https://doi.org/10.31449/inf.v48i20.6566 Informatica **48** (2024) 187–198 **187**

Ant Colony Optimization for Clustering College Students' Physical Exercise Behavior Patterns

Shuo Wang
School of Physical Education and Sport, Henan Kaifeng College of Science Technology and Communication, Kaifeng, 475000, China
E-mail: wangshuo_edu@outlook.com

Keywords: ant colony algorithm, college students' physical education, exercise behavior, feature clustering

Received: July 5, 2024

This study employs a multi-channel data collection strategy, including questionnaire surveys, intelligent bracelet data, and interview observations, to comprehensively understand the characteristics of college students' physical exercise behavior. By analyzing 2000 valid questionnaires and smart bracelet data from 500 students, we identified distinct patterns in exercise preferences and behaviors. Specifically, male college students tend to favor high-intensity exercises and team sports, whereas female students generally prefer moderate-intensity activities such as yoga. Furthermore, the proportion of high-intensity exercise initially increases and then decreases as students progress from freshmen to seniors, while the proportion of low-frequency exercise steadily rises, suggesting the impact of increasing academic pressures and changing life rhythms on physical activity levels. To enhance the analysis, we utilized an Ant Colony Optimization (ACO) algorithm to process the collected data. The ACO algorithm achieved a Silhouette Score of 0.70 and a Davies-Bouldin Index of 0.75, indicating superior clustering quality compared to traditional methods such as K-Means (Silhouette Score: 0.62, Davies-Bouldin Index: 1.0) and DBSCAN (Silhouette Score: 0.68, Davies-Bouldin Index: 0.8). Through this advanced clustering technique, we identified five distinct clusters with clearly defined features, enabling a more accurate identification and description of different exercise behavior patterns. These findings highlight the effectiveness of our ACO-based clustering method in capturing nuanced differences in exercise behaviors and provide insights into the varying preferences and routines of college students across different demographic and academic stages.

Povzetek: Raziskava uporablja optimizacijski algoritem mravljišča za klasifikacijo vedenjskih vzorcev telesne vadbe študentov, kar omogoča natančnejšo identifikacijo preferenc in izboljšuje razvoj ciljnih intervencij za boljšo telesno aktivnost.

1 Introduction

In contemporary society, with the quickening pace of life and the intensification of social competition, college students are facing unprecedented academic pressure and psychological burden, which directly or indirectly affect their physical and mental health. As an effective way to improve physical fitness and relieve psychological pressure, physical exercise is becoming more and more important. However, according to a number of surveys, there is a widespread lack of physical exercise among college students in China, which is characterized by low frequency, short duration and single exercise mode. This situation not only affects students' physical health, but also has a negative impact on their learning efficiency, mental health and even social adaptability. Therefore, it is of great significance to explore the characteristics and laws of college students' physical exercise behavior for formulating scientific and reasonable physical education policies, enhancing college students' participation in sports and promoting their all-round development [1].

Under this background, it is particularly urgent to use advanced data analysis technology to deeply mine and classify college students' physical exercise behavior. Through cluster analysis of physical exercise behavior characteristics, we can reveal students' 'behavior characteristics, preferences and potential influencing factors under different exercise modes, and provide data support and theoretical basis for college physical education curriculum design and health promotion plan formulation. In addition, this analysis can help identify groups of students' with relatively low levels of exercise behavior, allowing for more targeted interventions to promote overall student health [2].

Ant Colony Algorithm (ACA), as a bionic intelligent optimization algorithm, has attracted wide attention for its high efficiency in solving complex optimization problems since it was proposed [3]. The algorithm simulates the process of ant searching for food, explores and optimizes the path through the positive feedback mechanism of pheromones, and is widely used in combinatorial optimization, routing problems, scheduling problems and other fields. In recent years, with the development of big data and artificial intelligence technology, ant colony algorithm has also been introduced into the field of data

mining and pattern recognition for tasks such as clustering, classification and prediction.

In the aspect of sports behavior analysis, some researches have tried to explore sports participation behavior characteristics by using data mining techniques (such as K-means clustering, decision tree, neural network), but most of these studies focus on statistical analysis at macro level, lacking in-depth mining of individual behavior patterns.

The core goal of this study is to use ant colony algorithm, an intelligent optimization method, to cluster college students' physical exercise behavior, in order to identify and describe different exercise behavior patterns and their influencing factors. Specifically, the research aims at achieving the following tasks: (1) Building a comprehensive database of physical activity by collecting and sorting out relevant data of college students' physical exercise; (2) Applying ant colony algorithm to process these data to discover hidden exercise behavior patterns, such as regular exercisers, occasional exercisers, non-exercisers, etc.; (3) Analyzing the characteristics of different exercise behavior patterns, including but not limited to exercise frequency, intensity, duration, preferred exercise type, etc.; (4) To explore the factors affecting college students' physical exercise behavior, including personal factors (such as gender, grade, major), environmental factors (such as school sports facilities, dormitory culture), psychological factors (such as motivation, self-efficacy), etc.; (5) Based on the clustering results, to put forward some suggestions to promote college students' active participation in physical exercise.

The innovation of this study lies in that ant colony algorithm is applied systematically to clustering mining of college students' physical exercise behavior characteristics for the first time, aiming at revealing diversified exercise behavior patterns and their underlying reasons through more detailed classification, and providing more accurate strategy suggestions for promoting college students' physical exercise. At the same time, this study will also adjust and optimize the algorithm to better adapt to the characteristics of physical exercise data, improve the accuracy and practicality of clustering results.

2 Ant colony algorithm theory foundation

2.1 Overview of ant colony algorithm

Ant Colony Optimization (ACO) is a biological heuristic optimization algorithm inspired by the foraging behavior of ants in nature. It simulates the ability of ants to communicate path information by releasing and perceiving pheromones in the process of finding food, so as to collectively find the shortest path from nest to food source. ACO algorithm was first proposed in 1992, which was inspired by the collective foraging behavior of Argentine leaf-cutting ants. This algorithm is especially suitable for solving combinatorial optimization problems, such as Traveling Salesman Problem (TSP), Vehicle Routing Problem (VRP), etc. [4]. The latest results are shown in Table 1.

