IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 14, 2020, accepted December 23, 2020, date of publication January 5, 2021, date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049157

An Unsupervised Ensemble Clustering Approach for the Analysis of Student Behavioral Patterns

XIAOYONG LI[1,2], YONG ZHANG[1], (Member, IEEE), HUIMIN CHENG[2], FEIFEI ZHOU[1], AND BAOCAI YIN[1], (Member, IEEE)

[1]Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing Artificial Intelligence Institute, Beijing 100124, China
[2]Information Technology Support Center, Beijing University of Technology, Beijing 100124, China

Corresponding author: Yong Zhang (zhangyong2010@bjut.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62072015, Grant U19B2039, Grant 61632006, and Grant U1811463.

ABSTRACT Specialized services and management must understand students' behavioral patterns in a timely and accurate manner. Based on these patterns, we can make targeted rules, especially for unexpected patterns. To perform this type of work, a questionnaire method is usually used to collect data and analyze students' behavioral states. However, the effectiveness of this method is greatly influenced by the timeliness and validity of the feedback data. To address this problem, we propose an unsupervised ensemble clustering framework to use student behavioral data to discover behavioral patterns. Because the behavioral data produced by students on campus are available in real time without intentional bias, clustering analysis can be relatively efficient and reliable. The proposed framework extracts behavior features from the two perspectives of statistics and entropy and then combines density-based spatial clustering of applications with noise (DBSCAN) and k-means algorithms to discover behavioral patterns. To evaluate the performance of the proposed framework, we carry out experiments on six types of behavioral data produced by undergraduates in a university in Beijing and analyze the relationships between different behavioral patterns and students' grade point averages (GPAs). The results show that the framework can not only detect anomalous behavioral patterns but also find mainstream patterns. The findings from this research can assist student-related departments in providing better services and management, such as psychological consulting and academic guidance.

INDEX TERMS Specialized services and management, behavioral patterns, ensemble clustering, DBSCAN, k-means.

I. INTRODUCTION

An important task in the education field is discovering student behavioral patterns and taking the corresponding actions to optimize the educational process—for example, finding various behavioral factors that have strong correlations with academic performance [1]–[6], analyzing student learning behaviors to allow teachers to adjust teaching schedules for better outcomes and to give early warnings to students who may fail exams [7]–[10], modeling the mobility flow of students on campus to support the reasonable allocation of resources by administrators, detecting students' anomalous behaviors so that managers can take timely preventive measures, and determining social networks from behavioral data to identify solitary students. Related studies have shown that these measures can significantly improve the quality of education.

To complete these studies, most researchers use a questionnaire survey method to collect data from specific students in specific circumstances. However, the method of collecting data has some limitations. First, it is impossible to capture students' current state in a timely manner with this method because surveys are conducted on a scheduled basis, such as one per academic year or semester. If students with unexpected behavioral patterns cannot be identified in a timely manner, there may be serious consequences [11]–[13]. Second, students exhibiting anomalous behaviors may deliberately supply false information to make them appear normal, while normal students may not carefully fill out the survey, so the collected data could contain noise or false information

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

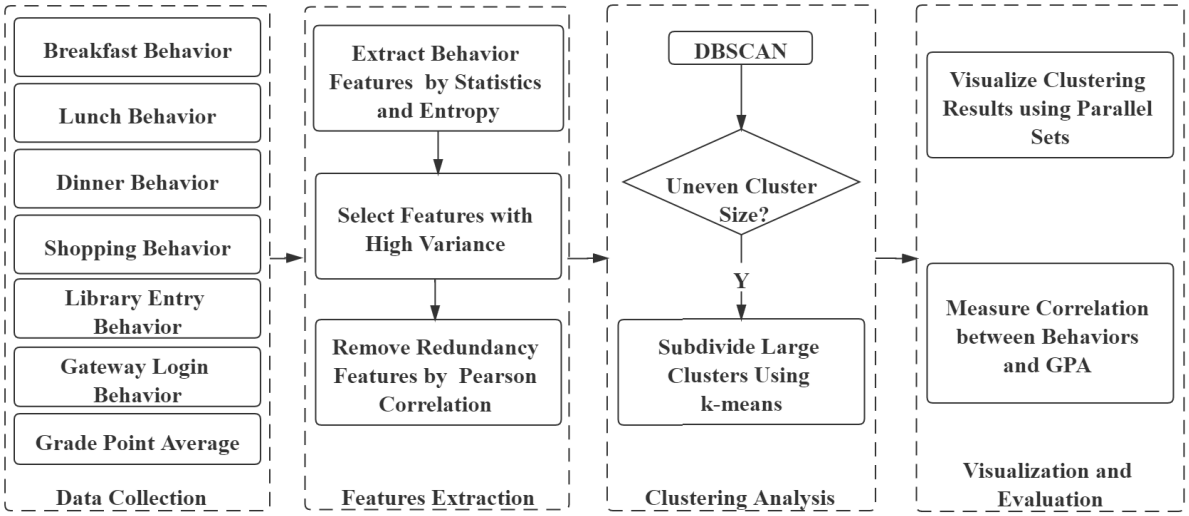


FIGURE 1. Framework of our study.

that bias the analysis results. Third, rich expert knowledge is needed to design a questionnaire that can capture enough information to comprehensively analyze students' behavioral patterns. These limitations make this method of data collection inefficient and costly. With the development of information technology, various types of accurate student behavioral data produced on campuses are stored in databases, and these data provide a more reliable and comprehensive source for real-time behavioral analysis.

The popular approaches adopted are based on machine learning algorithms, which can be categorized as supervised, semisupervised and unsupervised methods. Supervised approaches require labeled student data and the training of a classification model to determine which class an unseen student belongs to. Semisupervised approaches build a model to learn the representative features of students who belong to only one class. A student is marked as not belonging to the class when the difference between his or her features and the representative features exceeds the specified threshold. However, labeled student data, especially that of anomalous students, are not available because of privacy concerns. Additionally, student labels keep evolving, which means any model must be updated dynamically. These factors make supervised and semisupervised approaches difficult to apply in practice. In contrast, unsupervised approaches do not require labels and fully exploit the nature of datasets to cluster instances, so they are widely used in practical applications.

According to the above description, a promising direction for analyzing students' behavioral patterns is the use of unsupervised clustering algorithms to handle behavioral data produced on campus. To meet the requirements for specialized services and management, clustering algorithms should not only detect anomalous behavioral patterns for exception warnings, but also find several mainstream behavioral patterns for targeted management, and they should be easy to use. Density-based spatial clustering of applications with noise (DBSCAN) [14] and k-means algorithms are two classical unsupervised clustering methods, which are widely used in many fields. DBSCAN can automatically filter noise out of samples and find arbitrarily shaped clusters, and thus, it is applicable to cases where the distribution of the data space is unknown. However, the size of the clusters generated by DBSCAN is uneven; in extreme cases, the largest cluster contains almost all the samples, which makes it impossible to refine the understanding of the data space. And for k-means algorithm, it is a distance-based partition clustering method that works well for spherical data spaces in which the number of clusters must be specified according to the application requirements or partition metrics. However, k-means algorithm is sensitive to the noise in samples; the cluster centroids can be shifted toward noise, making them less representative. By analyzing the advantages and disadvantages of the two algorithms, and inspired by ensemble learning, we proposed an ensemble clustering methodology by combining DBSCAN and k-means algorithms, which can give full play to the advantages of the two algorithms, overcome their shortcomings, and meet the requirements of student management; its framework is shown in Fig. 1.

The proposed framework has four stages: data collection, feature extraction, clustering analysis, and visualization and evaluation. Six types of behavioral data produced on campus were collected from different information management systems using the extract-transform-load tool. These data are classical time series data composed of events with time stamps. The features of every type of behavior are extracted from the two aspects of statistics and entropy; the statistical information represents the central tendency and dispersion of the distribution of behavioral data, and entropy represents the regularity of behavior. To alleviate the curse of dimensionality, the features with small variance and redundant

features should be removed. In the clustering analysis stage, the DBSCAN algorithm is first applied to obtain initial clustering results; if a few very large clusters contain the vast majority of samples, then the k-means algorithm should be further used to subdivide them. In the final clustering results, the students that constitute noise and the students in small clusters discovered by DBSCAN can be considered anomalous, and the large clusters represent mainstream behavioral patterns. To evaluate the clustering results, parallel sets [15] are used to visualize the distribution of the features in each cluster, by which we can intuitively understand the differences among the clusters. In addition to visualizing the clustering results, we try to take students' semester grade point average (GPA) as the weak ground truth to measure the correlation between the clustering results of different behaviors and GPA.

Our main contributions are summarized as follows.

1) Six types of behavioral data in time series format were collected, and the features for each type of behavior are extracted from the perspectives of central tendency, dispersion and entropy, which provides a more reliable basis for the analysis of behavioral patterns.
2) An ensemble unsupervised clustering framework is proposed by fully taking advantage of the DBSCAN and k-means algorithms; this framework can detect unexpected behavioral patterns and discover mainstream behavioral patterns. The clustering results provide helpful information for specialized management.
3) GPA levels are taken as the ground truth to calculate extrinsic metrics to measure the correlation between different behaviors and academic performance.

The rest of this paper is organized as follows: in Section II, we provide an overview of the related work. We then describe the student behavioral data and privacy protection in Section III and extract the behavioral features in Section IV. Next, we describe the proposed clustering framework in Section V. Section VI shows the experimental results, and Section VII presents a detailed discussion. Finally, we conclude our work and propose future work.

II. RELATED WORK

In this section, we introduce related work from two perspectives.

A. ANALYSIS OF STUDENT BEHAVIORAL PATTERNS

Research on student behavior can be divided into three categories: supervised, semisupervised and unsupervised approaches. Supervised approaches aim at analyzing behavioral data to identify student labels such as academic performance and mental health. For example, [16] extracted two high-level behavioral features of orderliness and diligence from students' living behavior on campus, such as taking showers, eating meals, and fetching water in teaching buildings, and used the trained algorithm to predict the ranks of students' semester grades. Reference [17] is an updated version of [16]; this study designs a new behavior feature, sleep patterns, and applies social influence theory to build a multitask predictive framework to predict academic performance. Semisupervised approaches are usually used to detect anomalous patterns that differ greatly from normal patterns. Unsupervised approaches try to discover knowledge from massive amounts of behavioral data to support decision-making and are widely used in practice because they do not need labeled information. For example, [18] proposed a graph-based approach that aims to understand students' mobility behavioral patterns on campus. This approach constructs a behavior graph in which the nodes are dwell points extracted by the DBSCAN algorithm, and the edge values are the times from one dwell point to another point. Based on the graph, the k-core algorithm is used to recognize students' normal behavior, and the closeness center degree is used to detect abnormal behavior. Reference [19] proposed a kernel-based clustering method, called the outlier preserving clustering algorithm (OPCA), to identify both major and abnormal behaviors to completely characterize the data. To achieve the clustering objectives of compactness and separability, the OPCA combines the single linkage hierarchical clustering algorithm and the fuzzy c-means algorithm to identify well-separated clusters. The difficulty in hierarchical algorithms is setting the optimal merge or split conditions. Reference [20] discovers students' consumption capability distribution by extracting the two features of the total consumption amount and average consumption amount via the k-means algorithm and analyzes the relationship between behavioral features and academic performance.

