IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 4, August, 2020, pp. 465-470
Journal homepage: http://iieta.org/journals/ria

Prediction of English Scores of College Students Based on Multi-source Data Fusion and Social Behavior Analysis

Yanxia Zhao[1], Wei Ren[2], Zheng Li[3*]

[1] School of International Studies, Zhejiang Business College, Hangzhou 310053, China
[2] School of E-commerce, Zhejiang Business College, Hangzhou 310053, China
[3] College of Applied Science and Technology, Beijing Union University, Beijing 100101, China

Corresponding Author Email: nbl_zheng@buu.edu.cn

https://doi.org/10.18280/ria.340411

Received: 28 March 2020
Accepted: 16 June 2020

Keywords:
multi-source data fusion, social behavior analysis, machine learning (ML), student score, support vector machine (SVM)

ABSTRACT

Multi-source data fusion is the premise of applying big data technology in specific fields. Inspired by the theory on multi-source data fusion, this paper fuses various data on college students, including motion trajectories, consumptions, and social behaviors, and adopts support vector machine (SVM), a machine learning (ML) classifier to predict the English scores of college students. The behavior trajectories were taken into account, because this type of data represents the social similarity between students. Specifically, the behavioral features of college students were extracted, and subject to principal component analysis (PCA). Based on these features, the correlation between student score and social relation was analyzed, and used to predict the English scores of college students. Experimental results show that our method can accurately reflect the relationship between the social behaviors and course scores of college students.

1. INTRODUCTION

Our society is producing a myriad of data all the time. Taking the smart campus for example, every college student can generate a huge amount of multidimensional data on that campus. Effective data mining is needed to excavate valuable information from the massive data, revealing the potential law of things. This is now a possibility thanks to the latest development of artificial intelligence (AI) and big data mining. For instance, wireless network technology provides ubiquitous information services that fundamentally changes our lifestyle. The location technology based on mobile devices makes it easy to obtain the time and location information of others, enabling us to learn the living habits of others and prewarn the abnormal events based on mobile trajectories.

The course scores of students are the main yardstick of the teaching quality of a college. The main task of every college is to improve the learning quality of its students. Therefore, it is very important to thoroughly analyze student score in the complex learning environment, identify the factors that significantly affect the score, and make accurate prediction and prewarning of abnormal learning behaviors. Considering the colorful campus life and complex social relations of college students, student score could be studied more accurately and comprehensively in the light of their social situation.

This paper firstly fuses the behavioral data of college students, according to the theory on multi-source data fusion, and compares the data sparsity of student trajectories before and after the fusion process. Based on the trajectory data, the social similarity between students was calculated, and the correlation between student score and social relation was analyzed. Finally, the support vector machine (SVM), a machine learning (ML) classifier, was adopted to predict the English scores of college students. The effectiveness of the prediction method was proved through experiments.

2. LITERATURE REVIEW

More and more researchers have started to analyze the academic performance of college students based on their daily behaviors. For example, Cao et al. [1] proposed an early warning and intervention model of student score through outlier mining. Firstly, the enrollment data (i.e. current learning level, family background, and physical condition) of students were collected, together with the extracurricular learning data (e.g. attendance) indirectly related to course learning, and the abnormal features were defined in details. Then, the outlier mining algorithm was introduced to pinpoint the negative outlier objects, and determine the level of early warning based on time sequence and key attribute space. Furthermore, the negative outliers were intervened automatically or manually to promote the learning effect, and the intervention results were fed back timely to the intervention engine to complete early warning.

Considering the features of online teaching, Huang et al. [2] designed a three-layer backpropagation neural network (BPNN), and trained the network with actual scores of college students. The trained network was applied to predict student scores. The results show that the prediction accuracy of the network surpassed 80%. Basak and Krishnapuram [3] employed decision tree (DT) to analyze the learning situation of students: the students were rated against a 6-10 scale according to their course scores, and a classification and prediction model was constructed from the classification and regression trees (CART); taking the collected student scores as the inputs, the proposed model was adopted to predict the

unknown academic performance of college students. Sun and Bin [4] measured the behavioral differences between students with good and poor academic performance, using the data from online learning platforms.

