International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1858-1871

https://doi.org/10.70517/ijhsa464161

Research on the Construction of Psychological Profiles of College Students and the Practice of Precise Psychological Education Supported by Cluster Analysis Technology

Lingling Zhang[1,*]
[1] Mental Health Center, Yangzhou Polytechnic Institute, Yangzhou, Jiangsu, 225127, China
Corresponding authors: (e-mail: lingzhang20252025@163.com).

Abstract Currently, mental health issues among college students have become a complex and pressing reality, necessitating a systematic and scientific intervention strategy for addressing mental health crises among this population. This study first explored a method for obtaining optimal solutions by utilizing an improved artificial bee colony algorithm to derive initial cluster centers, followed by the application of the ABC-SC algorithm in the optimization process of fuzzy clustering algorithms. Subsequently, based on the characteristics of mental health data among college students, a user profile suitable for mental health questionnaire data was established. Finally, a 12-week experimental intervention was conducted on college students from a certain university to experimentally test the effectiveness of the proposed method in improving college students' mental health status. By using the Symptom Checklist-90 (SCL-90) to assess college students' mental health, it was found that the proposed method significantly improved symptoms of depression, anxiety, hostility, phobia, somatization, obsessive-compulsive disorder, paranoia, and interpersonal relationships.

Index Terms FCM algorithm, cluster analysis, artificial bee colony algorithm, college student psychological profile

I. Introduction

With the development of society and increasing attention to self-emotional well-being, mental health issues have become increasingly important [1], [2]. Among college students, mental health issues have become more prominent due to factors such as increased academic and life pressures, changes in family dynamics, and interpersonal relationships [3], [4]. In this context, implementing targeted mental health education practices for college students holds significant importance for enhancing their mental well-being [5].

In recent years, cluster analysis techniques have been widely applied in the field of psychology. By conducting cluster analysis on large datasets, mental health status can be determined, providing effective evidence for psychological intervention [6], [7]. For analyzing college students' mental health, cluster analysis techniques can utilize data such as students' personal information and campus mental health assessment forms [8]-[10]. Next, clustering is performed based on data similarity, grouping similar data together [11]. In calculating similarity, various strategies can be employed, such as Euclidean distance, cosine similarity, inner product, and Manhattan distance [12], [13]. Each data sample has a set of attributes, and the presence or absence of these attributes can indicate whether the sample belongs to a particular cluster [14], [15]. In this way, all data samples can be divided into several groups, with each group having high internal similarity and low inter-group similarity [16], [17]. By confirming students' psychological profiles through clustering analysis, targeted, precise, and personalized mental health education can be implemented [18]-[20].

This paper first utilizes global statistical information obtained through distribution estimation algorithms to improve the method of nectar source generation in the artificial bee colony algorithm, thereby proposing an artificial bee colony algorithm based on continuous distribution estimation to effectively enhance the global exploration capability of the artificial bee colony algorithm. Subsequently, adjustments were made to the fuzzy parameters in the FCM algorithm, increasing the algorithm's iteration paths to avoid getting stuck in local optima, and the artificial bee colony algorithm was applied to the fuzzy clustering algorithm to optimize the initial clustering. The performance of the improved algorithm is validated through experiments. Furthermore, based on the basic principles and framework of user profiling, questionnaire data on workers' perceptions of mental health are used to construct a psychological profile of college students, and the improved clustering algorithm is applied to create precise profiles for each category of individuals. Finally, after a 12-week experimental intervention on the experimental group of college students, the Symptom Checklist-90 (SCL-90) was used to compare and analyze the factors influencing the mental

health of the experimental and control groups. This explored the effectiveness of the proposed method in improving the mental health status of college students.