B. APPLICATION OF THE CLUSTERING ALGORITHMS

Clustering algorithms have emerged as a powerful approach for discerning information from massive volumes of data in many fields. These algorithms aim to partition data into clusters according to specific metrics, where similar objects are in the same cluster, and they can be roughly divided into five categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods. To effectively alleviate the confusion practitioners face when using clustering algorithms and automatically recommend the most suitable algorithms for an application, [21] conducted a survey of clustering algorithms from both theoretical and empirical perspectives, in which the most representative algorithms, including partitional algorithm fuzzy c-means (FCM), hierarchical algorithm Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), density-based algorithm DENsity-based CLUstEring (DENCLUE), grid-based algorithm Optimal Grid (OptiGrid) and model-based algorithm Expectation-Maximization (EM), were evaluated on 10 datasets via internal and external validity metrics, stability, runtime, and scalability. Due to the inherent property of not requiring labels, clustering methods are widely used in various applications; for example, [22] built an unsupervised system to capture user behavior, which partitions a similarity graph to identify clusters and leverages iterative feature pruning

to discover the natural hierarchy within user clusters. Reference [23] developed a trajectory clustering method in which an edit distance algorithm is designed to measure the similarity of the trajectories, and an adaptive hierarchical clustering algorithm is applied to distinguish regular and anomalous trajectories.

To further improve clustering quality, ensemble clustering methods [25] are applied in unsupervised learning scenarios, which are like ensemble methods in a supervised learning setting. According to the combination approach of clustering results of base learners, the ensemble clustering can be divided into two categories. In one category, diverse clustering results are parallelly obtained by running base methods, in which the base methods may be the same algorithm with different parameter configurations, different data samples, or different types of algorithms. These base clusters are then consolidated into the final cluster via a combination function. In the other category, the base methods are applied sequentially, and the results of the previous clustering can serve as the knowledge for the following clustering. It has been proven that a clustering ensemble can effectively reduce the adverse effects of factors in the clustering process, such as when clustering assumptions do not match the real data structure, and randomness of initial parameter configuration. For example, [18] and [19] integrated different clustering algorithms to obtain better results. The BIRCH algorithm integrates hierarchical clustering and other clustering methods, such as iterative partitioning, to cluster a large amount of numeric data, which greatly improves the quality of hierarchical agglomeration and can be used for clustering streaming and dynamic data. Note that clustering aims at finding the inherent structure of data, and there is no unified objective function to evaluate its performance, so we can design a new clustering method to accommodate to requirements of applications.

III. DATA DESCRIPTION AND PRIVACY PROTECTION

A. DATASET

The student behavioral data used in this paper include consumption behavioral data, library entry data, and gateway login behavioral data. These data were collected from different databases using extract-transform-load tools. Every behavioral data category is composed of a series of records indexed sequentially. The consumption behavior records have four attributes: time, location, transaction amount, and transaction type. Although there are many types of consumption behaviors, we collect only dining behavior and shopping behavior because they are the main consumption behaviors and offer a large volume of information. To further understand dining behavior, it is subdivided into breakfast behavior, lunch behavior and dinner behavior according to the consumption time; the time intervals are 6:00 am to 9:00 am, 11:00 am to 2:00 pm, and 4:30 pm to 8:30 pm, respectively. Entering a library is an important learning activity, and its record contains the two attributes of time and location. Because the dataset we used includes only one library, we remove the location attribute. The gateway system is a protocol converter deployed between the Internet and the campus local network. Students must log into it when they want to access the Internet from the campus network, so the gateway system can record the login time, logout time, login location, duration of Internet access, and flow of network traffic. In addition to the behavioral data, we collect students' GPAs to represent their academic performance.

B. PRIVACY PROTECTION

Privacy protection is a matter of great concern to us in the analysis process. First, in the enrollment stage, students are asked if they would like to share their behavioral data produced on campus to improve education quality. Second, the student ID is encoded as a unique anonymous identifier. Third, behavior time is transformed into an integer index. We uniformly divide one day into 48 bins and assign a bin an integer index starting from 1; the behavior time can be replaced with the index of the corresponding bin. For example, 8:10 am can be transformed into 17. Finally, the behavior location is converted to an anonymous symbol so that we cannot determine the specific location. After these processes, the behavior records with the same time and location are merged into one record. For consumption behavior, the transaction amounts in the merged records are added. For the gateway login behavior, the time duration and network traffic are added. For the library entry records, we remove duplicate records. As a result, the dataset does not contain any personal information, yet enough information is retained to support the behavior clustering analysis.

IV. BEHAVIOR FEATURES

A major challenge in the clustering analysis of behavioral patterns is the extraction of features from a large amount of behavioral data. In this paper, we use statistics and entropy to extract features and then select features using variance and correlation analyses.

A. FEATURE EXTRACTION

The attributes of behavioral data can be divided into two types: nominal and numeric. Except for behavioral location, which is nominal, all other attributes are numeric. To express the distribution of values of numeric attributes, we measure its central tendency using the range, mode and mean and measure its dispersion using a five-number summary consisting of the minimum, Q2, median, Q3, and maximum. For nominal attribute behavioral location, we calculate the Shannon entropy to measure the regularity of behavior from the spatial dimension. Because behavioral time has been transformed to an integer index, we also compute the entropy from the temporal dimension. The Shannon entropy is defined as (1):

$$H = -\sum_{i} p(i) log p(i) \tag{1}$$

where $p(i)$ is the probability of behavior event i occurring at a given time or location. The smaller the entropy is, the more

TABLE 1. Features of breakfast behavior.

No.	Feature Names	Description of Features
1	bf_frequency	Frequency of breakfast
2	bf_Loc_entropy	Shannon entropy of locations of breakfast
3	bf_Time_entropy	Shannon entropy of time of breakfast
4	bf_Time_mean	Average time of breakfast
5	bf_Time_mode	Most frequent time for breakfast
6	bf_Time_range	Difference between the earliest time and the latest time of breakfast
7	bf_Time_min	Earliest time of breakfast
8	bf_Time_Q1	25% of the breakfast time values are earlier than this time
9	bf_Time_median	50% of the breakfast time values are earlier than this time
10	bf_Time_Q3	75% of the breakfast time values are earlier than this time
11	bf_Time_max	Latest time of breakfast
12	bf_Trans_mean	Average transaction amount of breakfast
13	bf_Trans_mode	Mode of transaction amount of breakfast
14	bf_Trans_range	Difference between the minimum and maximum transaction amounts of breakfast
15	bf_Trans_min	Minimum transaction amount of breakfast
16	bf_Trans_Q1	25% of the transaction amount values of breakfast are less than this value
17	bf_Trans_median	50% of the transaction amount values of breakfast are less than this value
18	bf_Trans_Q3	75% of the transaction amount values of breakfast are less than this value
19	bf_Trans_max	Maximum transaction amount of breakfast
20	bf_Trans_sd	Standard deviation of the transaction amount of breakfast

regular the behavior. For instance, the entropy in the time dimension could be approximately zero if a student always has breakfast at the same time bin. However, it is difficult to compute the Shannon entropy for attributes with continuous values, such as transaction amounts, durations of Internet access and network traffic flows, so we use the standard variance instead of the entropy to measure the stability of these attributes. Similar to entropy, a lower variance indicates a more stable state. In addition, we determine the frequency to indicate how frequently every behavior occurs.

Table 1 lists the 20 extracted features of breakfast behavior. The other three types of consumption behavior, lunch behavior, dinner behavior, and shopping behavior have the same features as breakfast behavior. Their features are prefixed with '*lu*', '*di*' and '*sp*', respectively. Because there are fewer locations for shopping in our dataset, we ignore the entropy of location for shopping behavior. The library entry behavior has nine time-related features with the same meaning as those of breakfast behavior in addition to frequency, which are prefixed with '*lib*', such as *lib_frequency* and *lib_time_entropy*. For the gateway login behavior, we extract features for login time, logout time, login location, duration of Internet access, and network traffic flow attributes, which are prefixed with '*gw_intime*', '*gw_outtime*', '*gw_loc*', '*gw_dura_acces*', and '*gw_traf_flow*', respectively. There are a total of 38 features, including frequency and entropy of location.

B. FEATURE SELECTION

There are dozens of features for every behavior, which can result in the curse of dimensionality in clustering algorithms because the distance taken by the algorithm to measure the similarity between samples may be ineffective for high-dimensional data. To overcome this difficulty, we select optimal features by analyzing their variance and correlation.

1) VARIANCE ANALYSIS

Variance is a measure of data dispersion that indicates how spread out a data distribution is. A feature with low variance has values that tend to be very close to the mean, which can provide minimal useful clustering information. Therefore, we remove the features with low variance. The variance of the features of consumption behavior are shown in Fig. 2. In observing these features, we found three phenomena: (1) The variance of all features is less than 0.1, which indicates that there is little difference in the students' consumption behavior. (2) The features related to the transaction amount have lower variance than other features. Among the four types of consumption behaviors, the variance of the amount-related features of breakfast behavior is close to zero, which is easy to understand because the prices of breakfast foods are very close. For lunch behavior, the variance of the amount-related features is higher than that of the breakfast behavior because there are a variety of foods available for lunch, and their prices are more varied. Usually, the menu for dinner is the same as that for lunch; however, the variance of the amount-related features of dinner behavior is far lower than that of lunch behavior, which is very interesting. (3) The frequency, location entropy, and time-related features have relatively high variance, so these features can be used to express distinct behavioral patterns. Fig. 3 shows the variance of the features of library entry behavior and gateway login behavior. Nine of the 11 features of library entry behavior have variances greater than 0.03, which indicates that its behavioral patterns are very different. The features that have high variance in gateway login behavior are *gw_intime_mode, gw_intime_range, gw_intime_min* and *gw_outtime_mode*.

