework **AAP-EDM** in detail. The framework mainly contains multi-source heterogeneous data merging, feature extraction and academic achievement prediction which is illustrated in figure 1.

3.1 Multi-source Heterogeneous Data Merging

3.1.1 Raw Data Set

The raw data set contains Wi-Fi detection records, usage of campus network, usage of smartcards, class schedules and also students' academic achievements.

Through deploying Wi-Fi monitors at entrances of locations such as dormitories, canteens and teaching buildings, it is possible to detect smartphones' MAC addresses, providing a coarse-grained location trace for students who enter the coverage area of Wi-Fi monitors which is shown in Table 1.

Students' information of using campus network is shown in Table 2. Specifically, the locations where students access the network (building-level) can be inferred from the "IP Address", and the "Network Traffic" describes the traffic between login time and logout time in MBs, which includes uplink traffic and downlink traffic.

The information of students' devices while connected to the campus network is shown in Table 3. In the table, the "Device Type" can help us distinguish mobile devices from PC and the "Time" is recorded in days but not seconds compared with Table 1.

Table 4 demonstrates the usage of smartcards. The "Consumption Type" includes "Repast", "Shopping", "Bathing",

Table 2: Data format of usage of campus network

Anonymous ID	IP Address (Location)	Login/ logout Time	Network Traffic
E416**B2ED	10.210.**.***	20160301 08:00:00/ 20160301 09:00:00	200

Table 3: Data format of device information

MAC Address	IP Address (Location)	Device Type	Time
38BC******91	10.210.**.**	Mobile	20160301

"Network cost" and so on. Note that the consumption type will reveal the location where students consume with their smartcards.

Other than the data mentioned above, in this paper we also utilize students' class schedules to analyze students' class attendance and utilize students' academic achievements to train the classification model.

3.1.2 Trajectory Generation

We arranged the usage of campus network, the usage of smartcards and Wi-Fi detection records in chronological order to form students' semantic trajectories. In particular, we consider students to stay in the specific location during the periods between the login time and logout time according to campus network records, until records are captured in other locations. The semantic trajectories are shown in Table 5.

3.2 Feature Extraction

3.2.1 Trajectory Features

Daily wake-up time: Wake-up time can reflect the degree of diligence to a certain extent which is calculated as the first time in a day when a student logs in to the network in his dormitory.

Daily time of return to dormitory: Returning to dormitories at a later time in the evening usually means longer periods students spend in the classrooms or the library. We regard the last time in a day when a student logs in to the network in his dormitory as the time of return to dormitory.

Daily duration spent in the dormitory: Dormitories are usually not appropriate places for studying. We can estimate the duration of time spent in dormitory according to the time that students enter and leave the dormitory. Specially, only the time between 06:30 and 23:30 is under consideration.

Table 4: Data format of usage of smartcard

Anonymous ID	Time	Cost	Consumption Type (Location)
E416**B2ED	20160301 08:00:00	5.0	Repast

Table 5: Example of a semantic trajectory in one day

Id	Time	Location
1	07:30:00	Dormitory #13
2	07:33:14	Canteen #1
3	08:21:52	Teaching Building #3
4	11:49:39	Canteen #2
5	12:50:58	Dormitory #13
6	18:03:58	Canteen #2
7	18:35:34	Dormitory #13
8	20:39:16	Teaching Building #2
9	22:08:56	Super Market
10	22:15:15	Dormitory #13

Class Attendance: Given the daily trajectory $\{p_0 \rightarrow p_1 \rightarrow ... \rightarrow p_n\}$ where $p_n = (loc, time)$, the start time t_s and the end time t_e of the course according to class schedules, we will judge whether a student attends the class. Considering that students must appear in the classrooms and shouldn't have any records in other irrelevant places during the class, we propose the method according to two conditions. Eq.1 ensures that students have no records except in classrooms during the class periods. Eq.2 ensures that students are indeed in the classrooms.

$$\{p|t_s + \Delta t < time < t_e - \Delta t, loc \neq classroom\} = \emptyset \quad (1)$$
$$\{p|t_s - \Delta t < time < t_e + \Delta t, loc = classroom\} \neq \emptyset \quad (2)$$

Days outside of campus: Students who have no digital footprints in one day will be considered as not on campus. Students' academic achievements are supposed to be affected if they are often not on campus.