Sathick and Jaya [5] collected the information about the grade point average (GPA) of college students, such as attendance, learning level, social situation, geographical location, sleep duration, call record, and indoor activity frequency, and predict their academic performance by a linear regression model with a sparse operator. Pujianto et al. [6] recognized the value of learner feelings in the research of student behaviors and learning effect, and advised to understand the state and attitude of learners by analyzing learner feelings. Therefore, several volunteers were asked to express their feelings about various courses, and their psychological changes before, during, and after class of several volunteers were recorded. Then, two topic models, namely, probabilistic language analysis and latent Dirichlet distribution, were selected to find the topics that help to improve the prediction of student scores, and applied to predict the course scores of college students.

Salanova et al. [7] constructed a personalized analysis model for college students, drawing on the individual features of college students and the big data from the smart campus system. After analyzing education big data, Wu [8] built up a personalized adaptive online learning analysis model, which covers data, environment, stakeholders, methods, and objectives. The model provides students frequently engaged in learning activities with adaptive and personalized learning schemes. Hwang et al. [9] designed a learning analysis system with social behavior analysis as the core, and provided intelligent digital education services based on the social behaviors of students. Schroeder et al. [10] summarized that the data-based analysis methods of academic performance usually take one of the five perspectives: teaching and learning process, teaching resources, learner network, learner features, and learner behaviors/emotions.

3. MULTI-SOURCE DATA ACQUISITION AND FUSION

3.1 Data acquisition

The mainstream method to obtain user trajectories is to extract the relevant data from mobile devices. The trajectory data are mainly distributed in global positioning system (GPS), cellular mobile network, and wireless local area network (WLAN) [11].

On smart campus, the various wireless access points (APs) can be regarded as Wi-Fi probes to collect the data from mobile devices with Wi-Fi access function. Under the IEEE802.11 protocol, the data collected by Wi-Fi probes contain three fields: media access control (MAC) address, received signal strength indicator (RSSI), and timestamp of the received signal [12]:

$$< MAC, Location\_ID, Timestamp, RSSI > \quad (1)$$

Each Wi-Fi probe has its own fixed location, from which the location of the mobile device can be deduced. The trajectories of college students on campus could be obtained, because each MAC address corresponds to a user of mobile device.

The data collected by each Wi-Fi probe are sent to the server in real time. From the college gateway server, the daily Internet access of each student to the campus network can be obtained:

$$< Student_{ID}, IP_{address}, Login_{time}, Logout\_time, Flow > \quad (2)$$

where, *Student_ID* is student number; *IP_address* is the location that the student logs into the campus network; *Login_time* and *Logout_time* are the time that the student logs into and out of the campus network, respectively; *Flow* is the data flow consumed by the student online.

In addition, the hardware information of each mobile device logging into the campus network can be obtained:

$$< MAC, IP_{address}, Device\_name, Device\_type > \quad (3)$$

where, *MAC* is the unique identification of the mobile device; *Device_name* is the alias of the login device; *Device_type* is the type of the login device.

Further, the consumption data of each student can be obtained from his/her campus card:

$$< Student_{ID}, Timestamp, Amount, Consumption\_type > \quad (4)$$

where, *Timestamp* is the consumption time; *Amount* is the consumption amount; *Consumption_type* is the consumption type, including canteen consumption, supermarket consumption, bathhouse consumption, etc. The consumption type, which covers location information, is an important indicator of the daily trajectory of each college student.

The attendance is a key metric of academic performance and learning behaviors. Here, the attendance of each college student is extracted from the information on the courses of the entire semester, including the class time and break time of every mandatory course and elective course. In addition, the score ranking of each student was acquired as tag data to reflect the relationship between student score and learning behavior:

$$< Student\_ID, Grade\_ranking > \quad (5)$$

where, *Student_ID* is the student number; *Grade_ranking* is the score ranking of a student.