We set the variance threshold to 0.02 for the selected features, and the numbers of selected features are 10, 6, 7, 6, 9 and 4 for breakfast behavior, lunch behavior, dinner behavior, shopping behavior, library entry behavior and gateway login behavior, respectively.

2) CORRELATION ANALYSIS

A feature may be redundant if it can be derived from other features. In this section, we use the Pearson correlation coefficient to measure how strongly one feature implies another and then remove the redundant features. The coefficient matrix of features selected in section IV-B1 are shown in Figs. 4 and 5, through which we can clearly understand the

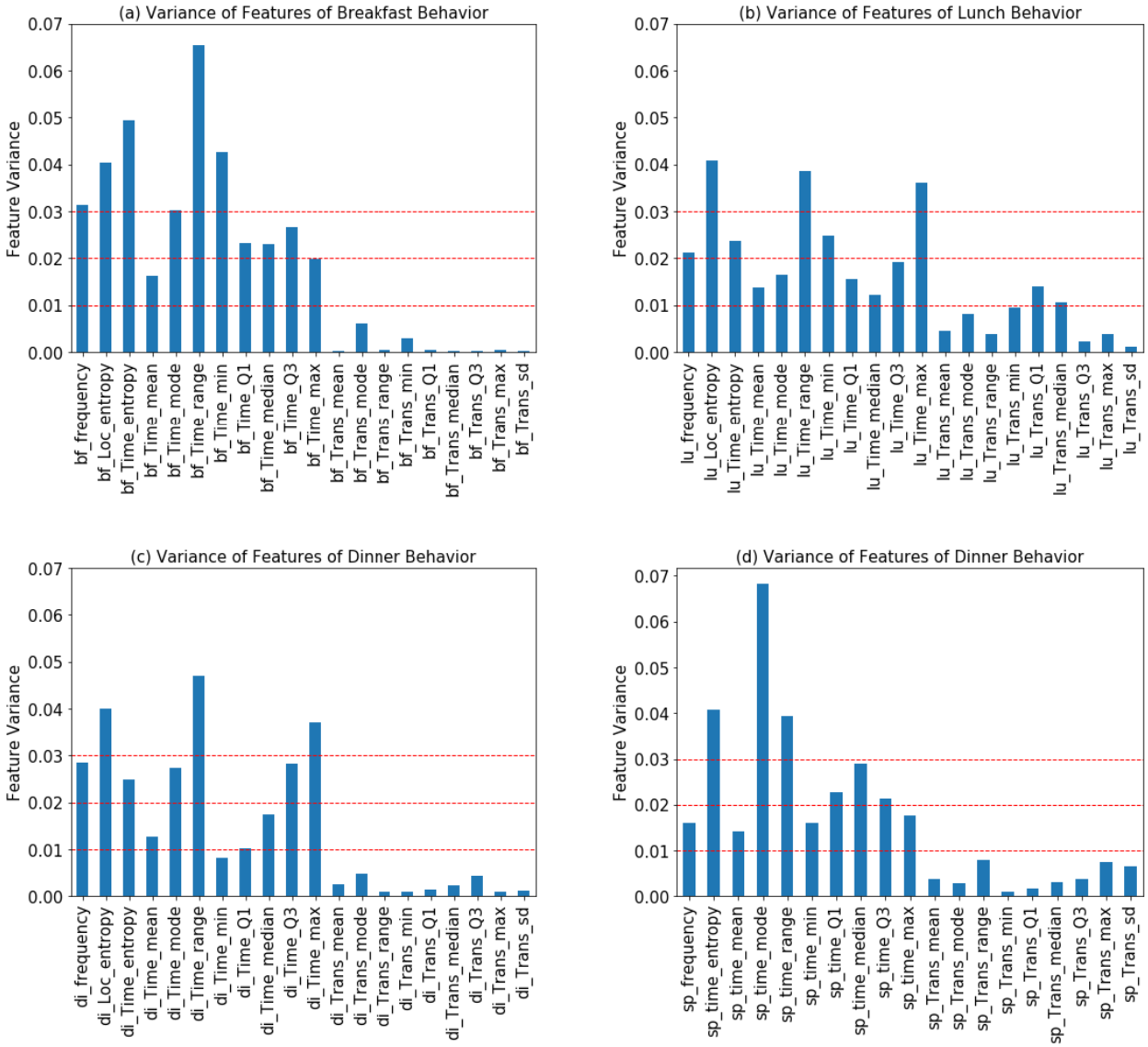


FIGURE 2. **Variances of the features of (a) breakfast behavior, (b) lunch behavior, (c) dinner behavior, and (d) shopping behavior.**

correlation between any two features. For example, features *bf_time_median* and *bf_time_mode* are highly correlated with a coefficient of 0.9, as shown in Fig. 4(a). Considering that feature *bf_time_median* has a lower variance than feature *bf_time_mode*, 0.023 vs 0.03, we remove *bf_time_median* to reduce redundancy.

In this paper, we set a correlation threshold to 0.8 to remove redundant features with lower variance. After the variance and correlation analysis, the reserved features are *bf_frequency, bf_Loc_entropy, bf_Time_entropy, bf_Time_mode, bf_Time_range, bf_Time_min, bf_Time_Q1, bf_Time_Q3* and *bf_Time_max* for breakfast behavior; *lu_frequency, lu_loc_entropy, lu_time_entropy, lu_time_range* and *lu_time_min* for lunch behavior; *di_frequency, di_loc_entropy, di_time_entropy, di_time_mode, di_time_range* and *di_time_Q3* for dinner behavior; *sp_time_entropy, sp_time_mode, sp_time_Q1, sp_time_median* and *sp_time_Q3* for shopping behavior; *lib_time_mode, lib_time_range, lib_time_min* and *lib_time_max* for library entry behavior; and *gw_intime_mode, gw_intime_range, gw_intime_min* and *gw_outtime_mode* for gateway login behavior.

V. PROPOSED CLUSTERING METHODOLOGY

Clustering algorithms are very beneficial for discovering students' behavioral patterns since they do not need labeled information from students. The existing popular clustering algorithms can be divided into three types: partitioning, hierarchical and density-based algorithms. Partitioning algorithms partition a data space into k clusters, but these algorithms are sensitive to noise, and the shape of all the clusters is

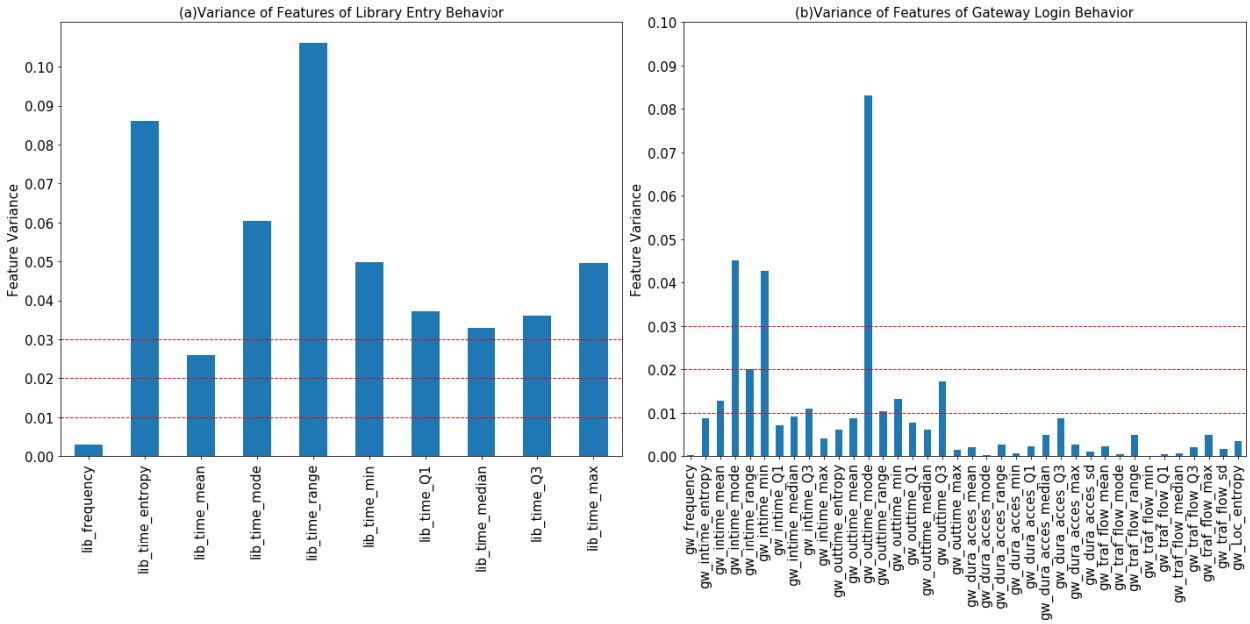


FIGURE 3. Variances of the features of (a) library entry and (b) gateway login.

convex. Hierarchical algorithms do not need k as an input and can discover nonconvex clusters; however, it is very difficult to define a termination condition for when to terminate the merge or division operation for these algorithms. Density-based algorithms can discover clusters of arbitrary shape with two given parameters and automatically filter out noise, but clusters of uneven size, especially very large clusters, may not meet the requirements of student services and management. As an effective way to improve the quality of clustering, ensemble clustering has received extensive attention. In this paper, we propose an ensemble clustering framework to determine student behavioral patterns. The basic idea is that framework first uses the density-based algorithm DBSCAN to filter out noise and form the initial clustering and then uses the k-means partitioning algorithm to subdivide the large clusters constructed by DBSCAN to obtain the final clustering result.

A. INITIAL CLUSTERING USING DBSCAN

The key idea of the DBSCAN algorithm is that the neighborhood of a given radius (*Eps*) for each sample of a cluster must contain at least a minimum number (*MinPts*) of samples. The neighborhood of a sample p, denoted by $N_{Eps}(p)$, is defined by $N_{Eps}(p) = \{q \in D | dist(p, q) \leq Eps\}$. Given the parameters *Eps* and *MinPts*, DBSCAN randomly chooses a core sample as a seed and retrieves all samples that are density-reachable from the seed to form a cluster; the samples that do not belong to a cluster are defined as noise. To determine the two parameters, [14] developed a simple but effective heuristic method. For a given *MinPts*, this method defines a function mapping each sample to the distance from its *MinPts*-th nearest neighbor and then sorts all the samples in descending order of *MinPts*-dist value and plots them. The graph can provide some hints about the distribution of the density. We usually plot the graphs with respect to different *MinPts*, and the optimal *MinPts* is set to the minimum value whose *MinPts*-dist graph does not significantly differ from others. The optimal *Eps* can be the *MinPts*-dist value of the sample at the first "valley" in the graph of optimal *MinPts*.