II. Clustering algorithm based on hybrid artificial bee colonies

II. A. Artificial Bee Colony Algorithm and Its Improvements

II. A. 1) Basic Artificial Bee Colony Algorithm

In the Artificial Bee Colony (ABC) algorithm, the model primarily consists of three fundamental components: nectar sources, employed bees, and unemployed bees [21]. This model defines three basic behavioral patterns for bees: when a forager bee discovers a rich nectar source, it recruits other bees to follow it to this location for foraging. When the nectar supply at a source becomes scarce, the bees abandon that source, and the forager bees associated with it transition into scout bees. Scout bees randomly search for a new nectar source, begin foraging there, and then transition into forager bees to carry out their tasks. The implementation process of the ABC algorithm is as follows:

1) Initialization phase. Initialize the maximum number of iterations MCN, the parameter limit for discarding nectar sources, and randomly generate SN initial solutions $X=\{x_1,x_2,\cdots,x_i,\cdots,x_{SN}\}$ (where SN is the number of follower bees or leader bees), where $x_i=(x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{id})$ is a d-dimensional vector, and initialize the position of nectar source i according to equation (1):

$$x_{ij}=x_{ij}^{\min}+rand(0,1)\left(x_{ij}^{\max}-x_{ij}^{\min}\right) \tag{1}$$

where, $x_{ij}^{\max}$ and $x_{ij}^{\min}$ represent the upper and lower limits of x_{ij}, respectively.

The position of each nectar source corresponds to a feasible solution to an optimization problem, and its "yield" corresponds to the fitness value $fit(x_i)$ of the feasible solution, which can be calculated according to equation (2):

$$fit(x_i)=\frac{1}{1+f(x_i)} \tag{2}$$

In the equation: $f(x_i)$ is the objective function value of the artificial bee colony algorithm.

2) Leader bee behavior phase. In this phase, the leader bee uses the "greedy selection" rule to compare the fitness values of the old nectar source and the new nectar source obtained through neighborhood search. When the new nectar source is better than the old one, it replaces the old one; otherwise, it remains unchanged. During neighborhood search, the new nectar source v_i is generated from the old nectar source x_i according to the following equation.

$$v_{ij}=\begin{cases}x_{ij}+\varphi_{ij}(x_{ij}-x_{kj}) & if j=j'\\ x_{ij} & otherwise\end{cases} \tag{3}$$

In the formula: $k\in\{1,2,\cdots,SN\}$ and $k\neq i$, i.e., the selected nectar source x_k is different from x_i. j' is a randomly selected dimension, $j'=1,2,\cdots,d$. $\varphi_{ij}\in[-1,1]$, which is a random number used to control the scope of the neighborhood search.

3) Follower bee behavior phase. When the leader bees return to the hive after completing their task, they communicate the nectar source information to the follower bees by dancing. The follower bees select the nectar source they will mine based on the nectar source information obtained through a roulette-style selection process. The higher the yield of a nectar source, the more follower bees it attracts. In the ABC algorithm, follower bees decide on nectar sources based on the probability value $P(x_i)$, so the probability $P(x_i)$ of follower bees selecting nectar source x_i is:

$$P(x_i)=\frac{fit(x_i)}{\sum_{i=1}^{SN}fit(x_i)} \tag{4}$$

When a follower bee selects a nectar source for foraging, it also performs a neighborhood search based on Equation (3) and applies the greedy principle for selection.

4) Scout bee behavior phase. When the yield of a nectar source has not been updated for limit consecutive times, it indicates that the nectar source has reached a local optimum, at which point it should be abandoned. At the same time, the leader bee transforms into a scout bee and generates a new nectar source randomly according to equation (1).

II. A. 2) Artificial Bee Colony Algorithm Based on Continuous Distribution Estimation

1) Algorithm Principles

The distribution estimation algorithm is a novel evolutionary algorithm based on probabilistic models. It combines genetic algorithms with statistical learning to guide the generation of the next generation of the population based on the current population's probabilistic model. This algorithm models the “macro” level of biological evolution and possesses excellent global exploration capabilities [22].