3.2 Data processing and multi-source fusion

Once a student enters its detection range, the Wi-Fi probe will detect the mobile device which has turned on the Wi-Fi switch. Because the detection signal is emitted at a certain frequency, several detection records might be generated for the same device in less than a second. Such a small interval is beyond the need for tracking student trajectories. Hence, the data collected by each Wi-Fi probe are inevitably redundant. The redundant data occupy a large storage space, and generate unnecessary computing load. To eliminate data redundancy, the multiple device records detected over 6s at the same location were merged into a single record.

Concerning the MAC address, cellphones were considered as the better mobile device for reflecting student location than tablet computers and personal computers (PCs), due to its

relatively high portability. Thus, the data on fixed devices collected by Wi-Fi probes were deemed as invalid.

To obtain high-quality trajectory data, Wi-Fi probe data were fused with the data from campus gateway server and campus cards. All three kinds of data can reflect the mobile location of the students. Together, the fused data could overcome the sparsity of trajectory data and improve the prediction accuracy.

The key of multi-source data fusion lies in determining the mapping between student numbers and mobile devices. Therefore, this paper designs a mapping algorithm between student numbers and MAC addresses, which fully utilizes the function of the Internet protocol (IP) address. The set *ID* of student numbers corresponding to MAC addresses can be expressed as:

$$ID = \{id_1, id_2, id_3, \dots\} \tag{6}$$

According to the data format of acquired data, the IP set corresponding to a MAC address can be defined as:

$$IP\_address = \{ip_1, ip_2, ip_3, \dots\} \tag{7}$$

Based on the corresponding *IP_address*, the set *ID* can be divided into *n* subsets, each of which contains *k* elements. The score of each element in the set *ID* within a period of *t* days can be calculated by:

$$Score_{ID} = \sum_t \sum_n \frac{Num_{ID}}{k} \tag{8}$$

where, Num_{ID} is the number that each element in the set $IP\_address$ appears in each *ID* subset. Finally, the student ID with the highest score was selected to match the MAC address.

Traditionally, data mining generally deals with the data from a single source. In reality, the original data may come from multiple sources in different fields [13]. The conventional multi-source data fusion is comparable to the integration of multiple datasets, which are consistent, accurate, and meaningful. The schema mapping of different datasets is extracted, and the datasets are integrated into a complete large dataset. This conventional strategy cannot effectively fuse the multi-source data from the big data environment, for these data are composed of multiple datasets, which are generated in different fields and connected through the same potential target.

In the big data environment, there are three ways to fuse multi-source data, namely, data fusion based on different scenes, data fusion based on feature level, and data fusion based on sematic learning [14]. This paper chooses the data fusion based on different scenes to merge the datasets of different stages and scenes for data mining. These datasets are loosely coupled and not necessarily unified in form. The Wi-Fi probe data, campus gateway data, campus card data, and other smart campus data were fused, and the daily trajectories of college students were extracted from the fused data. The multi-source data fusion process is explained in Figure 1 below.