3.2.2 Network Features

Daily Network Traffic in Dormitory: We sum up the network traffic that students upload and download in their dormitories. Compared with dormitories, the network traffic in teaching buildings is less, so we don't take this part into consideration.

Network Cost: Students don't need to pay for the campus network until their used traffic exceeds the upper limit of every month. The upper limit of network traffic is almost enough for normal usage, so students who exceed the limit may spend too much time on the internet accessing online videos or online games. We calculate the total network charges of each student.

Network top up Frequency: When the balance of students' network accounts is zero, students should recharge for continual usage.

Daily Network Traffic Peak: Daily network traffic peak is demonstrated as $L = \{l_0, l_1, ..., l_{23}\}$ where l_n represents an hour in a day and takes value of 0 or 1 shown in Eq.3 where $traffic_n$ is the traffic during the n_{th} hour and the $average$ is the average traffic per hour in one day.

$$l_n = \begin{cases} 1, & traffic_n >= average \\ 0, & traffic_n < average \end{cases} \quad (3)$$

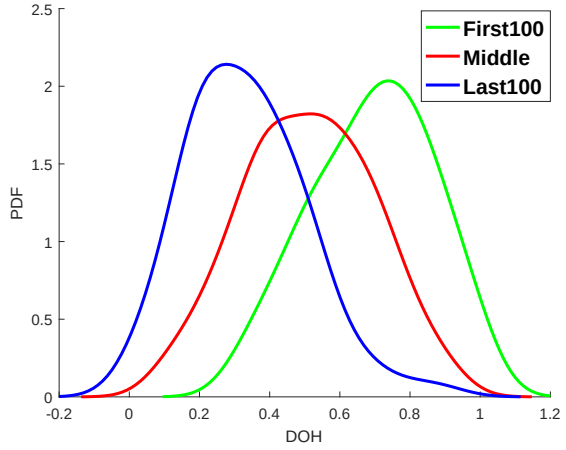


Figure 2: Probability density function of DOH

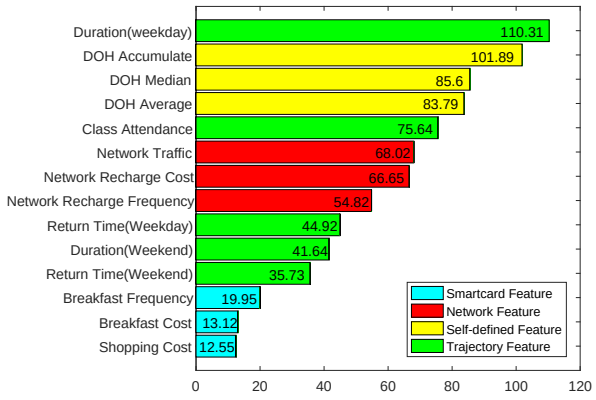


Figure 3: ANOVA F-test for binary classification

3.2.3 *Smartcard Features*

Students' consumption patterns are captured according to the usage of smartcards. In the campus, students will use their smartcards when having meals in the canteens, shopping in the supermarket and taking a shower in the bathhouse. Cumulatively, we calculate students' daily costs and frequency of consumption of breakfast, shopping and bathing.

3.2.4 *Self-defined Features*

In order to obtain a comprehensive evaluation of all features extracted above, we calculated the score of each feature for each student Eq.4. $Corr(X_k)$ is the Pearson correlation coefficient between the k_{th} feature X_k and student's academic achievements which is shown in Table 6. Note that the academic achievements are in the form of rankings when calculating the Pearson correlation coefficient. $Rank(x_n)$ means the ranking of the student u_n' features among N students. For example, there are three students (u_1, u_2, u_3), and their i_{th} feature (class attendances) are $(0.8, 0.5, 0.6)$, we have $Score_i^1 = 1, Score_i^2 = 0.66, Score_i^3 = 0.33$ because $Corr(X_i) < 0$ according to Table 6.