The distribution estimation algorithm can be categorized into two types based on encoding methods: discrete encoding and continuous encoding. To enhance the global search capability of the artificial bee colony (ABC) algorithm, this paper combines the ABC algorithm with the continuous distribution estimation algorithm UMDAc to propose a continuous distribution estimation-based ABC algorithm. This is achieved by improving the new nectar source generation formula of the ABC algorithm, utilizing the probability distribution of high-quality nectar sources to guide the generation of new nectar sources, thereby significantly improving the algorithm's convergence speed and robustness.

Among them, the process of UMDAc is as follows: each generation randomly generates M individuals from the probability vector $P(x)$, calculates the fitness value of these individuals, and then selects N optimal individuals from them, and updates the probability vector $P(x)(N \le M)$ with these N individuals. where $P_q(x)$ is used to represent the probability vector of the first q generation, that is, $P_q(x) = (p_q(x_1), p_q(x_2), \cdots, p_q(x_d))$, $x_q^1, x_q^2, \cdots, x_q^N$ represents the optimal N individuals selected, where $x_q^i = (x_q^{i1}, x_q^{i2}, \cdots, x_q^{id})$, assuming that $p(x_i)$ obeys a Gaussian distribution, $x_i \sim N(\mu_i, \sigma_i^2)$, the mean μ_i and the variance σ_i^2 are estimated as follows:

$$\hat{\mu}_i = \frac{1}{N}\sum_{k=1}^{N} x_q^{ki} \tag{5}$$

$$\hat{\sigma}_i^2 = \frac{1}{N}\sum_{k=1}^{N} (x_{ki} - \hat{\mu}_i)^2 \tag{6}$$

From equations (5) and (6) above, we obtain the mean vector $\overline{\mu} = (\mu_1, \mu_2, \cdots, \mu_d)$ and the covariance matrix $diag(\sigma_1, \sigma_2, \cdots, \sigma_d)$, thereby obtaining a multivariate independent normal distribution $N(\overline{\mu}, \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_d))$. Sampling is performed based on this normal distribution to generate a new generation of the population. The sampling method is as follows: first, generate two sets of random numbers $q_{ij}, p_{ij}\ (i = 1, 2, \cdots, SN; j = 1, 2, \cdots, d)$ that are uniformly distributed on [0, 1], then perform the following transformation: $w_{ij} = (-2\ln q_{ij})^{\frac{1}{2}} \cos(2\pi p_{ij})$; $i = 1, 2, \cdots, SN$ and $j = 1, 2, \cdots, d$. It can be seen that $w_{ij}\ (i = 1, 2, \cdots, SN; j = 1, 2, \cdots, d)$ follows a normal distribution $N(0,1)$. Further transformation yields $x_{ij'} = w_{ij}\sqrt{\sigma_j} + \mu_j$, $i = 1, 2, \cdots, SN$ and $j = 1, 2, \cdots, d$.

Therefore, the formula for updating the nectar source during neighborhood search by the leader bee becomes equation (7):

$$v_{ij} = x_{ij} + \varphi_{ij}\left(x'_{ij} - x_{kj}\right) \tag{7}$$

2) Specific implementation steps

The specific implementation steps of the UMDAc-ABC algorithm are as follows.

(1) Initialization. Initialize the three variables SN, MCN, and limit. Use equation (1) to initialize the nectar sources $X = \{x_1, x_2, \cdots, x_i, \cdots, x_{SN}\}$ using Equation (1), which are randomly generated according to a uniform distribution in the solution space. Simultaneously, calculate the fitness values corresponding to each nectar source using Equation (2).

(2) Execute the following loop:

a) Select the top $SN/2$ individuals with higher fitness values, and calculate their mean and variance: $\overline{\mu} = (\mu_1, \mu_2, \cdots, \mu_d)$ and $diag(\sigma_1, \sigma_2, \cdots, \sigma_d)$.

b) Leader bee stage: From the mean and variance calculated above, a multivariate normal distribution $N(\overline{\mu}, diag(\sigma_1, \sigma_2, \cdots, \sigma_d))$ using the improved neighborhood search method described above, i.e., generate a new nectar source v_i according to equation (7) and use the greedy principle to determine whether to replace the old nectar source x_i.