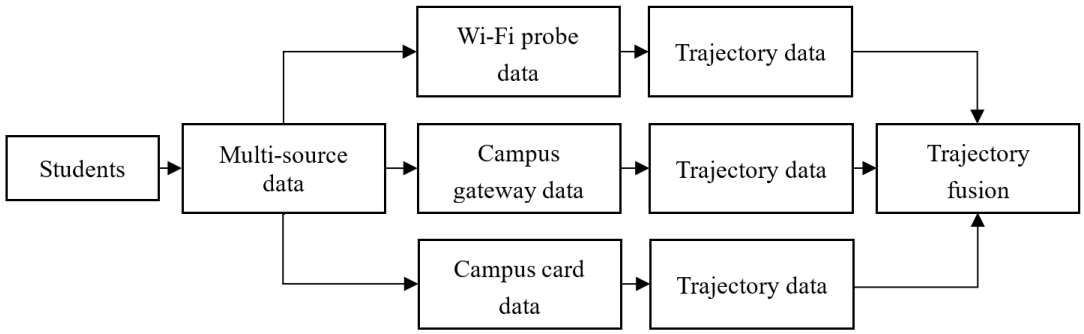

Figure 1. The process of multi-source data fusion

4. ANALYSIS ON THE CORRELATION BETWEEN SOCIAL RELATION AND STUDENT SCORE

4.1 Trajectory fusion

Based on the mapping between student number and MAC address obtained by the self-developed mapping algorithm, the trajectory data of each student can be obtained from Wi-Fi probes:

$$< Time, Loaction > \tag{9}$$

The mapping between each location and the IP of the local probe was acquired from the campus wireless network equipment. The trajectory data of each student thus derived can be expressed as:

$$< Time, Loaction, Login/Logout, IP\_address > \tag{10}$$

Moreover, the student location is indirectly affected by the consumption location contained in campus card data:

$$< Time, Loaction, Type, Amount > \tag{11}$$

The above trajectory data were sorted in chronological order to complete the trajectory data of each student (Table 1).

Table 1. The examples of trajectory data of college students

Time	Location	Login/Logout	IP	Type	Amount
07:22:01	Dormitory 1	Login	10.111.135.76	Null	Null
07:30:18	Dormitory 1	Logout	10.111.135.76	Null	Null
07:30:45	Dormitory 1	Null	Null	Null	Null
07:35:10	Canteen 2	Null	Null	Meal	5
		…….			
16:35:00	Teaching Building 3	Null	Null	Null	Null
17:20:05	Supermarket	Null	Null	Shopping	7.4
18:36:15	Bathhouse	Null	Null	Shower	1
19:30:00	Dormitory 1	Login	10.111.135.76	Null	Null

4.2 Construction of social network based on trajectory data

The longest subsequence common to all sequences in a set of sequences is called a longest common subsequence (LCSS) [15]. The most efficient way to find the LCSS is dynamic programming, which first recursively computes the length of common subsequences and then identifies the common subsequence with the greatest length.

Let $X = \{X_1, X_2, \ldots, X_m\}$ and $Y = \{Y_1, Y_2, \ldots, Y_m\}$ be two sequences. Then, the length $C[m,n]$ of the LCSS between X_m and Y_m can be recursively computed by:

$$C[m,n] = \begin{cases} 0 \quad if\ m = 0\ or\ n = 0 \\ C[m-1, n-1] + 1 \quad if\ m, n > 0,\ \ x_m = y_m \\ \max\{C[m, n-1], C[m-1, n]\} \quad if\ m, n > 0, x_m \neq y_m \end{cases} \tag{12}$$

It can be seen that the LCSS value of (X_m, Y_m) is contained in $C[m,n]$. To quantify the social similarity between students, the concept of social similarity can be defined as:

$$\mathrm{Sim}(stu_1, stu_2) = \frac{2LCSS(stu_1, stu_2)}{(|stu_1| + |stu_2|)} \tag{13}$$

where, *LCSS*() is the LCSS algorithm with continuously adjustable time threshold. The algorithm provides an adaptive tool to calculate the LCSS between two students. By adjusting the continuous time threshold, the LCSS algorithm can derive different types of social relations with high accuracy. For the similarity calculation between couples, the continuous time threshold should be properly lowered; For the similarity calculation between common students, the threshold should be properly increased.

In addition to the LCSS algorithm, the graph theory was introduced to build up the social network of college students. In the social network, each node represents a student, each edge represents the social relation between two students, and the weight of the edge represents the social similarity between the two students.