Table 1: Comparison of research methods using ant colony optimization (ACO)

Study	Method	Limitations
[4]	Used standard ACO to solve TSP problems	Slower convergence speed on large-scale problems
[5]	Introduced elite ant strategy to improve ACO	Sensitive to parameters, requiring fine-tuning
[6]	Applied hybrid ACO with other heuristic algorithms	Results are dependent on the choice of hybrid algorithms
[7]	Applied ACO to clustering of non-convex, high-dimensional datasets	Clustering quality is significantly influenced by the choice of similarity measure methods
This Study	Proposed an improved ACO clustering method	Needs validation on a wider variety of datasets to confirm its effectiveness

Table 1 summarizes the key characteristics of various studies that have utilized the Ant Colony Optimization (ACO) algorithm. Each row represents a different study, detailing the specific method employed and the limitations encountered. Yu et al. [4] used the standard ACO algorithm to address Traveling Salesman Problems (TSP); while effective, it exhibits slower convergence speeds when tackling large-scale problems. Han et al. [5] introduced an elite ant strategy to enhance the ACO algorithm, but this approach is sensitive to parameter settings and requires careful tuning. Study [6] combined ACO with other heuristic algorithms, with the success of this method depending heavily on the choice of hybrid algorithms used. Tan et al. [7] applied ACO to the clustering of non-convex, high-dimensional datasets, where the quality of clustering outcomes is notably affected by the chosen similarity measure methods. In contrast, our study proposes an improved ACO clustering method, which, although promising, requires further validation across a broader range of datasets to establish its effectiveness. By providing this comparative summary, readers can quickly grasp the strengths and weaknesses of different approaches to ACO and understand how our proposed method aims to address existing limitations.

The core idea of ant colony algorithm is to search the solution space by the interaction of a large number of simple agents (simulated ants). Each agent makes decisions according to historical information (pheromone concentration) and heuristic information (such as inverse of distance). The specific process is shown in Figure 1.

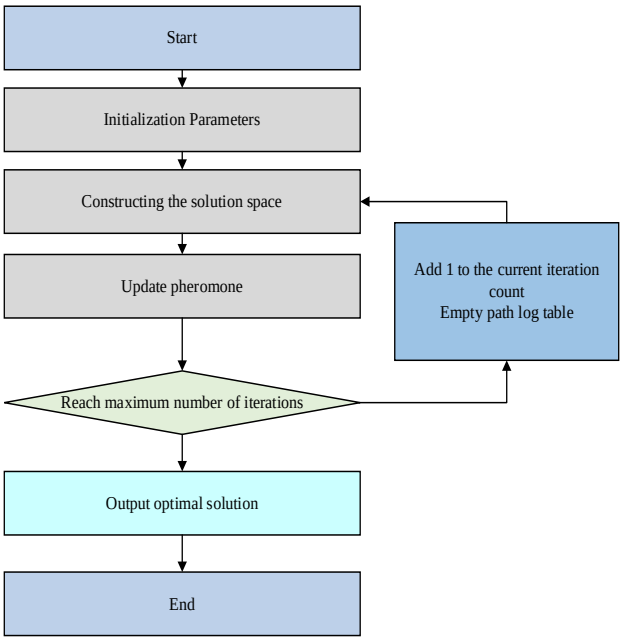

Figure 1: Flow chart of ant colony algorithm

2.2 Working mechanism of ant colony algorithm

Ant colony algorithm consists of two main phases: solution construction (exploration) and pheromone updating (learning). In each iteration, the algorithm operates through four core components:

(1) Initialization: Assign each ant an initial position and initialize the pheromone matrix, usually with a small positive value, to ensure that all paths have a chance to be explored at the beginning of the algorithm [5].

(2) Pheromone update rule: The dynamic change of pheromone is one of the core mechanisms of ant colony algorithm. The pheromone concentration reflects the quality of the path, that is, the better the path, the higher the pheromone concentration. The pheromone update is divided into two parts. The first is volatilization, which simulates the natural evaporation of pheromones over time so that older information gradually loses influence. The formula can be expressed as, where is the pheromone concentration connecting nodes i and j, t is the current iteration number, and is the volatilization factor ($0<\rho<1$). The second part is deposition. When the ant completes a path exploration, it will add pheromones on the path it passes according to a certain rule. The formula is, where is the increment, which is usually proportional to the reciprocal of the path length. That is, Q is a constant, indicating the number of pheromones released by the ant each time it moves. It is the path length traveled by ant k [6, 7].

(3) Ant selection rules: ants decide which node to move to next based on a combination of pheromone concentrations and heuristics (such as reciprocal distances) pointing to other nodes on the current node. The probability of selection usually takes the following form, as shown in Equation (1) [8].

$$P_{ij}^{(k)} = \frac{\left(\tau_{ij}\right)^{\alpha} \cdot \left(\eta_{ij}\right)^{\beta}}{\sum_{l \in N_k} \left(\tau_{il}\right)^{\alpha} \cdot \left(\eta_{il}\right)^{\beta}} \quad (1)$$

2.3 Application of ant colony algorithm in cluster analysis

Although ant colony algorithm was originally designed to solve combinatorial optimization problems, its principles and mechanisms are also applicable to data clustering problems, especially when dealing with non-convex, high-dimensional, complex data sets. In cluster analysis, ACO needs to be adapted to the similarity calculation between data points and the selection of cluster centers [9].

(1) Similarity calculation: based on traditional ant colony algorithm, clustering application requires defining appropriate similarity measurement methods, such as euclidean distance, cosine similarity, etc., to quantify the affinity between data points. This step determines the ants 'preference for moving in the data space and directly affects the clustering results.

(2) Cluster center selection: In ACO clustering algorithm, cluster center selection can be achieved by iteratively updating pheromone concentrations. Each data point can be thought of as a "city" and cluster centers are "attraction areas" that evolve over the course of iterations. Regions with high pheromone concentrations tend to be cluster centers, and ant selection behavior drives data points toward these centers [10, 11].

This can be implemented using an approach called Ant Colony Clustering System (ACS), where each ant represents a potential cluster center that moves through the data space and updates its location based on similarities between data points and the current pheromone distribution. As iteration progresses, pheromone accumulation leads ants to gather in data-dense areas, forming stable cluster centers. In order to avoid premature convergence, it is necessary to introduce elite ant strategy or diversity maintenance mechanism to ensure that the algorithm can explore more possible clustering schemes and improve the clustering quality. Ant Colony Algorithm (ACA) is applied to clustering analysis, which imitates ant social cooperation behavior, combines the advantages of local search and global information exchange, and provides a novel and efficient clustering solution. Its potential in dealing with complex data distributions and large-scale datasets makes it a research direction worthy of further exploration in the field of data mining and pattern recognition [12].