Considering that the distribution of the student behavioral feature space is unknown and that there is some noise, DBSCAN is selected to construct the initial clustering. The noise samples and samples belonging to clusters of small size can be considered anomalous, while the large clusters represent the mainstream behavioral patterns. However, DBSCAN may produce a very large cluster that contains almost all the samples, which does not meet the requirements of specialized services and management.

B. SUBDIVISION CLUSTERING USING K-MEANS

The K-means algorithm can partition the data space into the expected number of clusters, so it is the best complement to DBSCAN. The very large clusters generated by DBSCAN can be further subdivided using k-means. The elbow method is a popular method for determining the optimal number of clusters; this method calculates the sum of the within-cluster variance, also called inertia, when given a number of clusters k and then plots the curve of variance with respect to k. The k value at the first turning point of the curve can be the optimal number of clusters. In addition to the variance metric, the three intrinsic metrics of the silhouette coefficient (SC), the Calinski-Harabasz index (CHI) and the Davies-Bouldin index (DBI), can also be calculated for plotting the curves with respect to different k values. Better clustering results should be obtained with a higher silhouette score, higher CHI

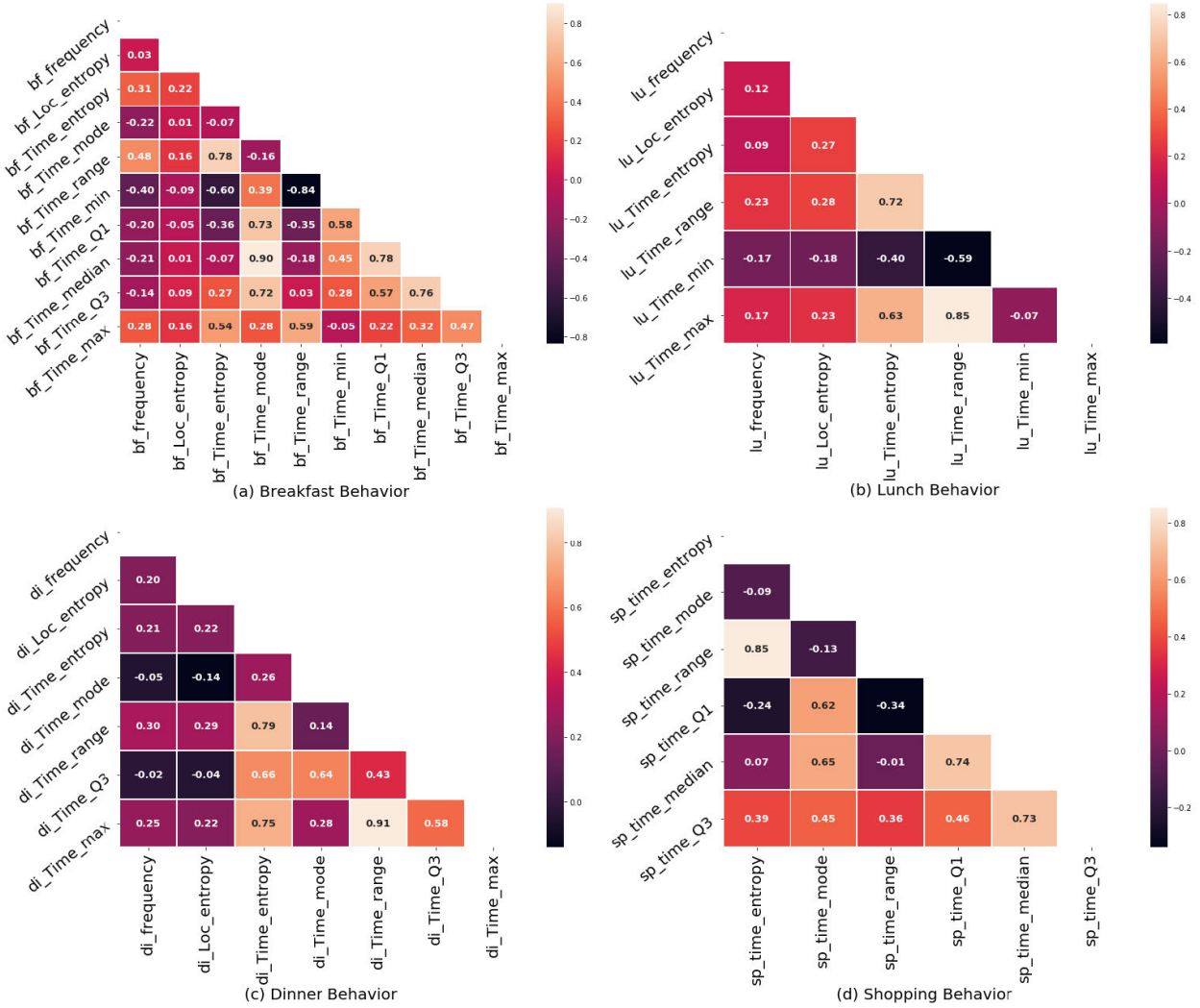


FIGURE 4. Correlation coefficients between features of (a) breakfast behavior, (b) lunch behavior, (c) dinner behavior, and (d) shopping behavior.

and lower DBI. In the subdivision process, we can simultaneously consider the four metrics and application requirements to select the number of subclusters. Note that if the clustering result of DBSCAN has met the application requirements, it is not necessary to further carry out subdivision. After subdivision, the very large clusters of DBSCAN can be replaced with the subdivided subclusters to obtain the final results.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

The six types of behavioral data analyzed in this paper were collected from 9024 undergraduates at a university in Beijing during the spring of 2019. The experiments are implemented using Python and scikit-learn libraries.

A. CLUSTERING RESULTS USING DBSCAN

To determine the parameters *Eps* and *MinPts* of DBSCAN, we plot a *MinPts*-dist graph for each type of behavior, where *MinPts* is set from 2 to 24. In the six graphs, the curves do not significantly change when *MinPts* is greater than 8, so we set *MinPts* to 8. The 8-dist graphs show that the optimal *Eps* values are 0.231 for breakfast behavior, 0.14 for lunch behavior, 0.175 for dinner behavior, 0.124 for shopping behavior, 0.082 for library entry behavior, and 0.09 for gateway login behavior. The clustering results of DBSCAN with the given values of *Eps* and *MinPts* are shown in Figs. 6 and 7, where -1 is the label of the noise cluster, the normal clusters are labeled with numbers starting from 0, and the number of students in each cluster is above its bar. For example, there are a total of 19 clusters numbered from -1 to 17 for breakfast behavior, as shown in Fig. 6(a); noise cluster -1 contains 184 students who can be identified as those with unexpected behavioral patterns; clusters numbered 2, 4, 12, 13, 14, 15, 16 and 17 all contain relatively few students, less than 200, so the behavioral patterns they represent should be in the minority; clusters 0, 1, 3, 5, 6, 7, 8, 9, 10 and 11 all contain relatively large numbers of students, and they can represent

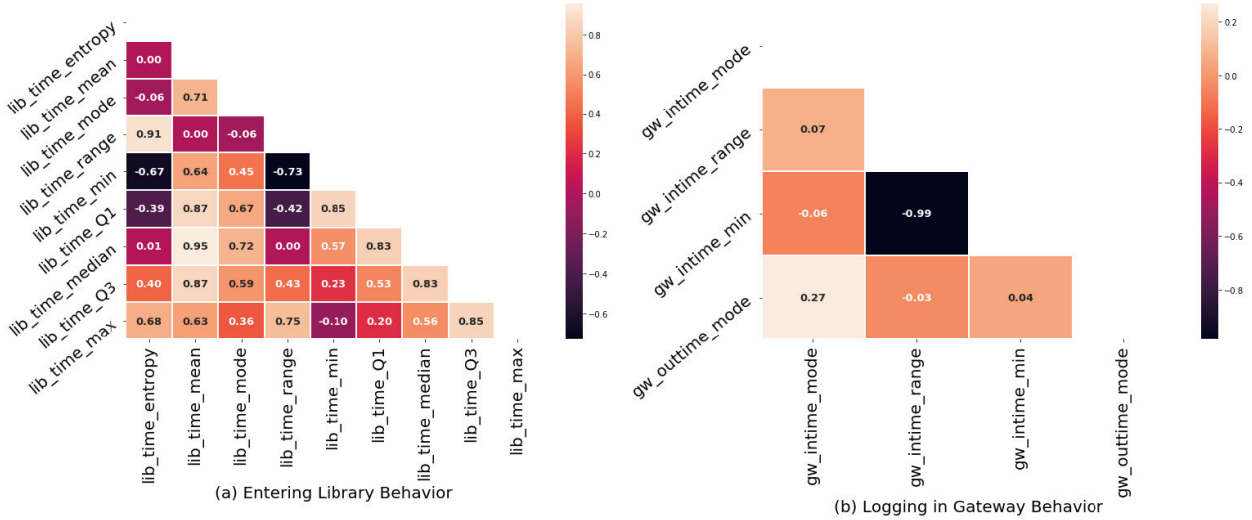


FIGURE 5. Correlation coefficients between features of (a) library entry and (b) gateway login.

students' mainstream behavioral patterns, especially clusters 0, 1, and 3. Based on the results, student services and management departments should pay more attention to the noise clusters and minority clusters for early warnings and provide targeted services and management according to mainstream patterns. Lunch behavior has similar clustering results as breakfast behavior, as shown in Fig. 6(b). However, the clustering results of the other four types of behavior are not ideal; as shown in Fig. 6(c) and (d) and Fig. 7(a) and (b), their clusters 0 contain more than 90% of students. Although this phenomenon indicates that the behavioral patterns of the majority of students are relatively similar, it is necessary to further subdivide these clusters to understand behavioral patterns in detail. However, there is no one threshold that can be applied to all applications to determine which clusters need to be subdivided. It should be specified according to specific application requirements, here we set it to 80%.

B. SUBDIVISION RESULTS USING K-MEANS

We use k-means for subdivision because the number of clusters k can be specified in advance by observing the curves of the four metrics, as well as the management requirements and prior knowledge. Additionally, this method can obtain more representative behavioral patterns than direct application of k-means to the original dataset because DBSCAN has filtered out the noise and very small clusters.