Some students have higher social similarity than others. Thus, the edge weights below the similarity threshold should be filtered out, along with the corresponding edges. To set a reasonable similarity threshold, some students which are known to be friends were chosen, and the trajectory similarity between every two of them was calculated to obtain an empirical threshold. On this basis, the established social network was binarized into the final social network.

4.3 Correlation analysis

A total of 500 English majors and other students were selected for correlation analysis. First, a trajectory similarity matrix (500×13,885) and a social network were established. Then, the social activity (number of friends) and student score (score ranking) was analyzed.

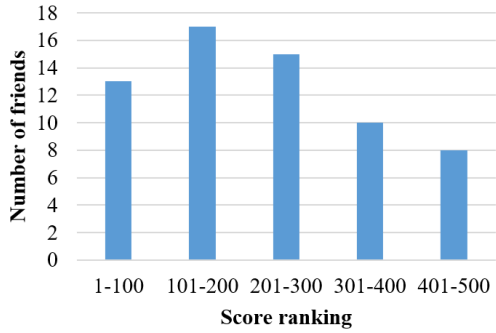

Figure 2. The relationship between social activity and student score

As shown in Figure 2, the students fall into four categories by score ranking. The students with medium scores have a relatively large social circle, i.e. keep many friends. By contrast, the students with high scores have relatively poor social skills.

Next, the social preferences of students in different categories were discussed. The results are recorded in Figure 3.

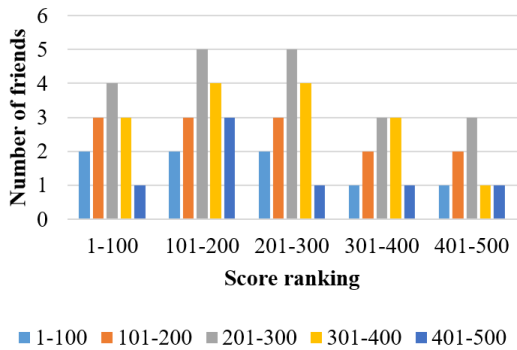

Figure 3. The number friends in different categories

As shown in Figure 3, regardless of the category, the most popular friends are students with medium scores, indicating that the students with medium scores make more friends than those with high or low scores. It can also be seen that, with the decrease of student score, the number of friends with high score gradually drops; with the increase of student score, the number of friends with low score gradually grows.

It is also interesting to examine whether the relation between couples promotes or inhibits the academic

performance of students. For this purpose, the first step is to identify the gender of students. Here, the gender of each student is determined based on the allocation of male and female dormitories on campus. Specifically, the number of days with sleep record in a semester was counted for each student. Then, the location where that number surpasses half of the days in the semester was determined as the dormitory. Finally, the gender of that student was identified based on whether the dormitory is for males or females.

Considering the lifestyle and habits of student couples in college, two students of different genders were considered couples if they have ultrahigh trajectory similarity, often appear in the dormitory of the opposite gender at night, and use the other's account to log in the campus network. Concerning trajectory similarity, the LCSS algorithm was employed to discriminate couples from other students. Because couples are more intimate than ordinary students, the continuous time threshold was set as smaller than or equal to 3min. The campus network account is relatively private. Couples are arguably the only students that frequently borrow each other's account. In the college network, there is a limit on the Internet traffic of each student. If a student uses the account of another student of the opposite gender multiple times (greater than or equal to 5 times), and if he/she does not yet reach the traffic limit, the two students will have a high probability of being couples.

To verify the accuracy of our method and maintain the privacy of students, the results of couple discrimination were validated with the consent of the relevant students. The validation shows that the couple discrimination method achieved an accuracy of 87.1%.

Next, the student scores were divided into five intervals according to the score ranking. Then, the distributions of student couples and ordinary students in each interval were calculated (Figure 4) to disclose the effects of couple relation on students with different scores.