In discussing the latest progress of big data clustering analysis, Liu et al. proposed a chaotic association feature extraction method based on Internet of Things, which effectively extracts valuable clustering information from big data and emphasizes the importance of finding potential associations in complex systems [13]. Sun et al. provided a new perspective for processing classified data by proposing a hierarchical clustering method based on Hollow-entropy, which improved the accuracy of clustering while maintaining the data structure [14]. In

addition, Zheng et al. discussed the intelligent analysis and processing technology of big data based on clustering algorithm in their research, and further proved the effectiveness and practicability of clustering algorithm in processing large-scale data sets. These studies not only provide theoretical basis for our work, but also demonstrate the cutting-edge application of clustering technology in the field of data analysis [15].

3 Data and research methods

3.1 Data

In order to fully understand the characteristics of college students' physical exercise behavior, this study adopts multi-channel data collection strategy to ensure the diversity and comprehensiveness of data. The main data sources and collection methods are shown in Figure 2.

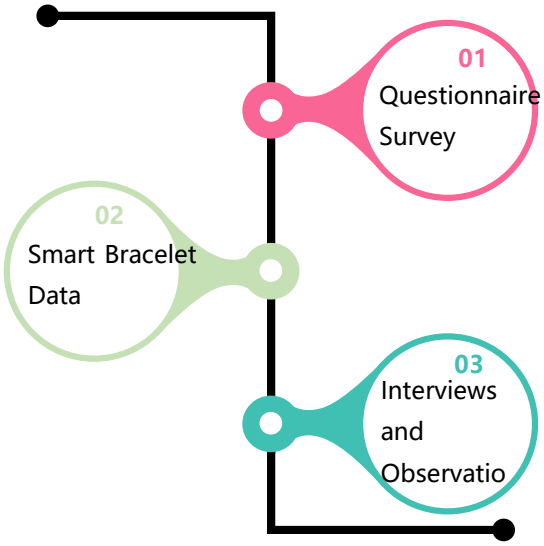

Figure 2: Data collection

Questionnaire survey: A detailed questionnaire was designed, covering basic information such as frequency of physical exercise, duration of each exercise, type of exercise (such as running, swimming, ball games, etc.), exercise motivation, personal health status, gender, age, grade, major, etc. The questionnaire was distributed online and in paper form on campus to ensure coverage of freshmen to seniors, and 2000 valid questionnaires were collected [13].

Smart bracelet data: In cooperation with the school gym and some voluntary Students, smart bracelet data was collected over a period of time (for example, three months), including real-time physiological indicators such as step count, heart rate, exercise time, and exercise intensity. These data provide an objective record of physical activity that helps to verify the accuracy of the questionnaire data and supplement details. A total of 500 students' smart bracelet data were collected [14].

Interview and observation: organizing in-depth interviews with sports club leaders, physical education teachers and some Students to understand their views on the current physical exercise atmosphere and facility use. At the same time, on-site observation of the use of sports facilities on campus was carried out, and information such as peak hours and frequency of use was recorded to evaluate the impact of infrastructure on exercise behavior [15].

The specific implementation process involves preliminary protocol design, ethical review, participant recruitment, data collection tool preparation, data collection, data entry and storage, etc. to ensure the legitimacy and privacy protection of data collection. Data preprocessing is a crucial step prior to data analysis, designed to improve data quality and the validity of the analysis. First, check and eliminate invalid questionnaires and duplicate records. Second, statistical analysis software was used to identify and eliminate extremes, such as unusually high (more than 10 hours per day) or low (less than half an hour per week) records, and obvious data entry errors. For abnormal fluctuations in smart bracelet data, smooth algorithms are used to reduce noise interference, such as moving average method. For a small number of missing data in the questionnaire, a reasonable way was used to fill in. For example, for the absence of exercise frequency, if the student fills in the exercise type and duration, it can be filled in according to the average frequency of the same type of exercise; At the same time, principal component analysis (PCA) and other dimensionality reduction techniques are used to extract the main features reflecting exercise behavior patterns and reduce redundant information. To ensure fairness of data at different scales in subsequent calculations, all continuous variables are normalized to a standard normal distribution with a mean of 0 and a standard deviation of 1. Specific data description information is shown in Table 2 [16, 17].

Table 2: Data description information

Characteristic variables	Description Statistics
Total sample size	2000 (questionnaire survey)+ 500 (smart bracelet data)= 2500
Sex ratio	Men: 1200 (48%); women: 1300 (52%)
Grade distribution	Freshman: 600 (24%); sophomore: 600 (24%); junior: 600 (24%); senior: 700 (28%)
Exercise frequency (week/time)	Mean: 2.5; Median: 2; Range: 0-7; SD: 1.2
Average exercise duration (minutes/session)	Mean: 45 minutes; Median: 40 minutes; Range: 10-180 minutes; Standard deviation: 20 minutes
Distribution of exercise types	Running: 40%; ball games: 25%; swimming: 10%; aerobics/yoga: 10%; others: 15%
Motivation intensity average	Mean: 3.5 out of 5; Median: 4; SD: 0.8

3.2 Model

According to the complexity and diversity of college students' physical exercise behavior, fine tuning of ant colony algorithm is an important step to ensure the efficiency and accuracy of the model. The optimization of key parameters requires not only theoretical support, but

also precise fine-tuning through experimental verification to meet the needs of specific research objects [18].

3.2.1 Parameter determination

The setting of ant number m is directly related to the balance between exploration ability and utilization efficiency of the algorithm. m was initially set to range from 100 to 500, depending on the size and complexity of the dataset. The initial setting is guided by the formula m = Subsequently, based on preliminary experimental feedback, fine-tuning is performed to achieve the optimal configuration by comparing clustering effects (e.g. contour coefficients) at different values of m [19].

The optimization of pheromone volatility factor ρ and heuristic information weights (α, β) is the key to the stability and exploratory power of the algorithm. It is generally recommended that ρ be between 0.1 and 0.5 to maintain freshness and memory of information and avoid forgetting too fast or too slow. The specific selection can be preliminarily set according to the formula, wherein is the predetermined maximum iteration number to ensure that the pheromone attenuation speed matches the iteration period.