Here, we take dinner behavior as an example to illustrate how to determine the number of subclusters. The line charts of the four metrics are plotted as shown in Fig. 8, where k is set from 2 to 50. The inertia metric decreases as k increases, as shown in Fig. 8(a), and its scope becomes smooth when k is greater than 10, which indicates that it cannot significantly reduce the inertia value when the dataset is divided into more than 10 clusters, so the proposed number of clusters ranges from 2 to 10. Fig. 8(b) shows the curve of the silhouette score. The proposed k values range from 2 to 6 since their silhouette scores are higher than others. The curve of CHI is shown in Fig. 8(c); its shape is similar to the inertia metric, and we can take the values from 2 to 10 as the candidates for k. The curve of DBI fluctuates considerably with respect to k and reaches the two lowest values when k equals 6 or 10. Simultaneous consideration of the four metrics shows that the optimal number of subclusters is six; the corresponding metric values are highlighted using a red vertical line in these graphs. In the same way, the optimal numbers of subclusters for cluster 0 of shopping behavior, library entry behavior, and gateway login behavior are determined to be six, five and four, respectively. In practice, we can also introduce management requirements to determine the optimal number.

The final clustering results of these four types of behaviors after subdividing cluster 0 with the given k are shown in Figs. 9 and 10, in which the clusters suffixed with '_DBSCAN' are noise clusters and minority clusters generated by DBSCAN, while clusters suffixed with '_KMEANS' are the subclusters subdivided using k-means. The number of students in each cluster is above the bar. The final result not only retains the noise and small clusters but also subdivides the large clusters into basically uniform subclusters.

C. VISUALIZATION OF THE CLUSTERING RESULTS

To intuitively understand the clustering results, parallel sets are introduced to visualize them. Parallel sets are a method for the visualization of categorical data, in which an axis represents a behavioral feature, the boxes in the axis represent the feature value categories, and the thickness of each curved line represents a quantity that is repeatedly subdivided by category. By observing the result, we can understand the distribution of the behavioral features of every cluster and the difference between clusters. We take dinner behavior as an example to illustrate the visualization effect,

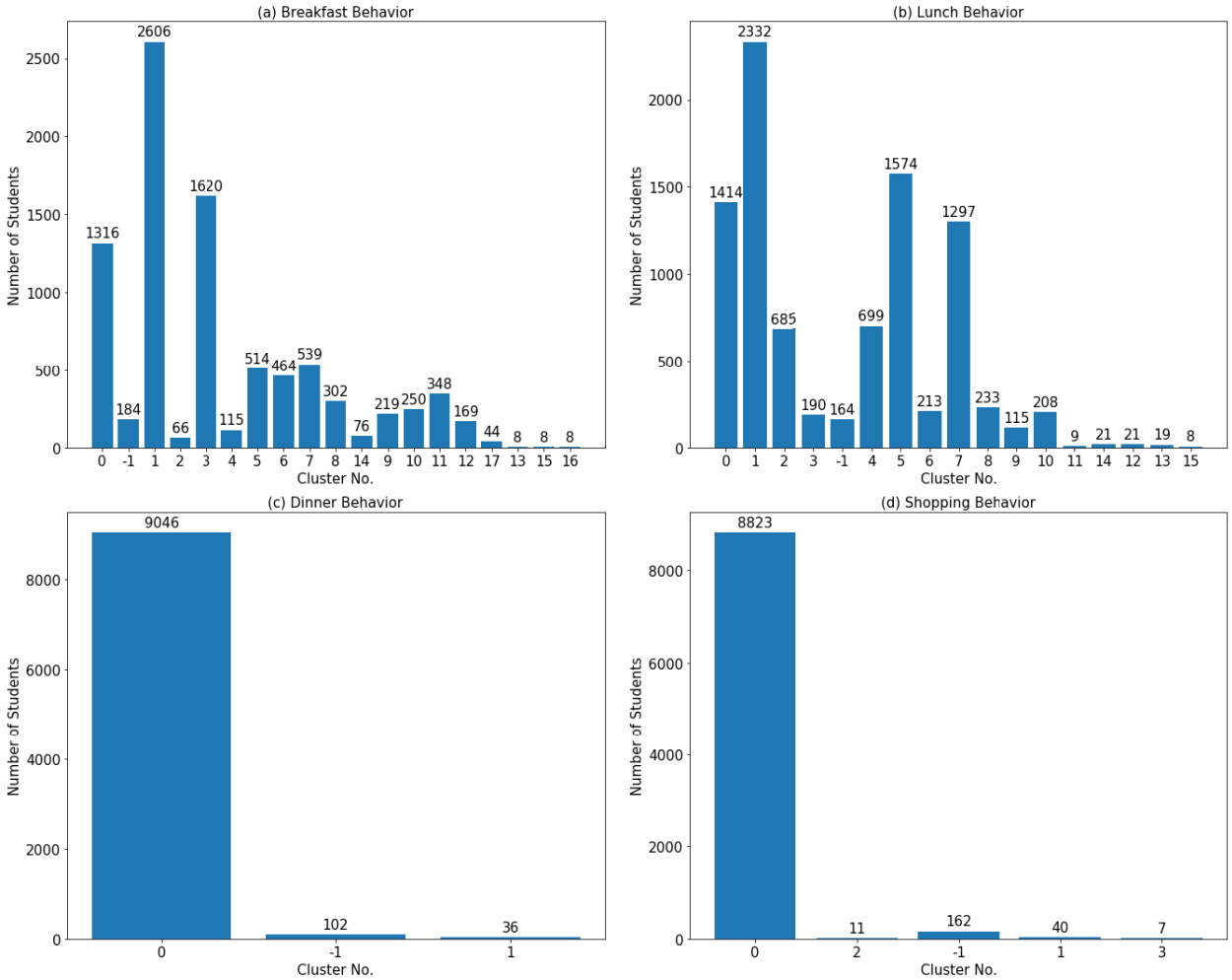


FIGURE 6. Initial clustering results of (a) breakfast behavior, (b) lunch behavior, (c) dinner behavior, and (d) shopping behavior using DBSCAN.

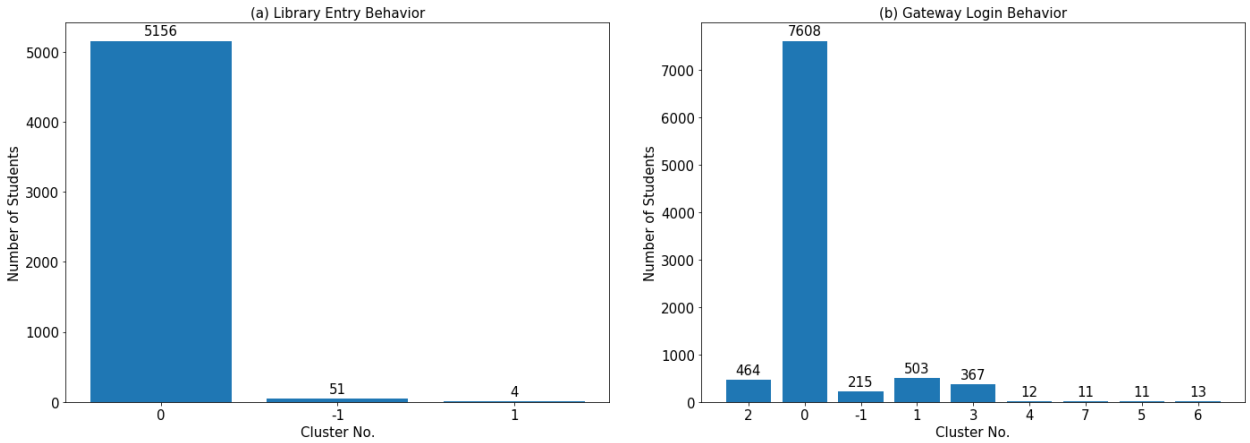


FIGURE 7. Initial clustering results of (a) library entry and (b) gateway login using DBSCAN.

as shown in Fig. 11. The continuous features *di_frequency, di_loc_entropy, and di_time_entropy* are converted to discrete ranges taken as categories. By observing the graph, we can find that these clusters have distinct characteristics. For example, there are 1870 students in cluster *0_KMEANS*, the frequency of having dinner varies from 20 to 80, the time

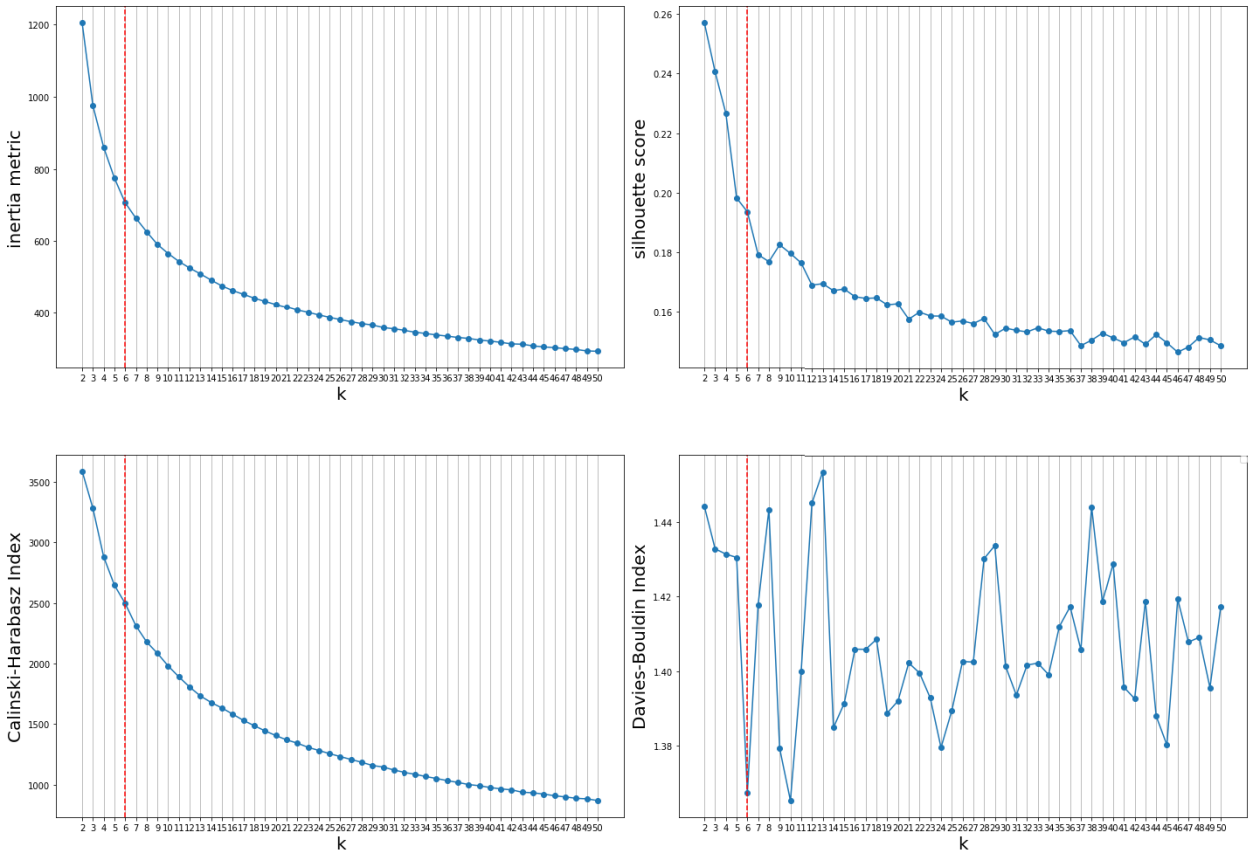


FIGURE 8. Line charts of the four metrics for determining the number of subclusters of dinner behavior.