Then we defined the degree of hardworking(**DOH**) utilizing the feature scores Eq.5 where K is the count of all features we have extracted. We plot the probability density function of **DOH** (Min-Max normalized) of three groups of students separated by their rankings of academic achievements as shown in Figure 2. From the figure we can find that the distributions of DOH are similar to the normal distribution and the averages are approximately 0.2, 0.5 and 0.8. The apparent distinction among three groups proves that our defined feature is a strong factor for prediction. Essentially the DOH is the weighed mean of feature scores and the weighs are the correlation coefficients. Besides DOH , self-defined features also include other statistics characteristics of feature scores such as average and median.

$$Score_k^n = \begin{cases} (N - Rank(x_n))/N, & Corr(X_k) > 0 \\ Rank(x_n)/N, & Corr(X_k) < 0 \end{cases} \quad (4)$$

$$DOH^n = \sum_{k=1}^{K} (|Corr(X_k)| * Score_k^n) \quad (5)$$

3.3 Academic Achievement Prediction

We separate the whole semester into four periods, the first three periods last for four weeks respectively and the last one lasts for six weeks. We calculate the mean of daily features respectively in four periods due to the fact that students' behaviours may change along with the whole semester and generate different impacts on their academic achievements. Moreover, it is necessary to distinguish weekdays and weekends in each period for different behavioural patterns.

The academic achievement prediction is essentially a binary classification problem which can be used in academic precaution. For that the values of features vary greatly, in order to increase the speed of gradient descent and the accuracy of classifiers, we limited all the feature values to the range of 0 to 1 using Min-Max normalization. We have 100 students who performed the worst according to their school reports to be positive labels and other 428 students to be negative labels. The dataset is split into training set and test set according to the ratio of 7:3.

There might be relevancies among features which will decrease the performance of classifiers. For example, students who spend long time surfing the campus network can possibly bear high network charges. In this work, we implement the state-of-the-art methods, Principal Component Analysis, to solve this problem.

We trained various classification models such as Logistic Regression, Support Vector Machine, Naive Bayes and Decision Tree using cross-validation and evaluated on the test set. Moreover, we implemented the voting classifier to combine conceptually different machine learning classifiers and use a majority vote to predict the class labels.

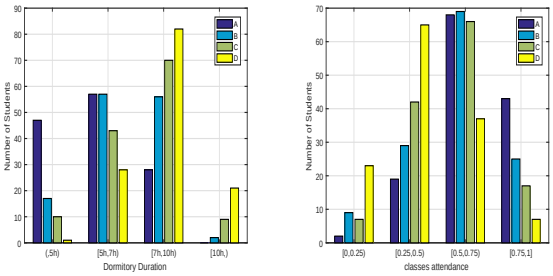
4. EXPERIMENTAL RESULTS

4.1 Experimental Data

We collect 1673706 records totally of 528 undergraduates in their third year from 19 classes in one faculty. The period we selected lasted for a complete semester of 140 days from Feb. 29^{th}, 2016 to Jul. 17^{th}, 2016. The academic achievements

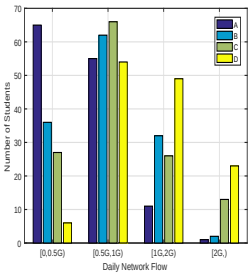
Table 6: Correlation Coefficient and P-value

Feature	Correlation coefficient	P-value
Class attendance	-0.430	3.39e-25
Time spent in dormitory(Weekday)	0.565	7.71e-46
Time spent in dormitory(Weekend)	0.411	5.84e-23
Time of return to dormitory(Weekday)	-0.394	4.22e-21
Time of return to dormitory(Weekend)	-0.348	1.60e-16
Wake-up time(Weekday)	0.222	2.69e-7
Wake-up time(Weekend)	0.204	2e-6
Shopping cost	0.215	6.09e-7
Breakfast Frequency	-0.337	1.9e-15
Breakfast cost	-0.266	5.55e-10
Days out of campus	0.068	0.117
Network traffic	0.406	2.11e-22
Network cost	0.362	8.3e-18
Network top up frequency	0.361	1.02e-17
Feature score average	-0.551	3.3e-43
Feature score median	-0.547	1.64e-42
DOH	-0.561	3.84e-45

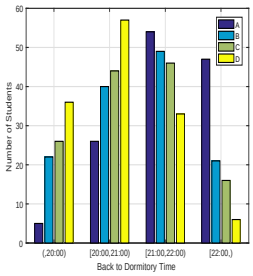


(a) Duration in Dormitory

(b) Class Attendance



(c) Daily Network Traffic



(d) Returning Dormitory Time

Figure 4: Features statistics among different grades

are the weighted average scores of all courses in a semester which includes quizzes, midterns and finals.