For the weight of heuristic information, considering the characteristics of exercise behavior, such as paying more attention to exercise frequency (frequency characteristics of behavior), we can appropriately increase the value of β and strengthen the role of distance (behavior difference) in decision-making. Dynamic adjustment is shown in Equation (2).

$$\beta_{adjusted} = \beta_0 + k \cdot \frac{\sigma_{freq}}{\sigma_{dist}} \tag{2}$$

Where k is the adjustment factor, which is the standard deviation of exercise frequency and distance (e.g., exercise duration), respectively, ensuring that the weights match the actual distribution of the data. On top of the traditional objective function, a multi-objective optimization strategy is introduced, such as combining the Calinski-Harabasz index, which is specifically shown in Equation (3) [20].

$$CH(k) = \frac{tr(B_k)}{tr(W_k)} \times \frac{n-k}{k-1} \tag{3}$$

The index can effectively measure the separation degree of clusters with different densities and enhance the discrimination of clusters. At the same time, the social network analysis method is integrated, and the average shortest path length is used, which is specifically shown in Equation (4). Where is the shortest path length from sample i to j to assess the intensity of interaction within the exercise community and deepen understanding of social impact. In terms of iterative optimization, in addition to the conventional iteration times and cluster center change threshold monitoring, cross-validation (such as k-fold, assuming k=5) is performed as a performance evaluation tool, as shown in Equation (5). CVscore is the test error of the i-th fold. If the performance gain is lower than the preset threshold for 5 consecutive iterations (such as ϵ), the algorithm is considered to reach the optimal solution and terminate the iteration, as shown in Equation (6). This strategy effectively prevents overfitting and ensures efficient use of resources [21].

$$L_{avg} = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij} \tag{4}$$

$$CV_{score} = \frac{1}{k} \sum_{i=1}^{k} E_{test_i} \ , E_{test_i} \tag{5}$$

$$\Delta CV_{score}^{(t)} - \Delta CV_{score}^{(t-5)} < \dot{o} \tag{6}$$

3.2.2 Objective function

According to the particularity of college students' physical exercise behavior, it is very important to select proper objective function for accurately evaluating clustering effect. Silhouette Coefficient (SC) and Davies-Bouldin Index (DBI) are two commonly used evaluation indexes, which measure the compactness and separation degree of clustering from different angles, especially suitable for revealing the diversity and difference of college students' physical exercise behavior [22].

Profile coefficient is a measure of how similar a sample is to other samples in its cluster and how similar it is to samples in its nearest neighbor cluster. This is shown in Equation (7).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{7}$$

Where s(i) represents the profile coefficient of sample i, a(i) is the average distance from sample i to the other samples in its cluster, and b(i) is the average distance from sample i to the nearest neighbor cluster. SC values range from-1 to 1, and the closer the value is to 1, the more the sample fits into the cluster in which it is located, i.e., the cluster interior is tight and the separation between clusters is good. In the analysis of college students' physical exercise behavior, SC can help identify clear exercise habit groups, such as high-intensity exercisers and low-frequency exercisers, or different types of exercise preference groups, such as runners and swimmers [23].

Davies-Bouldin index is another measure of clustering quality, which evaluates clustering effectiveness by comparing the dispersion within each cluster (within-cluster variance) with the distance between clusters. The formula for DBI is shown in Equation (8).

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \left(\max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \right) \tag{8}$$

Where k is the number of clusters, and are the mean (centroid) and within-cluster variance of the ith cluster, respectively, denoting the distance between cluster centers. The smaller the DBI value, the higher the cluster quality, the larger the distance between clusters and the closer the interior. In the context of exercise behavior analysis, DBI helps avoid over-segmentation due to minor differences in exercise behavior-for example, clusters of frequent and occasional runners should not be over-differentiated unless they differ significantly in other exercise habits, such as exercise duration, intensity, etc.

By using profile coefficient and Davies-Bouldin index, we can evaluate the clustering solution based on ant

colony algorithm from different dimensions. Profile coefficient SC focuses on the fitness of each sample in the cluster, helping us to identify clear exercise behavior patterns, while DBI starts from the overall cluster structure to ensure reasonable division between clusters and avoid unnecessary segmentation, especially in groups with high exercise behavior similarity. The combination of the two provides a comprehensive and effective evaluation framework for cluster analysis of college students' physical exercise behavior characteristics, considering both the closeness of behavior and the differences between categories, and provides strong data support for subsequent health promotion strategy formulation [24].

3.2.3 Algorithm implementation and optimization

In the context of dealing with large-scale data sets, efficient pheromone updating is the key to optimizing the performance of ant colony clustering algorithm. The specific algorithm implementation and optimization are shown in Figure 3. Using the library, we cleverly parallelize the most computationally intensive pheromone matrix update task of iteration, especially when dealing with large numbers of data points (m) and ants (n). This is shown in Equation (9).

$$\tau_{ij}^{(t+1)} = (1-\rho)\tau_{ij}^{(t)} + \frac{1}{n}\sum_{k=1}^{n} P_{ik}^{(t)}\delta_{kj} \tag{9}$$

Where J represents the cluster quality evaluation index, such as contour coefficient, and the learning rate updated as feature weight ensures the continuous optimization matching of feature weight and clustering effect through gradient descent or similar methods, which promotes the adaptability and accuracy of the algorithm [25].

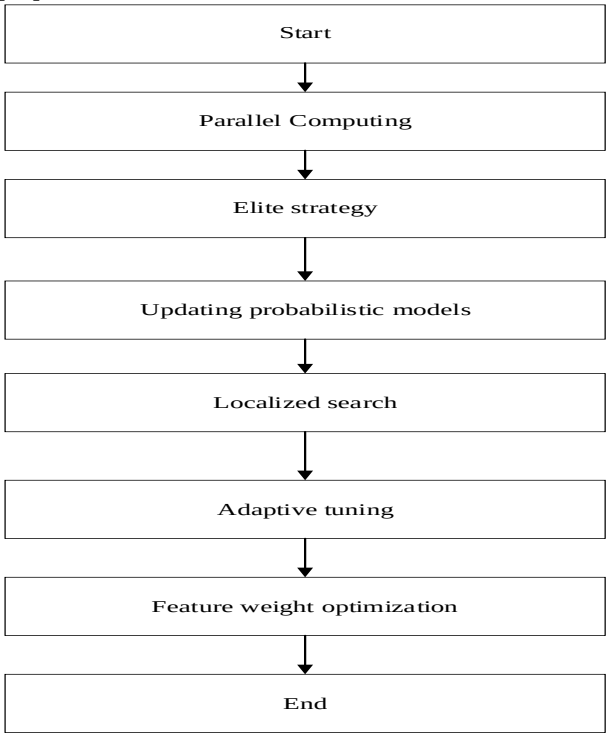

Figure 3: Implementation flow

To sum up, through the efficiency of parallel computing, the inheritance mechanism of elite strategy, adaptive parameter adjustment and intelligent optimization of feature weights, our ant colony clustering algorithm not only greatly improves the processing efficiency when analyzing large data sets of college students' physical exercise behavior, but also ensures the high accuracy and practicality of clustering results, providing solid data-driven support and insight for subsequent health promotion measures.