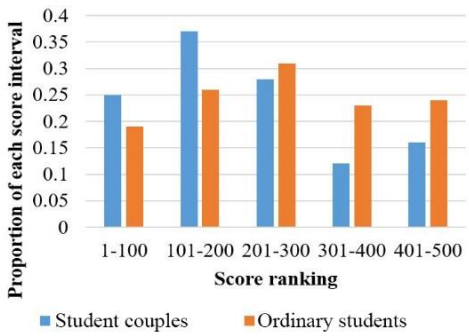

Figure 4. The effects of couple relation on student score

As shown in Figure 4, most students with medium scores are student couples, while virtually no student with low score has a couple. Besides, the majority of students in the first half of score ranking have a couple, while few in the lower half have a couple. To a certain extent, the couple relation has a positive effect on the academic performance of college students.

5. PREDICTION OF ENGLISH SCORES OF COLLEGE STUDENTS

To characterize the learning state of students in different periods of the same semester, each semester was divided into four intervals. Considering the influence of weekends and weekdays in each interval, the features of student scores were split into weekend and non-weekend classes.

Then, the learning behaviors of the students were quantified based on the multi-source data. To present an intuitive picture, the features of student scores were categorized into computational features and statistical features (Table 2).

Table 2. The features of student scores

Computational features	Statistical features
Attendance of compulsory courses	Internet traffic
Total hours spent in the classroom on holidays	Supermarket consumption
Number of visits to classrooms on holidays	Campus network recharge
Number of autonomous learning in the evenings	Social activity
Total hours spent in dormitory on weekdays	Student couple/ordinary student
	Gender

Some of the above original features are redundant and useless. Thus, principal component analysis (PCA) was performed to transform the multiple features into a few representative features through dimensionality reduction. Through the PCA, the original features were compressed into 2 dimensions. The k-means clustering (KMC) results on the original and compressed features are presented in Tables 3 and 4, respectively.

Table 3. The accuracy of KMC on the original features

Score ranking	Cluster 1	Cluster 2
1-250	167	81
251-500	92	156
Accuracy	64.88%	66.27%

Table 4. The accuracy of KMC on the compressed features

Score ranking	Cluster 1	Cluster 2
1-250	152	99
251-500	60	187
Accuracy	71.02%	69.72 %

From Tables 3 and 4, it is learned that feature selection greatly promotes the clustering accuracy of the KMC. The clustering process does not involve the private data on score ranking. However, the discrimination accuracy between students with different scores is not satisfactory. To solve the problem, the SVM classifier was adopted to analyze and predict student scores again, based on the features of student scores obtained through multi-source data extraction and attribute subset selection in a semester (1-16 weeks). Multiple classification criteria were adopted to thoroughly demonstrate the prediction performance. The classification results before and after feature selection are shown in Tables 5 and 6, respectively.

Table 5. The classification results before feature selection

Class	Precision	Recall	F1	Support
1	0.72	0.81	0.76	48
2	0.79	0.68	0.72	52

Table 6. The classification results after feature selection

Class	Precision	Recall	F1	Support
1	0.75	0.87	0.78	54
2	0.82	0.66	0.73	46

As shown in Tables 5 and 6, feature selection can greatly improve the prediction accuracy of the SVM classifier. Besides, the SVM achieved better accuracy than the KMC, providing a suitable method to predict student scores, and prewarn abnormalities in learning behaviors.

6. CONCLUSIONS

Inspired by the theory on multi-source data fusion, this paper collects the data on student trajectories from multiple sources in the campus network, and fuses the collected data based on the Wi-Fi positioning technology, trajectory data mining algorithm, and social network mining algorithm. From the fused data, the daily trajectories of college students were extracted. On this basis, a model was established to analyze the correlation between social relation and student score, and predict the English scores of college students. Experimental results show that our method can accurately identify the said correlation, and predict student scores. The research findings lay a solid basis for follow-up research in education big data.

ACKNOWLEDGMENT

This article is the research result of the project funded by Department of Education of Zhejiang Province (Grant No.: Y201942517).