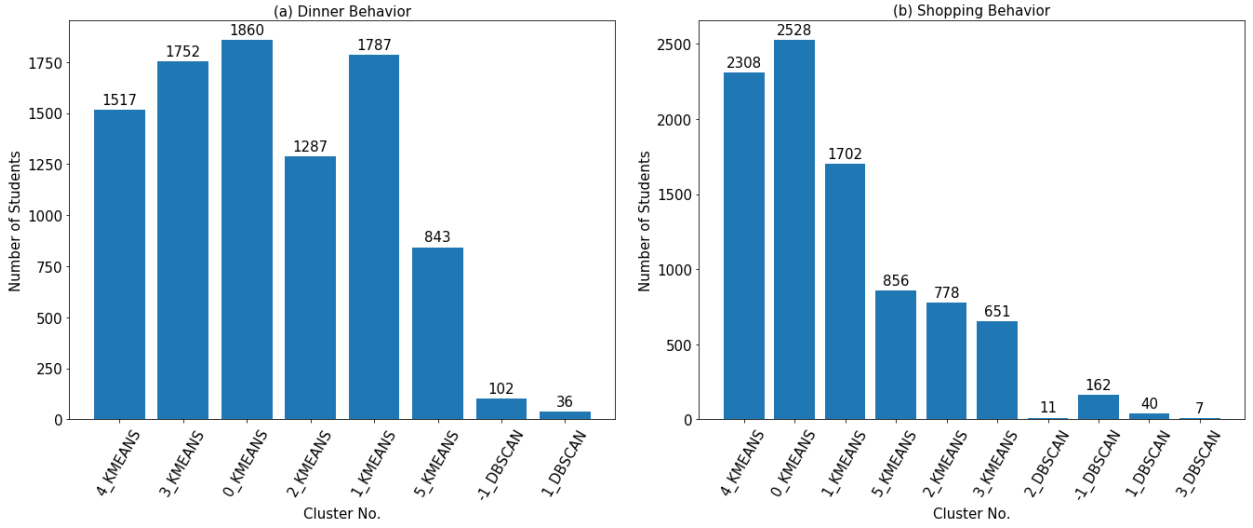


FIGURE 9. Final clustering results of (a) dinner behavior and (b) shopping behavior.

mode is between 5:30 pm and 6:00 pm, the Q3 of time is between 6:30 pm to 7:30 pm, and the time entropy and location entropy are very high. This result indicates that these students can have dinner on time, but the location and time are more diverse. In this way, the characteristics of the clusters of dinner behavior are summarized in Table 2. Students in cluster *1_KMEANS* always have dinner at the same canteen; students in cluster *2_KMEANS* stably have dinner at the regular time; students in cluster *3_KMEANS* usually have dinner very late, and the time is unstable; students in cluster

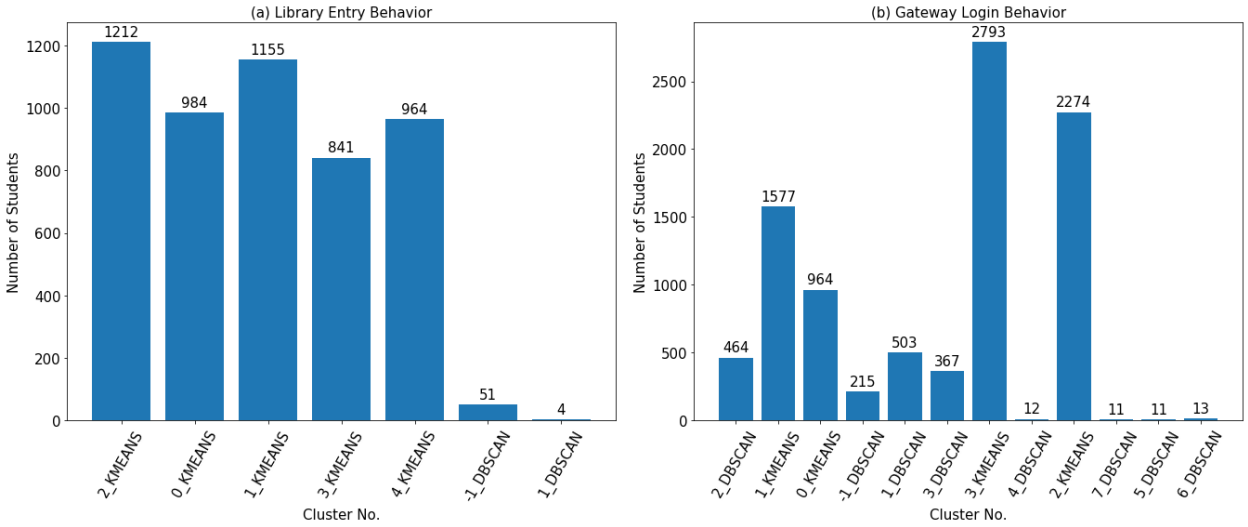


FIGURE 10. Final clustering results of (a) library entry behavior and (b) gateway login behavior.

TABLE 2. Characteristics of the clusters of dinner behavior.

Cluster	Frequency	Location Entropy	Time Entropy	Time Mode	Time Range	Time Q3
0_KMEANS	disperse	high	high	moderate	high	high
1_KMEANS	disperse	low	moderate	moderate	moderate	moderate
2_KMEANS	disperse	moderate	low	low	median	low
3_KMEANS	disperse	disperse	high	high	high	high
4_KMEANS	high	high	high	low	high	low
5_KMEANS	disperse	low	low	low	low	low

4_KMEANS often have dinner at the regular time on campus, but the time and location are unstable; most of the students in cluster *5_KMEANS* have dinner less often, but they have dinner at the regular time, and the time and location of dinner are very stable.

VII. DISCUSSION

A. ADVANTAGES OF THE PROPOSED METHOD

As described above, the proposed method can meet the requirements of student services and management, and it is easy to implement. By observing the results in section VI, it is clear that why k-means is applied after DBSCAN. And to explain the role of DBSCAN in the proposed method, we adopt the principal component analysis (PCA) method to reduce the dimensionality of the behavior feature space to two and then use a scatter chart to visualize the clustering results. Fig. 12(a) visualizes the clustering result of dinner behavior generated by the proposed method, in which the light blue points represent samples in small clusters and noise clusters, the other six bright colors represent the six sub-clusters of cluster 0, and the red solid circles indicate the centroids of the six sub-clusters. Note that the cumulative variance explained by the two selected components in Fig. 12(a) is 68.6%, so they can basically express the distribution of students in the original space. Fig. 12(b) shows the clustering result of all samples of dinner behavior only using k-means, in which six colors represent six clusters, and the blue crosses are their centroid. In order to compare the centroid of the two clustering results, we also plot the centroid of the six sub-clusters generated by the proposed method in Fig. 12(b). It can be seen that there is a deviation between the two types of centroids, which is obvious in the purple area of Fig. 12(b). Apparently, the noise samples make the centroid less representative. This illustrates the importance of filtering noise using DBSCAN, which can make the centroids of sub-clusters more representative.

B. HOW TO CLUSTER MULTISOURCE BEHAVIORS

In the previous sections, the proposed clustering method is applied only to single-source behavioral data. We want to determine whether this method can be used to cluster students' multisource behaviors as a whole. To test this, the reserved features of six types of behaviors are concatenated to form 33 features of multisource behavioral data. In section IV-A, we use the Pearson coefficient to measure the correlation between the features of single-source behavior. Here, we take the average value of the coefficient to measure the correlation between different behaviors. The coefficient matrix is shown in Table 3. We find that the correlation between different behaviors is very weak; among them, the highest value of 0.164 is between lunch behavior and dinner behavior. Therefore, we do not remove any more features. However, because the Euclidean distance measure taken by DBSCAN and k-means can be ineffective on high-dimensional data, the proposed method may not work well on multisource behavioral data.