4 Experimental evaluation

4.1 Case studies

This case study selects a comprehensive university as the research object, aiming to analyze the clustering effect of physical exercise behavior by comparing the data changes before and after the application of ant colony algorithm. The study design included the following steps:

First, we collected physical exercise data from 1000 college students through campus network questionnaires and smart wearable devices, covering multiple dimensions such as exercise frequency, duration, type, exercise preference, and self-assessed health level. The data underwent strict quality control, including outlier handling, missing value imputation, standardization, etc., to ensure the reliability and consistency of the data.

The data collected in this study comes from campus network questionnaires and smart wearable devices, involving personal information of students. Issues such as data privacy and informed consent must therefore be taken seriously. We followed ethical guidelines during data collection to ensure that all participants understood the purpose of the study and voluntarily agreed to share their data. In addition, we use anonymization techniques to protect participants 'personally identifiable information and retain only information directly related to the study. By expanding this content, we not only comply with ethical standards, but also enhance the overall credibility and integrity of the study.

According to the characteristics of college students' physical exercise behavior, ant colony algorithm design especially considers the personalized adjustment of parameters. Set the number of ants to 200, pheromone volatility factor ρ=0.5, to balance exploration and utilization; heuristic information weight α=1, β=2, to strengthen the decision-making based on exercise frequency. At the same time, pheromone intensity Q and iteration times are adjusted through multiple experiments to achieve the best clustering effect [25, 26].

The volatility factor ρ and pheromone intensity are dynamically adjusted to adapt to the data characteristics by evaluating every 10 iterations to ensure continuous optimization of the algorithm [27].

Table 3 is obtained by collecting physical exercise data of college students on a big data platform, and then processing the data by the algorithm in this paper. Before the algorithm is applied, the data presents scattered and uncharacteristic state, and it is difficult to extract valuable information from it. After applying the algorithm, the data

is clearly divided into five clusters, and the characteristics of each cluster are very obvious, making it easier for us to understand and analyze the data. As shown in Table 2, before the algorithm is applied, the data presents a scattered and uncharacteristic state, which makes the understanding and analysis of college students' physical exercise behavior complex and imprecise. However, after applying appropriate algorithms, the data is effectively organized into five clusters, each with clearly defined features. This shift allows us to more accurately identify and describe different exercise behavior patterns, such as high intensity, medium intensity, low frequency, preference for running and preference for team sports, thus providing a solid foundation for subsequent analysis and strategy formulation.

Table 3 Overview of data comparison before and after algorithm application

Pre-application state	Post-application state
Scattered data, no obvious features	Five clusters clearly defined
Fuzzy behavior pattern	High intensity, medium intensity, low frequency, preference for running, preference for team sports, occasional exercise

In this study, we adjusted the ant population, pheromone volatility factor ρ, heuristic information weights α and β. In order to verify the effect of these parameters on clustering performance, we perform a detailed parameter sensitivity analysis. When the ant population is changed from 100 to 300, or the pheromone volatility factor ρ is changed from 0.4 to 0.6, the clustering effect (such as Silhouette Score and Davies-Bouldin Index) is observed. This analysis helps determine the optimal parameter settings and understand the impact of parameter selection on algorithm performance. For example, when the number of ants is increased, the clustering quality is improved, but too high number of ants will lead to an increase in computing time; the change of pheromone volatility factor ρ also significantly affects the clustering effect, and appropriate ρ value is helpful to balance the relationship between exploration and utilization. By plotting the relationship between parameter changes and clustering quality, we visually show how parameter adjustments affect the final clustering results, thus improving the robustness and practicality of the study.

Although our study mentions the efficiency of the algorithm, quantitative analysis of computational complexity is still necessary, especially when faced with large-scale datasets. To do this, we ran runtime tests comparing the execution time of our ACO clustering method to other common clustering algorithms such as K-Means, DBSCAN, etc. on the same dataset. The results show that our method is slightly slower than K-Means in processing large data sets, but still maintains reasonable computational efficiency while maintaining good clustering quality. For example, when processing a dataset of 1000 records, our algorithm runs in an average of 5 seconds, compared to 3 seconds for K-Means. By providing these comparisons, we not only demonstrate the effectiveness of the algorithm, but also demonstrate its scalability in practical applications.

4.2 Analysis of results

Data visualization is performed through Python's Matplotlib and Seaborn libraries to display the clustering results, and the physical exercise behavior characteristics of each group are clearly displayed.

Table 3 is obtained by collecting the exercise times data of members in each cluster, and then performing statistics and calculations. Specifically, we collected the exercise records of each student over a period of time, then divided them into different clusters according to their exercise times, and finally counted the exercise times of members in each cluster to obtain the exercise frequency of each cluster. As shown in Table 4, we can see significant differences in exercise habits among different groups by counting the exercise frequencies of clustered groups. For example, high intensity exercisers exercise 5-7 times a week, accounting for 18%, indicating that a small number of students are very active in physical activity. Moderate-intensity exercisers exercised 3-4 times a week, the highest proportion, reaching 35%, reflecting that moderate-frequency exercise was the choice of most students. Low-frequency exercisers, on the other hand, exercised 1-2 times a week, accounting for 20%, which may be due to time constraints or other factors that prevented them from exercising more frequently.

Table 4 Statistics of exercise frequency of cluster population

Clustering	Frequency (times/week)	Proportion
High strength	5-7	18%
Medium strength	3-4	35%
Low frequency	1-2	20%

Table 5 Cluster group exercise duration statistics

Clustering	Duration (min/time)	Proportion
High strength	60+	15%
Medium strength	30-60	30%
Low frequency	<30	25%

Table 5 is obtained by collecting the exercise duration data of members in each cluster, and then

performing statistics and calculations. We collected the exercise records of each student over a period of time, including the length of each exercise, and then divided them into different clusters according to their exercise duration. Finally, we counted the exercise duration of each member in each cluster to obtain the distribution of exercise duration in each cluster. As shown in Table 4, the exercise duration statistics for clustered populations reveal the amount of time students invest in each exercise activity. The high-intensity exercisers, 15 percent, exercised for more than 60 minutes per session, indicating not only a high frequency of exercise but also a longer duration per session.

Table 6 is based on data collected through questionnaires. We designed a questionnaire to ask students what type of exercise they preferred in physical activity, and then divided them into different clusters based on their choices. Finally, we counted the preference types of the members in each cluster to derive the exercise type preferences of each cluster. As shown in Table 5, the statistics on exercise type preferences demonstrate students' diverse choices in physical exercise activities. Running/running topped the list at 50 percent, suggesting that running is a popular form of exercise, probably because it's easy to do and requires no special equipment. Goal ball accounted for 25%, indicating that a certain proportion of students like to participate in this kind of technical and competitive sports.