REFERENCES

[1] Cao, H., Si, G., Zhang, Y., Jia, L. (2010). Enhancing effectiveness of density-based outlier mining scheme with density-similarity-neighbor-based outlier factor. Expert Systems with Applications, 37(12): 8090-8101. https://doi.org/10.1016/j.eswa.2010.05.079

[2] Huang, L., Wang, C.D., Chao, H.Y., Lai, J.H., Philip, S.Y. (2019). A score prediction approach for optional course recommendation via cross-user-domain collaborative filtering. IEEE Access, 7: 19550-19563. https://doi.org/10.1109/ACCESS.2019.2897979

[3] Basak, J., Krishnapuram, R. (2005). Interpretable hierarchical clustering by constructing an unsupervised decision tree. IEEE Transactions on Knowledge and Data Engineering, 17(1): 121-132. https://doi.org/10.1109/TKDE.2005.11

[4] Sun, G., Bin, S. (2018). Construction of learning behavioral engagement model for MOOCs platform based on data analysis. Educational Sciences: Theory & Practice, 18(5): 2206-2216. https://doi.org/10.12738/estp.2018.5.120

[5] Sathick, K.J., Jaya, A. (2013). Extraction of actionable knowledge to predict students' academic performance using data mining technique-an experimental study. International Journal of Knowledge Based Computer System, 1(1): 28-32.

[6] Pujianto, U., Azizah, E.N., Damayanti, A.S. (2017). Naive Bayes using to predict students' academic performance at faculty of literature. In 2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), pp. 163-169. https://doi.org/10.1109/ICEEIE.2017.8328782

[7] Salanova, M., Schaufeli, W., Martínez, I., Bresó, E. (2010). How obstacles and facilitators predict academic performance: The mediating role of study burnout and engagement. Anxiety, Stress & Coping, 23(1): 53-70. https://doi.org/10.1080/10615800802609965

[8] Wu, L.Y. (2017). Research on innovation of personalized adaptive online learning model for computer major students. Boletin Tecnico/Technical Bulletin, 55(4): 582-591.

[9] Hwang, W.Y., Chen, H.R., Chen, N.S., Lin, L.K., Chen, J.W. (2018). Learning behavior analysis of a ubiquitous situated reflective learning system with application to life science and technology teaching. Journal of Educational Technology & Society, 21(2): 137-149. https://www.jstor.org/stable/26388388

[10] Schroeder, C.M., Scott, T.P., Tolson, H., Huang, T.Y., Lee, Y.H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 44(10): 1436-1460. https://doi.org/10.1002/tea.20212

[11] Sun, G., Bin, S. (2018). A new opinion leaders detecting algorithm in multi-relationship online social networks. Multimedia Tools and Applications, 77(4): 4295-4307. https://doi.org/10.1007/s11042-017-4766-y

[12] El Yassini, A., Ali Jallal, M., Ibnyaich, S., Zeroual, A., Chabaa, S. (2020). A miniaturized wide-band antenna based on the epsilon negative transmission line for wireless communication devices. Instrumentation Mesure Métrologie, 19(2): 83-90. https://doi.org/10.18280/i2m.190202

[13] Senousy, Y., Hanna, W.K., Shehab, A., Riad, A.M., El-Bakry, H.M., Elkhamisy, N. (2019). Egyptian social insurance big data mining using supervised learning algorithms. Revue d'Intelligence Artificielle, 33(5): 349-357. https://doi.org/10.18280/ria.330504

[14] Song, J., Shi, Z., Du, B., Han, L., Wang, Z., Wang, H. (2019). The data fusion method of redundant gyroscope system based on virtual gyroscope technology. IEEE Sensors Journal, 19(22): 10736-10743. https://doi.org/10.1109/JSEN.2019.2930314

[15] Jin, K. (2019). The length of the longest common subsequence of two independent mallows permutations. The Annals of Applied Probability, 29(3): 1311-1355. https://arxiv.org/abs/1611.03840