There are many clustering algorithms that handle high-dimensional data. In this paper, we introduce the spectral clustering algorithm [24], which projects the original data into a low-dimension embedding of the affinity matrix between samples and then runs a clustering algorithm such

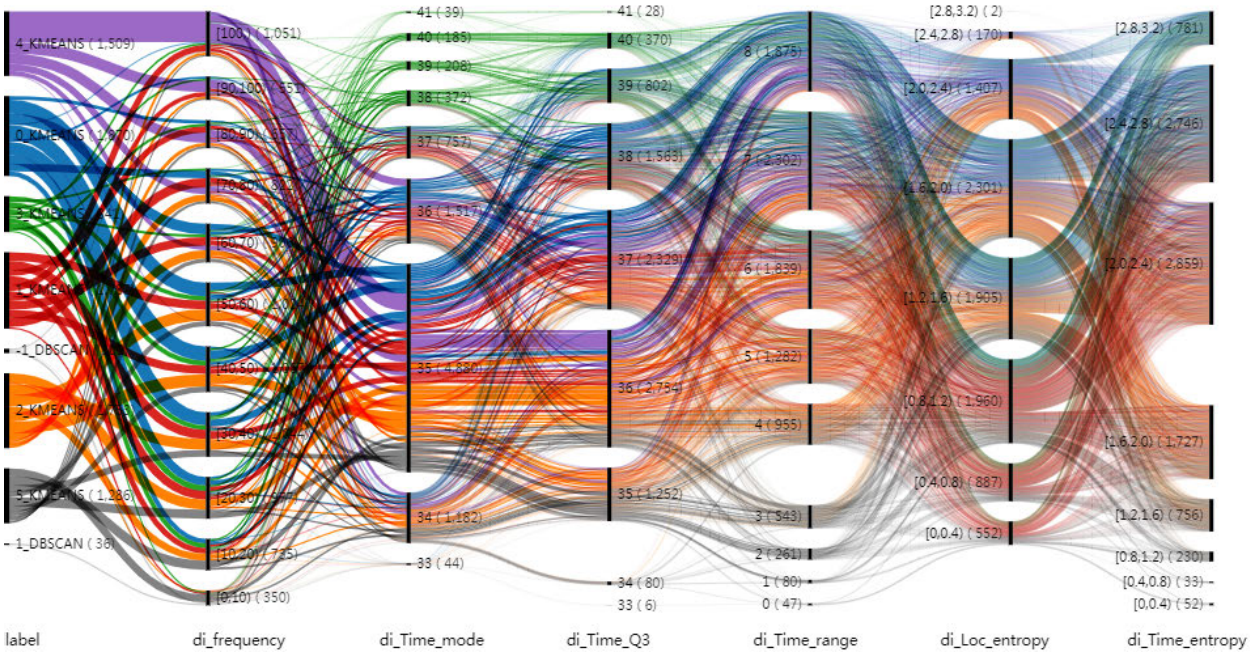


FIGURE 11. Visualization of the clustering result of dinner behavior using parallel sets.

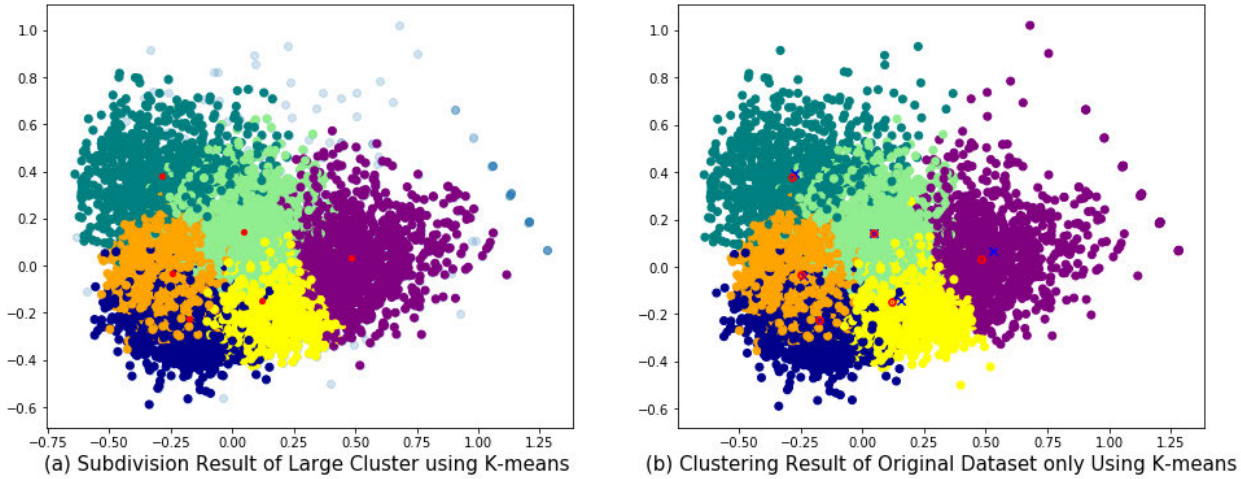


FIGURE 12. Subdivision results vs the clustering results of dinner behavior using only *k*-means.

as k-means in the low-dimensional space. The dimensionality of the new space should be the same as the desired number of clusters. Because the k-means algorithm is used after dimensionality reduction, we can use the method described in section V-B to determine the optimal number of clusters. According to the curves of the silhouette score, CHI, and DBI with respect to different k values, the k value is set to 3. The clustering result is shown in Fig. 13, in which cluster 1 can be considered anomalous and clusters 0 and 2 represent two mainstream behavioral patterns. Because only 34.3% of the cumulative variance is attributed to the top two components of the PCA, the clustering result cannot be visualized via a scatter chart.

C. CORRELATION ANALYSIS BETWEEN BEHAVIORAL PATTERNS AND ACADEMIC PERFORMANCE

Relevant studies have shown that behavioral patterns have an important impact on academic performance. To analyze the correlation between different behavioral patterns and academic performance, we use six metrics, the adjusted Rand index (ARI), normalized mutual information (NMI), homogeneity, completeness and Fowlkes-Mallows Index (FMI), to measure the similarity between the clustering results of different behaviors and academic performance levels. The upper bound of all these metrics is 1, and a higher value indicates better similarity. Students' academic performance is represented by their GPA and divided into four levels. The

TABLE 3. Pearson correlation coefficient between behaviors.

	Breakfast behavior	Lunch behavior	Dinner behavior	Shopping behavior	Library entry behavior	Gateway login behavior
Breakfast behavior	0.365					
Lunch behavior	0.063	0.230				
Dinner behavior	0.063	0.164	0.391			
Shopping behavior	0.010	0.022	0.060	0.556		
Library entry behavior	-0.005	0.081	0.057	-0.014	0.668	
Gateway login behavior	0.014	0.010	0.027	0.018	-0.006	0.163

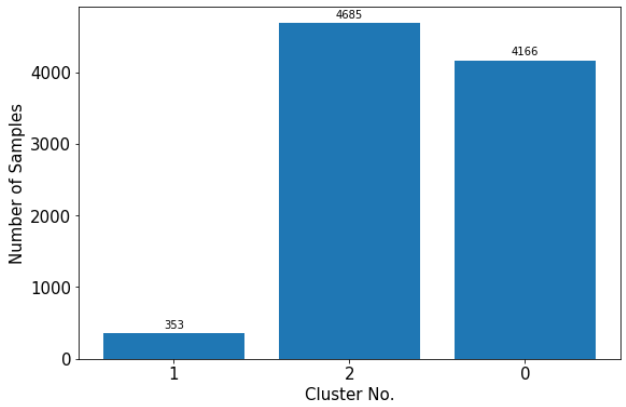


FIGURE 13. Clustering result of multisource behavior by spectral clustering.

TABLE 4. The correlation between different behavioral patterns and academic performance.

	ARI	AMI	Homogeneity	Completeness	V-measure	FMI
Breakfast Behavior	0.019	0.004	0.017	0.005	0.008	0.33
Lunch Behavior	0.001	0.001	0.006	0.002	0.003	0.307
Dinner Behavior	0.013	0.007	0.019	0.008	0.011	0.37
Shopping Behavior	- 0.009	0.003	0.009	0.004	0.005	0.344
Library Entry Behavior	0.005	0.003	0.011	0.004	0.006	0.375
Gateway Login Behavior	0.004	0.004	0.014	0.005	0.008	0.357
Multisource Behavior	0.008	0.015	0.019	0.015	0.017	0.541

metric values are shown in Table 4. We find that the clustering results of multisource behavior are more consistent with GPA than the single behaviors and that the clustering results of dinner behavior have a stronger correlation with GPA than other single-source behaviors. This finding suggests that student management departments can focus on dinner behavior to improve students' academic performance.

D. COMPARISON WITH OTHER METHODS

In order to further illustrate that the proposed method can better meet the requirements of student services and management, in this section, we select five popular algorithms from the five clustering categories stated in II-B, namely, partitioning algorithm K-means, density-based algorithm DBSCAN, hierarchical algorithm BIRCH, grid-based algorithm CLIQUE, and model-based algorithm EM, and use them to cluster dinner behavioral data, shopping behavioral data, library entry behavioral data, and gateway login behavioral data. As for the related methods in II-A, because the implementation of these methods is not available and the data used are completely different, we do not compare the proposed method with them.

BIRCH is a multiphase hierarchical clustering algorithm, in which a feature tree (CF-tree) is used to store the clustering features and makes the clustering method effective for incremental and dynamic data. BIRCH generally has two phases. In the initial microclustering stage, it dynamically builds an in-memory CF-tree that represent the data's inherent clustering structure. Once the CF-tree is built, any clustering algorithm such as a typical partitioning algorithm can be applied to cluster the leaf nodes of the CF-tree, which remove sparse clusters as outliers and groups dense clusters into large ones. This phase is called the macroclustering stage. CLIQUE (CLustering In QUEst) is a simple grid-based algorithm that finds density-based clusters in subspaces. It performs clustering in two steps. In the first step, CLIQUE partitions each dimension into nonoverlapping intervals, thereby partitioning all data objects into cells. CLIQUE uses a density threshold to identify dense cells and sparse ones. A cell is dense if the number of objects belonging to it exceeds the density threshold. In the second step, CLIQUE uses the maximal regions to cover connected dense cells. The maximal regions can be viewed as clusters, and the cells not belonging to any cluster can be viewed as noise. The EM algorithm is a model-based algorithm, which estimates the maximum likelihood parameters of a statistical model by iteratively performing two steps: the E step and the M step, until the clustering cannot be improved. In the E step, each object is assigned to the cluster based on the posterior distribution; in the M step, the parameters are re-estimated by maximizing the likelihood rule.

In general, finding the best clustering result is NP-Hard, and the clustering result is sensitive to the parameters of the clustering algorithm. In this experiment, we use SC as an evaluation metric of clustering results and the grid search method to find the optimal parameters for every algorithm. The implementation is based on the pyclustering library. Table 5 shows the clustering results of different algorithms on different behavioral data, in which we can see the number of clusters and the number of students in each cluster. For example, the gateway login behavior data are divided into 14 clusters by BIRCH, and there are 2221 students in cluster No. 0, 409 students in cluster No. 1, and so on. Usually, it is impossible to define a general standard to determine

TABLE 5. Clustering results of different algorithms on different behavioral data.