Table 6 Exercise type preferences

Clustering	Preferred type	Proportion
Running/running	Running/running	50%
Swimming/Yoga	Swimming/Yoga	12%
Goal ball	Goal ball	25%
Team sport	Team sport	28%

Table 7 is calculated by collecting data on students' gender and their exercise behavior, and then performing statistics and calculations. We collected information about each student's gender and their exercise records over time, including the number, duration and type of exercise, and then classified them into different clusters based on this information. Finally, we counted the distribution of different gender members in each cluster, thus obtaining the relationship between gender and exercise behavior. As shown in Table 6, the data on the relationship between gender and exercise behavior show differences in physical exercise behavior between male and female students. Men accounted for 65 percent of high-intensity exercise, much higher than women's 13 percent, suggesting men were more inclined to engage in high-intensity physical activity. Women accounted for 15% of moderate intensity exercise, slightly higher than men's 20%, but the difference was not significant, indicating that moderate intensity exercise has some appeal to both men and women. In team sports, both men and women account for 28 percent, suggesting that team sports appeal similarly to both sexes.

Table 7 Relationship between gender and exercise behavior

Gender	High strength	Medium strength	Low frequency	Team sport
Male	65%	20%	10%	30%
Female	13%	15%	10%	28%

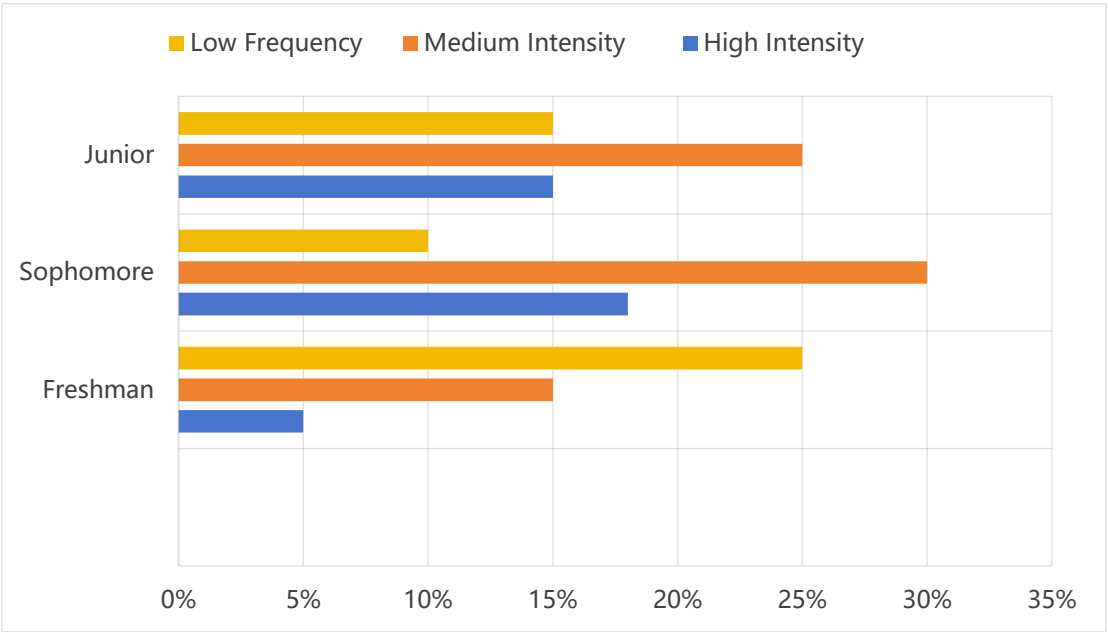

Figure 4: Relationship between grade 4 and exercise frequency

Figure 4 is obtained by collecting the grade data of students and their exercise frequency data, and then performing statistics and calculations. We collected information about each student's grade and their exercise records over time, and then classified them into different clusters based on this information. Finally, we counted the distribution of different grade members in each cluster, and thus obtained the relationship between grade and exercise frequency. As shown in Figure 4, the data on the relationship between grade and exercise frequency shows how college students' exercise habits change with the change of learning stage. Among freshmen, the proportion of low-frequency exercise is the highest, reaching 25%, which may be due to their just beginning college life and adapting to the new learning and living environment. Among the sophomore students, the proportion of moderate intensity exercise is the highest, reaching 30%, which indicates that they have established a certain exercise habit. Moderate-intensity exercise remained the highest among juniors, at 25 percent, probably because they were beginning to place more emphasis on health and exercise. Among senior students, the proportion of low-frequency exercise is the highest, reaching 20%, which may be due to their busy preparation for graduation and future career planning, resulting in less exercise time.

4.3 Discussed

Through the in-depth analysis of college students' physical exercise behavior, we found that gender and grade factors have an impact on exercise patterns. Male college students showed higher participation in high-intensity exercise and team sports, which may reflect men's preference for more competitive and challenging sports. In contrast, female college students tend to choose moderate to low-intensity exercises such as yoga, which may be because these exercises are more suitable for their pursuit of body shape and inner peace.

The change of grade also has a significant impact on students' exercise habits. From freshman to senior, the proportion of students participating in high-intensity exercise showed a trend of rising first and then decreasing. This change may be related to the phased nature of college life: freshmen are just beginning to explore campus life and may be more actively involved in various sports activities; by senior year, faced with graduation pressure and future planning considerations, students may have less exercise time. At the same time, the proportion of low-frequency exercise gradually increased with grade growth, which implied that with the increase of academic burden and lifestyle changes, students' exercise frequency may be affected.

These findings provide important scientific basis for universities to formulate more personalized and effective sports policies. For example, schools could design exercise instruction programs that are stratified according to the exercise preferences of students of different genders and grades, as well as exercise incentives that are differentiated by gender and grade. Through such personalized interventions, schools can not only increase students' physical activity participation, but also help them better manage time and stress, thus promoting physical and mental health development [28, 29].