c) Similar to step a), re-select $SN/2$ high-quality nectar sources and calculate their mean and variance.

d) Follow-up bee phase: Follow-up bees calculate the probability of each nectar source being selected using equation (4) based on the fitness values obtained in (1), and select the nectar source they will mine using a roulette wheel method.

e) Repeat step c).

f) Scout bee phase: Determine whether any nectar sources have reached a local optimum. If so, abandon that nectar source, and the corresponding leader bee is converted into a scout bee. Use equation (1) to begin searching for the next new nectar source and collect nectar, after which the scout bee is converted back into a leader bee.

g) Record the current optimal nectar source, check whether the termination condition is met, and if so, output the optimal nectar source (solution). Otherwise, continue the loop.

II. B.FCM algorithm optimized based on improved artificial bee colony algorithm

II. B. 1) Description of the improved artificial bee colony algorithm for optimizing the GFCM algorithm

The GFCM algorithm makes the following improvements to the main formula in the fuzzy C mean clustering algorithm:

1) The fuzzy parameter in the objective function is m_1 and $m_1 > 0$. The formula is:

$$J_{GFCM}(U,v) = \sum_{i=1}^{n}\sum_{k=1}^{c}(\mu_{ik})^{m_1}(d_{ik})^2 \tag{8}$$

2) The fuzzy parameter in the cluster center calculation is m_2 and $m_2 > 0$. The formula is:

$$v_j = \frac{\sum_{i=1}^{n}(\mu_{ik})^{m_2}x_i}{\sum_{i=1}^{n}(\mu_{ik})^{m_2}} \tag{9}$$

3) The fuzzy parameter in the membership degree calculation is m_3 and $m_3 > 1$. The formula is:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_i - v_k\|^2}{\|x_i - v_j\|^2}\right)^{\frac{2}{m_3-1}}} \tag{10}$$

The algorithm flow is as follows:

Step 1: Initialize the control parameters, the number of clusters $c\left(2 \le c \le n\right)$, the number of samples n, the fuzzy parameters m_1, m_2, m_3, the stopping threshold ε, and the iteration counter to 0.

Step 2: Use formula (10) to calculate or update the membership matrix U such that $\sum_{i=1}^{c}\mu_{ij} = 1$.

Step 3: Use formula (9) to calculate the cluster centers V and calculate the distance d_{ij} between each data sample x_i and the cluster center c_j according to formulas (2-11).

Step 4: Use formula (8) to calculate the objective function J_{GFCM} and determine whether its change is less than ε or the algorithm has reached the specified number of iterations. If so, the algorithm stops; otherwise, proceed to step 2.

II. B. 2) Clustering Approach Based on the ABC-SC Algorithm

The quality of nectar sources is equivalent to the quality of potential solutions, the speed of nectar collection represents the speed of solution, and the optimal nectar source indicates the best clustering result. Each honey source represents a set of cluster centers, and the data object set $X = \{x_1, x_2 \cdots x_n\}$ is defined, where the variable x_i is a d-dimensional vector, and the cluster centers form the set $V = \{v_1, v_2 \cdots v_c\}$ where v_k is also a d-dimensional vector. Thus, the position of each artificial bee is a $c*d$-dimensional vector. Currently, most clustering algorithms use floating-point number encoding, so this paper also adopts this encoding method. Then, each artificial bee can be represented in the following form:

$$x_i = \{v_{11}, v_{12} \cdots v_{ij} \cdots v_{c1} \cdots v_{cd}\} \tag{11}$$

The fitness function measures the fitness of potential solutions to clustering problems. In the FCM algorithm, the higher the fitness of an individual, the smaller the objective function, and the better the clustering effect. Therefore, the fitness function defined in this paper is as follows:

$$fit_i = \frac{1}{1 + J_{GFCM}(U,V)} \tag{12}$$

Among these, $J_{GFCM}(U,V)$ is the objective function. The smaller its value, the larger the fitness value fit_i, and the better the clustering effect.