Algorithms	Dinner Behavior	Shopping Behavior	Library Entry Behavior	Gateway Login Behavior
K-Means	*0*:1180, *1*:1565, *2*:1828, *3*:878, *4*:1748, *5*:1985	*0*:1899, *1*:693, *2*:2395, *3*:2339, *4*:1115, *5*:602	*0*:1174, *1*:1011, *2*:1185, *3*:967, *4*:874	*0*:3672, *1*:2147, *2*:1011, *3*:2374
DBSCAN	*-1*:102, *0*:9046, *1*:36	*0*:8823, *1*:11, *2*:162, *3*:40, *4*:7	*-1*:51, *0*:5156, *1*:4	*-1*:215, *0*:7608, *1*:503, *2*:464, *3*:367, *4*:12, *5*:11, *6*:13, *7*:11
BIRCH	*0*:8422, *1*:762	*0*:5571, *1*:3472	*0*:2046, *1*:3165	*0*:2221, *1*:409, *2*:124, *3*:197, *4*:6, *5*:66, *6*:225, *7*: 1903, *8*:2916, *9*:1, *10*:599, *11*:4, *12*:315, 13:218
CLIQUE	*0*:9183, *1*:1	*-1*:31, *0*:9012	*-1*:167, *0*:5047	*-1*:17, *0*:9187
EM	*0*:49, *1*:3927, *2*:5144, *3*:64	*0*:5807, *1*:3236	*0*:4072, *1*:1139	*0*: 9204

TABLE 6. The ability to detect anomalous and mainstream patterns of different algorithms.

Algorithms	Dinner Behavior		Shopping Behavior		Library Entry Behavior		Gateway Login Behavior	
	Anom.	Main.	Anom.	Main.	Anom.	Main.	Anom.	Main.
K-Means	No	Yes	No	Yes	No	Yes	No	Yes
DBSCAN	Yes	No	Yes	No	Yes	No	Yes	No
BIRCH	No	No	No	Yes	No	Yes	Yes	Yes
CLIQUE	Yes	No	Yes	No	Yes	No	Yes	No
EM	Yes	Yes	No	Yes	No	Yes	No	No
Proposed Method	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

which clusters are anomalous and which are mainstream. Here, if a clustering algorithm can find clusters with less than 5% of the total number of students in specific behavioral data, we believe that the algorithm has the ability to detect anomalous patterns for given data, and if there is no single cluster with more than 80% of the total number of students, we see that the algorithm can find several mainstream patterns for better services and management. According to the defined specific rules, the ability to detect anomalous patterns and mainstream ones of different algorithms are shown in Table 6. The experimental results show that the ensemble method proposed in this paper is more stable and flexible because it can always detect anomalies and can obtain the mainstream patterns by manually setting the amount of clustering.

VIII. CONCLUSION

This paper proposed an ensemble unsupervised clustering framework for the analysis of students' behavioral patterns by combining DBSCAN and k-means algorithms. To evaluate the effect of the proposed method, we collect six types of behavioral data produced by 9024 undergraduates on campus and extract behavioral features through the two aspects of statistics and entropy. The experimental results demonstrate that the proposed method can not only detect anomalous behavioral patterns but also more precisely identify mainstream behavioral patterns. Based on the clustering results, student departments can adopt more targeted measures for intervention and specialized services. At the end of the paper, we discuss three issues: whether we can cluster the behavioral feature space using only the k-means algorithm, the difficulty of applying the proposed method to high-dimensional multisource behavior features, and the relationship between different behavioral patterns and academic performance levels. For better clustering analysis, future work should include the following: (1) extract more meaningful features by fully fusing multisource behavioral data; (2) design a new distance measure to make the proposed method effective for high-dimensional feature spaces; and (3) further study the relationship between behavioral patterns and student labels, such as academic performance, psychological state, and employment domain.

REFERENCES

[1] A. H. Eliasson, C. J. Lettieri, and A. H. Eliasson, "Early to bed, early to rise! Sleep habits and academic performance in college students," *Sleep Breathing*, vol. 14, no. 1, pp. 71–75, Feb. 2010, doi: 10.1007/s11325-009-0282-2.

[2] X. D. Keating, D. Castelli, and S. F. Ayers, "Association of weekly strength exercise frequency and academic performance among students at a large university in the united states," *J. Strength Conditioning Res.*, vol. 27, no. 7, pp. 1988–1993, Jul. 2013, doi: 10.1519/JSC.0b013e318276bb4c.

[3] M. Valladares, E. Duran, A. Matheus, S. Duran-Agueero, A. M. Obregon, and R. Ramirez-Tagle, "Association between eating behavior and academic performance in university students," *J. Amer. College Nutrition*, vol. 35, no. 8, pp. 699–703, 2016, doi: 10.1080/07315724.2016.1157526.

[4] J. Filippou, C. Cheong, and F. Cheong, "Modelling the impact of study behaviours on academic performance to inform the design of a persuasive system," *Inf. Manage.*, vol. 53, no. 7, pp. 892–903, Nov. 2016, doi: 10.1016/j.im.2016.05.002.

[5] S. Ghosh and S. K. Ghosh, "Exploring the association between mobility behaviours and academic performances of students: A context-aware trajgraph (CTG) analysis," *Prog. Artif. Intell.*, vol. 7, no. 4, pp. 307–326, Dec. 2018, doi: 10.1007/s13748-018-0164-6.

[6] Z. Yang, X. Mo, D. Shi, and R. Wang, "Mining relationships between mental health, academic performance and human behaviour," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, USA, Aug. 2017, pp. 1–8.

[7] T. Phan, S. G. McNeil, and B. R. Robin, "Students' patterns of engagement and course performance in a massive open online course," *Comput. Edu.*, vol. 95, pp. 36–44, Apr. 2016, doi: 10.1016/j.compedu.2015.11.015.

[8] I. JO, Y. Park, J. Kim, and J. Song, "Analysis of online behavior and prediction of learning performance in blended learning environments," *Educ. Technol. Int.*, vol. 15, no. 2, pp. 71–88, 2014.

[9] G. Kostopoulos, S. Kotsiantis, N. Fazakis, G. Koutsonikos, and C. Pierrakeas, "A semi-supervised regression algorithm for grade prediction of students in distance learning courses," *Int. J. Artif. Intell. Tools*, vol. 28, no. 4, Jun. 2019, Art. no. 1940001, doi: 10.1142/S0218213019400013.

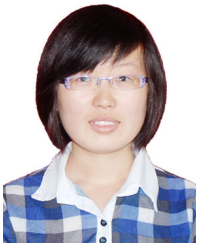
[10] D. Hooshyar, M. Pedaste, and Y. Yang, "Mining educational data to predict Students' performance through procrastination behavior," *Entropy*, vol. 22, no. 1, p. 12, Dec. 2019, doi: 10.3390/e22010012.

[11] N. Iam-On and T. Boongoen, "Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 2, pp. 497–510, Apr. 2017, doi: 10.1007/s13042-015-0341-x.
[12] I. HarwatiR Virdyanawaty and A. Mansur, "Drop out estimation students based on the study period: Comparison between naive Bayes and support vector machines algorithm methods," in *Proc. ICET4SD*, Yogyakarta, IN, USA, 2015.
[13] P. Aparicio-Chueca, I. Maestro-Yarza, and M. Domínguez-Amorós, "Academic profile of students who drop out a degree. A case study of faculty of economics and business, UB," in *Proc. EDULEARN*, Barcelona, Spain, Jul. 2016.
[14] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.
[15] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 4, pp. 558–568, Jul. 2006, doi: 10.1109/TVCG.2006.76.
[16] Y. Cao, J. Gao, D. Lian, Z. Rong, J. Shi, Q. Wang, Y. Wu, H. Yao, and T. Zhou, "Orderliness predicts academic performance: Behavioural analysis on campus lifestyle," *J. Roy. Soc. Interface*, vol. 15, no. 146, Sep. 2018, Art. no. 20180210, doi: 10.1098/rsif.2018.0210.
[17] H. Yao, D. Lian, Y. Cao, Y. Wu, and T. Zhou, "Predicting academic performance for college students: A campus behavior perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 3, pp. 1–21, May 2019, doi: 10.1145/3299087.
[18] F. Li, X. Long, S. Du, J. Zhang, Z. Liu, M. Li, F. Li, Z. Gui, and H. Yu, "Analyzing campus mobility patterns of college students by using GPS trajectory data and graph-based approach," in *Proc. 23rd Int. Conf. Geoinformatics*, Wuhan, China, Jun. 2015.
[19] M. J. Lesot, "Outlier preserving clustering for structured data through kernels," in *Proc. 29th Annu. Conf. German-Classification-Soc.*, Magdeburg, Germany, 2005, pp. 462–469.
[20] S. Fan, P. Li, T. Liu, and Y. Chen, "Population behavior analysis of chinese university students via digital campus cards," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Atlantic, NJ, USA, Nov. 2015, pp. 72–77.
[21] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014, doi: 10.1109/TETC.2014.2330519.
[22] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, San Jose, CA, USA, May 2016, pp. 225–236.
[23] Y. Wang, K. Qin, Y. Chen, and P. Zhao, "Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data," *ISPRS Int. J. Geo-Information*, vol. 7, no. 1, p. 25, Jan. 2018, doi: 10.3390/ijgi7010025.
[24] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.
[25] Z.-H. Zhou, "Clustering ensembles," in *Ensemble Methods Foundations and Algorithms*, 1st ed. Boca Raton, FL, USA: CRC Press, 2012, pp. 135–155.

XIAOYONG LI was born in Shanxi, China. He received the M.S. degree in computer science from the Beijing University of Technology, Beijing, China, in 2009, where he is currently pursuing the Ph.D. degree in computer science with the Beijing Key Laboratory of Multimedia and Intelligent Software Technology.

Since 2009, he has been an Engineer with the Information Technology Support Center, Beijing University of Technology. His research interests include big data analysis and visualization.

YONG ZHANG (Member, IEEE) received the Ph.D. degree from the Beijing University of Technology, Beijing, China, in 2010. He is currently an Associate Professor with the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology. His research interests include intelligent transportation systems, big data analysis and visualization, and computer graphics.



HUIMIN CHENG received the B.S. degree in computer science and technology from Yantai University, Yantai, China, in 2011, and the M.S. degree in computer technology from the Beijing University of Technology, Beijing, China, in 2014. She currently works with the Information Technology Support Center, Beijing University of Technology. Her research interests include data analysis and software development.

FEIFEI ZHOU is currently pursuing the M.S. degree with the Beijing University of Technology, Beijing, China. Her current research interests include data analysis and visualization.

BAOCAI YIN (Member, IEEE) received the M.S. and Ph.D. degrees in computational mathematics from the Dalian University of Technology, Dalian, China, in 1988 and 1993, respectively. He is currently a Professor with the Faculty of Information Technology, BJUT. He has authored or coauthored more than 200 academic articles in prestigious international journals, including the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FORVIDEO TECHNOLOGY, and top-level conferences, such as INFOCOM and ACM SIGGRAPH. His research interests include multimedia, image processing, computer vision, and pattern recognition.