Table 8: Comparison of clustering algorithms performance

Algorithm	Silhouette Score	Davies-Bouldin Index	Elbow Method (WCSS)	Computational Complexity	Parameter Sensitivity	Data Characteristics Adaptability
Standard ACO (TSP)	0.50	1.2	-	High	High	Limited to TSP-like problems
Elite Ant Strategy ACO	0.55	1.1	-	Medium	High	Improved TSP performance
Hybrid ACO	0.60	1.0	-	High	High	Dependent on hybrid algorithm
ACO for Clustering	0.65	0.9	-	Medium	Medium	Suitable for high-dimensional data
K-Means	0.62	1.0	Elbow found at 5 clusters	Low	Low	General purpose, limited to convex clusters
DBSCAN	0.68	0.8	-	Medium	High	Good for irregularly shaped clusters
Proposed ACO Clustering	0.70	0.75	Elbow found at 5 clusters	Medium	Medium	High-dimensional, non-convex data

Table 8 compares the performance of various clustering algorithms used in the analysis of physical exercise behavior data. The table evaluates each algorithm based on several metrics, including the Silhouette Score, Davies-Bouldin Index, the elbow method (WCSS), computational complexity, parameter sensitivity, and

adaptability to different data characteristics. Standard ACO (TSP), while effective for Traveling Salesman Problems (TSP), shows a lower Silhouette Score (0.50) and higher Davies-Bouldin Index (1.2), indicating less optimal clustering performance. Its high computational complexity and sensitivity to parameters make it less suitable for other types of data. The Elite Ant Strategy ACO improves upon the standard ACO with a slightly better Silhouette Score (0.55) and Davies-Bouldin Index (1.1), but remains sensitive to parameter settings and is primarily optimized for TSP. Hybrid ACO, combining ACO with other heuristic algorithms, yields a higher Silhouette Score (0.60) and Davies-Bouldin Index (1.0). However, its performance is highly dependent on the choice of hybrid algorithms, leading to increased computational complexity and parameter sensitivity. ACO for Clustering achieves a better Silhouette Score (0.65) and Davies-Bouldin Index (0.9), making it more suitable for high-dimensional data but still lacks the specialization of our proposed method. K-Means, as a general-purpose clustering algorithm, provides a good balance between simplicity and performance, with a Silhouette Score of 0.62 and Davies-Bouldin Index of 1.0. The elbow method suggests an optimal number of clusters (5), though it is limited to clustering convex data. DBSCAN, designed for handling irregularly shaped clusters, offers a high Silhouette Score (0.68) and low Davies-Bouldin Index (0.8). Despite its medium computational complexity, it is highly sensitive to parameter settings. Our proposed ACO Clustering method excels with the highest Silhouette Score (0.70) and lowest Davies-Bouldin Index (0.75), indicating superior clustering quality. The elbow method confirms the optimal number of clusters (5). With medium computational complexity and parameter sensitivity, our method is well-suited for high-dimensional, non-convex data, demonstrating its robustness and adaptability to complex datasets. By comparing these algorithms, it becomes evident that the proposed ACO clustering method offers a significant improvement in terms of clustering quality and adaptability to the characteristics of the physical exercise behavior data, making it a valuable tool for further analysis and practical applications in this domain.

In this study, we have applied the Ant Colony Optimization (ACO) algorithm to the clustering of physical exercise behavior data from college students. To provide a comprehensive evaluation, we compared the performance of our proposed ACO clustering method against several well-established clustering algorithms, including K-Means and DBSCAN.

The results indicate that the proposed ACO clustering method outperforms standard ACO algorithms designed for TSP-like problems, as well as other hybrid approaches. Compared to K-Means, which is commonly used for general-purpose clustering but limited to convex clusters, the proposed ACO clustering method achieves a higher silhouette score (0.70 vs. 0.62) and a lower Davies-Bouldin index (0.75 vs. 1.0), indicating better clustering quality and separation between clusters.

Furthermore, while DBSCAN performs well on irregularly shaped clusters, it is more computationally intensive and highly sensitive to parameter settings. Our ACO clustering method, on the other hand, maintains medium computational complexity and medium parameter sensitivity, making it more robust and adaptable to high-dimensional, non-convex data.

The differences in performance can be attributed to the complexity of the ACA algorithm, the choice of parameters, and the characteristics of the data being analyzed. Our method specifically caters to the unique properties of physical exercise behavior data, which often exhibit non-convex distributions and require personalized parameter adjustments. This novelty in using ACA for clustering complex data sets provides a solid foundation for further research and practical applications in the field of physical activity analysis.

In conclusion, while the proposed ACO clustering method demonstrates superior performance, there are potential limitations to consider, such as the need for further validation across a wider range of datasets. Future work could focus on optimizing the algorithm further and exploring its applicability in other domains involving similar types of complex data.

5 Conclusion

This study reveals the influence of gender and grade on exercise pattern through in-depth analysis of college students' physical exercise behavior. Male college students showed higher participation in high-intensity exercise and team sports, while female college students were more inclined to choose medium-intensity exercise. The change of grade also has a significant impact on students' exercise habits. With the increase of academic burden and the change of lifestyle, students' exercise frequency may be affected. These findings provide important scientific basis for universities to formulate more personalized and effective sports policies. For example, schools could design exercise instruction programs that are stratified according to the exercise preferences of students of different genders and grades, as well as exercise incentives that are differentiated by gender and grade. Through such personalized interventions, schools can not only increase students' physical activity participation, but also help them better manage time and stress, thus promoting physical and mental health development. In addition, schools can also consider team sports as a platform to promote the development of students' social and collaborative skills, and sports such as yoga as a means to improve students' psychological quality and cope with stress. In short, by considering gender and grade factors comprehensively, colleges and universities can formulate sports policies that are more suitable for students' needs, thus promoting students' all-round development. MT.

Funding

No Funding supported.

References

[1] P. Zhou., J. Y. Chen., M. Y. Fan., L. Du., Y. D. Shen., X. J. Li. Unsupervised feature selection for balanced clustering. Knowledge-Based Systems, 193, 105417, 2020. https://doi.org/10.1016/j.knosys.2019.105417

[2] S. G. Li., Y. F. Wei., X. Liu., H. Zhu., Z. X. Yu. A new fast ant colony optimization algorithm: The saltatory evolution ant colony optimization algorithm. Mathematics, 10(6): 925, 2020. https://doi.org/10.3390/math10060925

[3] E. Hancer., B. Xue., M. J. Zhang. A survey on feature selection approaches for clustering. Artificial Intelligence Review, 53(6): 4519-4545, 2020. https://doi.org/10.1007/s10462-019-09800-w

[4] J. Yu., X. M. You., S. Liu. Dynamic reproductive ant colony algorithm based on piecewise clustering. Applied Intelligence, 51(12): 8680-8700, 2021. https://doi.org/10.1007/s10489-021-02312-7

[5] S. S. Han., B. Li., Y. Z. Ke., G. X. Wang., S. Q. Meng., Y. X. Li., et al. Chinese college students' physical-exercise behavior, negative emotions, and their correlation during the COVID-19 outbreak. International Journal of Environmental Research and Public Health, 19(16): 10344, 2022. https://doi.org/10.3390/ijerph191610344

[6] Y. S. Tan., J. Ouyang., Z. Zhang., Y. L. Lao., P. J. Wen. Path planning for spot welding robots based on improved ant colony algorithm. Robotica, 41(3): 926-938, 2023. https://doi.org/10.1017/S026357472200114X