What we should emphasize is that our experiment was completely conducted under an anonymous situation. In our experiment, students' IDs which reveal the true identities were mapped to anonymous IDs.

4.2 Feature Analytics

It is meaningful for educators to find out how students' behaviours influence the academic achievements. We performed correlation coefficient analysis and ANOVA F-test to compare different features' contributions to students' study. The correlation coefficients are shown in Table 6. Note that time spent in dormitory on weekdays reaches the highest value of 0.565 with smallest P-value which is a novel finding. Our self-defined features also reach high correlation coefficients. Many previous studies have shown that class attendance is a significant and positive predictor of academic achievements which is also true in our study. Specifically, the self-defined features indicate high correlation coeffiients which is also proved in the ANOVA F-values for binary classification shown in Figure 3. Thus it can been seen that our proposed method for new features is effective which will improve the performance of the prediction. Other than the self-defined features, the overall F-values of network features are relatively high while the smartcard features are slightly irrelevant. Note that the wake-up time and the days leaving campus which don't achieve sufficient significance ($p \geq 0.001$) are omitted in the Figure 3.

To observe the differences of behaviours among students in detail, we display the distributions of four features which are highly relative with academic achievements in Figure 4. We divide all the students into four groups in the order of their academy achievements. Group A represents the best performers and group D represents the worst performers.

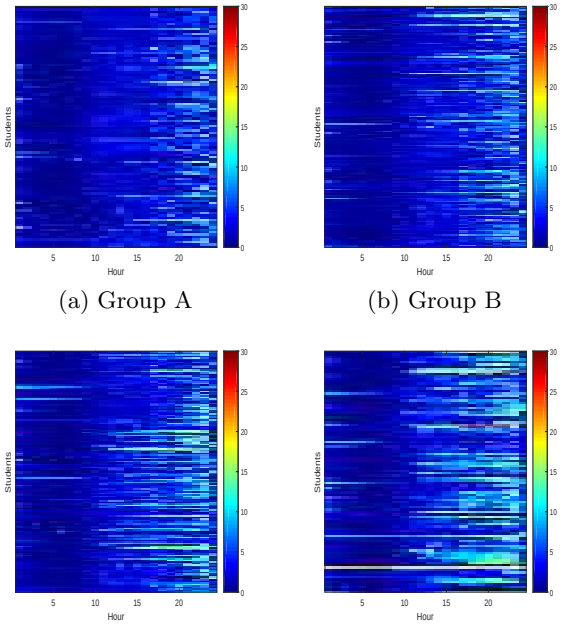
As we can see in subgraph Figure 4a, more than 70% students of group A spend less than 7 hours in dormitories. On the contrary, most students in group C and D stay in dormitories for longer than 7 hours, some even staying for more than 10 hours. In subgraph Figure 4b, we find that class attendance is mainly distributed from 0.5 to 0.75 except group D in which more than 60% students' attendance is less than 0.5. Nearly 90% students of group A have a high attendance rate. Whether class attendance has influence on academic achievements is controversial.[9, 1] We discover that it is a relatively strong factor in our research. Daily network traffic is shown in subgraph Figure 4c, it is obvious that more than 90% students spend less than 1 GB traffic daily in group A. Bad performers may spend more time for online gaming and

Table 7: Classification Results

Model	Class0 Precision	Class0 Recall	Class0 F1-score	Class1 Precision	Class1 Recall	Class1 F1-score	Accuracy
SVM	0.92	0.86	0.89	0.55	0.69	0.61	0.82
SVM(PCA)	0.87	**0.97**	**0.92**	0.78	0.44	0.56	0.86
LR	0.92	0.77	0.84	0.45	0.75	0.56	0.77
LR(PCA)	0.88	**0.97**	**0.92**	**0.79**	0.47	0.59	0.87
NB	0.92	0.71	0.80	0.39	0.75	0.52	0.72
NB(PCA)	0.87	0.96	0.91	0.72	0.41	0.52	0.85
DT(PCA)	0.91	0.93	**0.92**	0.73	0.69	0.71	0.87
SVM+LR(PCA)	**0.94**	0.91	**0.92**	0.69	**0.75**	**0.72**	**0.88**

movies which results in more network traffic. Subgraph Figure 4d shows students' time of return to dormitory. The left two groups of data tend to show an ascending trend while the right ones show a descending trend which depict that most students of group A and B come back to dormitories after 21:00 and are therefore more diligent.