The idea behind the ABC-SC-based GFCM (abbreviated as ABC-SGFCCM) algorithm is as follows: First, the ABC-SC algorithm is used to optimize the initial cluster centers. The optimal solution output by the ABC-SC algorithm is used as the best cluster center. Then, the GFCM algorithm is used to optimize this cluster center to obtain the clustering result.

II. B. 3) Algorithm Flow

Step 1 Initialization of relevant control parameters, including population size NP, number of nectar sources (i.e., initial solution count) SN, iteration control parameter limit, maximum evolution count MaxCycle, number of clusters cn, error parameter ε, etc.

Step 2 Initialize the bee swarm by randomly sampling SN data samples from the dataset, which serve as the initial cluster centers. Generate the initial positions of the nectar sources based on the encoding method. Calculate the fitness of each nectar source using Equation (12) and sort them by value. Select the top half as leading bees and the remaining as following bees. Initialize the membership matrix U.

Step 3 The leader bee searches for new nectar sources in its neighborhood according to the formula and performs boundary handling on the new nectar sources. Calculate the fitness of the new nectar sources. If it is better than the original nectar sources, correct the original nectar sources and set the trial counter to 0. Otherwise, increment the counter by 1.

Step 4 The follower bees evaluate the quality of the nectar sources, calculate the following probability according to the formula, and select which leader bee to follow. Then, the follower bee searches for new nectar sources around the nectar source according to the formula, compares the fitness values, and retains the better nectar source. It also records the optimal solution for each generation of the population.

Step 5 Determine whether the counter has reached the threshold limit. If so, discard the nectar source, and the corresponding leader bee changes its role to a scout bee and randomly generates a new nectar source according to the formula.

Step 6 If the optimal value of the population does not change after three consecutive iterations, it indicates that the population has reached a local optimum. Apply chaotic mutation to each nectar source according to the formula. If the fitness of the new nectar source is better than the original one, replace the original nectar source with the mutated new one. Record the global optimal solution after chaotic mutation and compare it with the global optimal solution before mutation. If it is better, replace it.

Step 7 Determine whether the error has reached the set threshold. If so, obtain the optimal cluster center and output it; otherwise, go to Step 3.

Step 8 Update the membership matrix according to formula (10) to generate a new membership matrix U.

Step 9 Update the cluster centers according to formula (9). If $\left|U^{t+1}-U^{t}\right|<\varepsilon$, stop; otherwise, go to step 8.

The ABC-SGFCM algorithm flow is shown in Figure 1.

II. C.Simulation and analysis based on improved clustering algorithms

II. C. 1) Determination of Fuzziness Parameters

This section conducts experiments on the proposed algorithm using MATLAB. First, simulation experiments are performed on the classic test datasets Iris, Seeds, and Glass from the UCI database to evaluate the clustering performance of FCM, KFCM, and the improved clustering algorithm proposed in this paper. The maximum number of iterations is set to 100, the population size is 20, and the Gaussian kernel parameter is 150. When determining the fuzzy parameter m, its value must be greater than 1. However, m should not be too large, as values too large cause membership degrees to approach 1/c, where c is the number of clusters. The membership degree change plots are shown in Figure 2 (Figures (a), (c), and (e) represent the membership degrees of the three datasets when m = 15). Figures (b), (d), and (f) show the membership degree plots of the three datasets using the algorithm in this paper at their optimal m values. It can be seen that when m is too large, the membership degree partitioning becomes unclear. Therefore, the value range of the FCM algorithm should be (1, +), and when m > 6, the membership degrees tend to be equal. This paper determines the range of m values as [1.1, 6], and it can be seen that the membership degree partitioning is clearer at the optimal m value.