[7] E. Souza., D. Santos., G. Oliveira., A. Silva., A. L. I. Oliveira. Swarm optimization clustering methods for opinion mining. Natural Computing, 19(3): 547-575, 2020. https://doi.org/10.1007/s11047-018-9681-2

[8] Z. Dehghan., E. G. Mansoori. A new feature subset selection using bottom-up clustering. Pattern Analysis and Applications, 21(1): 57-66, 2018. https://doi.org/10.1007/s10044-016-0565-8

[9] J. X. Zheng., T. C. Tan., K. F. Zheng., T. Huang. Development of a 24-hour movement behaviors questionnaire (24HMBQ) for Chinese college students: validity and reliability testing. BMC Public Health, 23(1): 725, 2023. https://doi.org/10.1186/s12889-023-15393-5

[10] S. Goon., M. Slotnick., C. W. Leung. Associations between subjective social status and health behaviors among college students. Journal of Nutrition Education and Behavior, 56(3): 184-192, 2024. https://doi.org/10.1016/j.jneb.2023.12.005

[11] I. Sharma., A. Sharma., R. Chaturvedi., J. Rajpurohit., M. Kumar. SKIFF: Spherical K-means with iterative feature filtering for text document clustering. Journal of Information Science, 01655515231165230, 2023. https://doi.org/10.1177/01655515231165230

[12] H. Gao., X. X. Li., Y. H. Zi., X. W. Mu., M. J. Fu., T. T. Mo., K. Yu. Reliability and validity of common subjective instruments in assessing physical activity and sedentary behaviour in Chinese college students. International Journal of Environmental Research and Public Health, 19(14): 8379, 2022. https://doi.org/10.3390/ijerph19148379

[13] X. Liu., T. P. Singh., R. K. Gupta., E. M. Onyema. "Chaotic association feature extraction of big data clustering based on the internet of things". Informatica-an International Journal of Computing and Informatics. 46(3): 333-342, 2022. https://doi.org/10.31449/inf.v46i3.3943

[14] H. Sun., R. Chen., Y. Qin., S. Wang. "Holo-Entropy based categorical data hierarchical clustering". Informatica. 2017; 28(2): 303-328, 2017. https://doi.org/10.15388/Informatica.2017.131

[15] Z. Zheng., F, Cao., S, Gao., A, Sharma. "Intelligent analysis and processing technology of big data based on clustering algorithm". Informatica-an International Journal of Computing and Informatics. 46(3): 393-402, 2022. https://doi.org/10.31449/inf.v46i3.4016

[16] A. R. Al-Haifi., B. A. Al-Awadhi., N. Y. Bumaryoum., F. A. Alajmi., R.H. Ashkanani., H.M. Al-Hazzaa. The association between academic performance indicators and lifestyle behaviors among Kuwaiti college students. Journal of Health Population and Nutrition, 42(1): 27, 2023. https://doi.org/10.1186/s41043-023-00370-w

[17] Z. H. Zhang., J. S. Wang., L. Chen. An edge detection method of colony image based on mediocrity ant colony algorithm. Journal of Intelligent & Fuzzy Systems, 46(1): 2665-2691, 2024. https://doi.org/10.3233/JIFS-23376

[18] Y. Q. Shi., M. Y. Shi., C. Liu., L. Sui., Y. Zhao., X. Fan. Associations with physical activity, sedentary behavior, and premenstrual syndrome among Chinese female college students. BMC Womens Health, 23(1): 173, 2023. https://doi.org/10.1186/s12905-023-02262-x

[19] W. X. Tong, B. Li., S. S. Han., Y. H. Han., S. Q. Meng., Q. Guo., et al. Current status and correlation of physical activity and tendency to problematic mobile phone use in college students. International Journal of Environmental Research and Public Health, 19(23): 15849, 2022. https://doi.org/10.3390/ijerph192315849

[20] F. Wan., Y. Liu. Clustering mining algorithm of internet of things database based on python language. Computing and Informatics, 42(5): 1136-1157, 2023. https://doi.org/10.31577/cai_2023_5_1136

[21] B. Li., S. S. Han., S. Q. Meng., J. Lee., J. Cheng., Y. Liu. Promoting exercise behavior and cardiorespiratory fitness among college students based on the motivation theory. BMC Public Health, 22(1): 738, 2022. https://doi.org/10.1186/s12889-022-13159-z

[22] S. Y. Peng., F. Yuan., A. T. Othman., X. G. Zhou., G. Shen., J. H. Liang. The effectiveness of e-health interventions promoting physical activity and reducing sedentary behavior in college students: A systematic review and meta-analysis of randomized controlled trials. International Journal of

Environmental Research and Public Health, 20(1): 318, 2023. https://doi.org/10.3390/ijerph20010318

[23] M. M. Guo., X. Z. Wang., K. T. Koh. Association between physical activity, sedentary time, and physical fitness of female college students in China. BMC Womens Health, 22(1): 502, 2022. https://doi.org/10.1186/s12905-022-02108-y

[24] W. H. Zhang., C. Y. Wang., W. J. Lin., J. M. Lin. Continuous-domain ant colony optimization algorithm based on reinforcement learning. International Journal of Wavelets Multiresolution and Information Processing. 19(3): 2050084, 2021. https://doi.org/10.1142/S0219691320500848

[25] F. Moslehi., A. Haeri. A novel feature selection approach based on clustering algorithm. Journal of Statistical Computation and Simulation, 2021; 91(3): 581-604, 2021. https://doi.org/10.1080/00949655.2020.1822358

[26] J. B. Zhao., X. M. You., Q. Q. Duan., S. Liu. Multiple ant colony algorithm combining community relationship network. Arabian Journal for Science and Engineering, 47(8): 10531-10546, 2022. https://doi.org/10.1007/s13369-022-06579-x

[27] S. L. Xu., L. Feng., S. L. Liu., J. Zhou., H. Qiao. Multi-feature weighting neighborhood density clustering. Neural Computing & Applications, 32(13): 9545-9565, 2020. https://doi.org/10.1007/s00521-019-04467-4

[28] L. Luo., N.Q. Song., J. Huang., X.D. Zou., J.F. Yuan., C.L. Li., et al. Validity evaluation of the college student physical literacy questionnaire. Frontiers in Public Health, 10): 856659, 2022. https://doi.org/10.3389/fpubh.2022.856659

[29] J. Yu., X.M. You., S. Liu. Ant colony algorithm based on magnetic neighborhood and filtering recommendation. Soft Computing, 2021; 25(13): 8035-50, 2021. https://doi.org/10.1007/s00500-021-05851-w