Figure 5 shows the distribution of students' daily network rush hours in one month. The horizontal axis represents the 24 hours in one day. The vertical axis represents students in the specific group according to academic achievements. Each student is represented by a row vector ($\boldsymbol{v} \in \mathbb{R}^{24}$) accumulated in one month according to Eq.3. The color bar shows the numbers in vectors which are between 0 and 30 (30 days in one month). Therefore, the brighter areas mean students always spend more time online during the specific periods. From the figure we can see, students of group A and B have a shorter span of rush hours and they always login the network near to 22:00 after they come back from classrooms, while rush hours of students of group C and D last for a longer time from about 15:00 to 23:00.



(a) Group A (b) Group B

(c) Group C (d) Group D

Figure 5: Daily network rush hours

4.3 Results of Prediction

In our research the prediction task is an unbalanced classification problem. According to students' academic achievements, the dataset is composed of 428 good performers (negative samples) and 100 bad performers (positive samples). We conducted four different supervised learning algorithms consisting of Support Vector Machine, Logistic Regression, Decision Tree and Naive Bayes. The highest classification accuracy can be up to 88%. However it is not convincing enough for unbalanced classification problems to just inspect the classification accuracy. In this paper, we used precision, recall and F1-score to evaluate the performance of our models. The average classification results of 10-Fold cross validation are shown in Table 7. Specially we ensemble the Support Vector Machine and Logistic Regression through voting classifier and realize the highest accuracy 88%. The principle of the voting classifier is that the students are classified as negative samples when the two classifiers conflict with each other.

5. CONCLUSIONS

In this paper, we predicted that students' academic achievements to identify students who perform worse in their study based on our proposed framework **AAP-EDM**. Firstly, multi-source heterogeneous data is merged to generate semantic trajectories. Then we extracted features consisting of trajectory features, network features and smartcard features. Furthermore, self-defined features are proposed to explore features comprehensively. At last, we have evaluated the framework through multiple classification models using students' real world data. The results show that our proposed framework is feasible and meaningful for educational supervision and warning. Our research provides promising approaches to transform the collage education from traditional descriptive analytics to predictive analytics. We will improve our framework through further research and concentrate on realizing the prescriptive analytics in college education.

6. ACKNOWLEDGMENTS

This paper is supported by "the Fundamental Research Funds for the Central Universities",(No.2018XKRK03).

7. REFERENCES

[1] J. Brocato. How much does coming to class matter? some evidence of class attendance and grade performance. *Educational Research Quarterly*,

19(3):2–6, 1989.

[2] T. Devasia, V. T P, and V. Hegde. Prediction of students performance using educational data mining. In *Data Mining and Advanced Computing (SAPIENCE), International Conference on*, 2016.

[3] R. R. Halde, A. Deshpande, and A. Mahajan. Psychology assisted prediction of academic performance using machine learning. In *IEEE International Conference On Recent Trends In Electronics Information Communication Technology*, 2016.

[4] V. Kassarnig, A. Bjerre-Nielsen, E. Mones, S. Lehmann, and D. Dreyer Lassen. Academic performance and behavioral patterns. Website, 2017. https://arxiv.org/abs/1706.09245.

[5] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan. Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103:1–15, 2016.

[6] A. Musa and J. Eriksson. Tracking unmodified smartphones using wi-fi monitors. In *SenSys '12 Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, pages 281–294, 2012.

[7] A. Pardo, F. Han, and R. A. Ellis. Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies*, 10(1):82–92, 2016.

[8] C. Romero, M. LÃşpez, J. Luna, and S. Ventura. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, pages 458–472, 2013.

[9] R. Wang, G. Harari, P. Hao, X. Zhou, and C. T. Smartgpa: How smartphones can assess and predict academic performance of college students. In *UBICOMP '15,OSAKA, JAPAN*, 2015.

[10] S. Zhao, Z. Zhao, Y. Zhao, R. Huang, S. Li, and G. Pan. Discovering peopleâĂŹs life patterns from anonymized wifi scanlists. In *Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom)*, 2014.