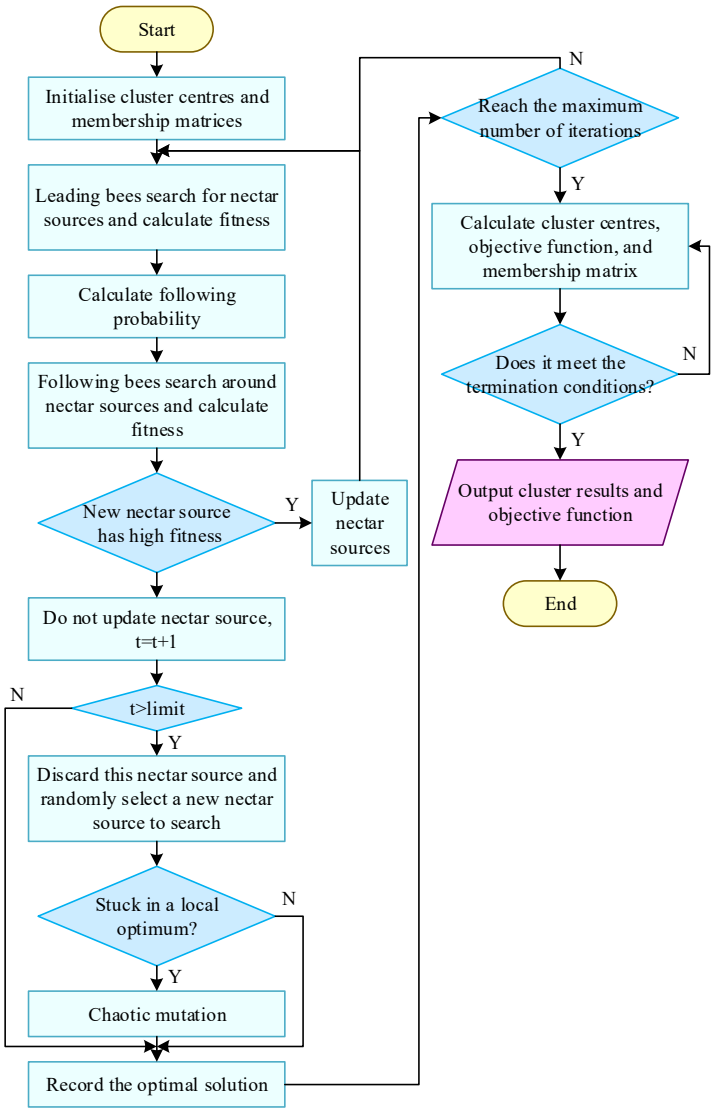
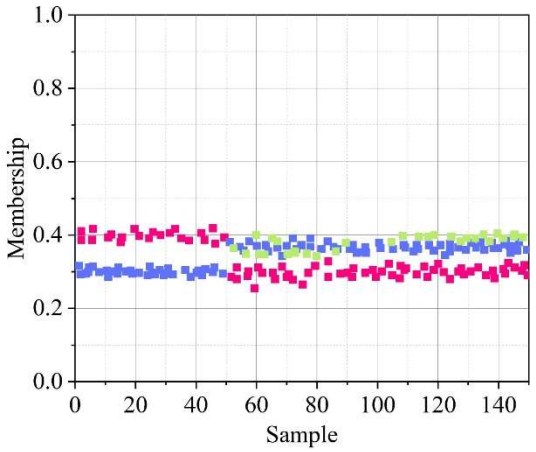
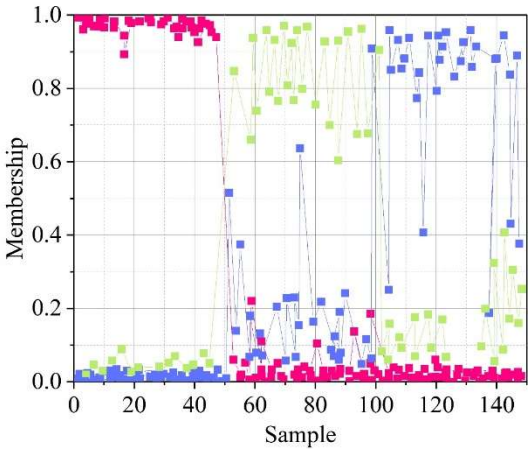


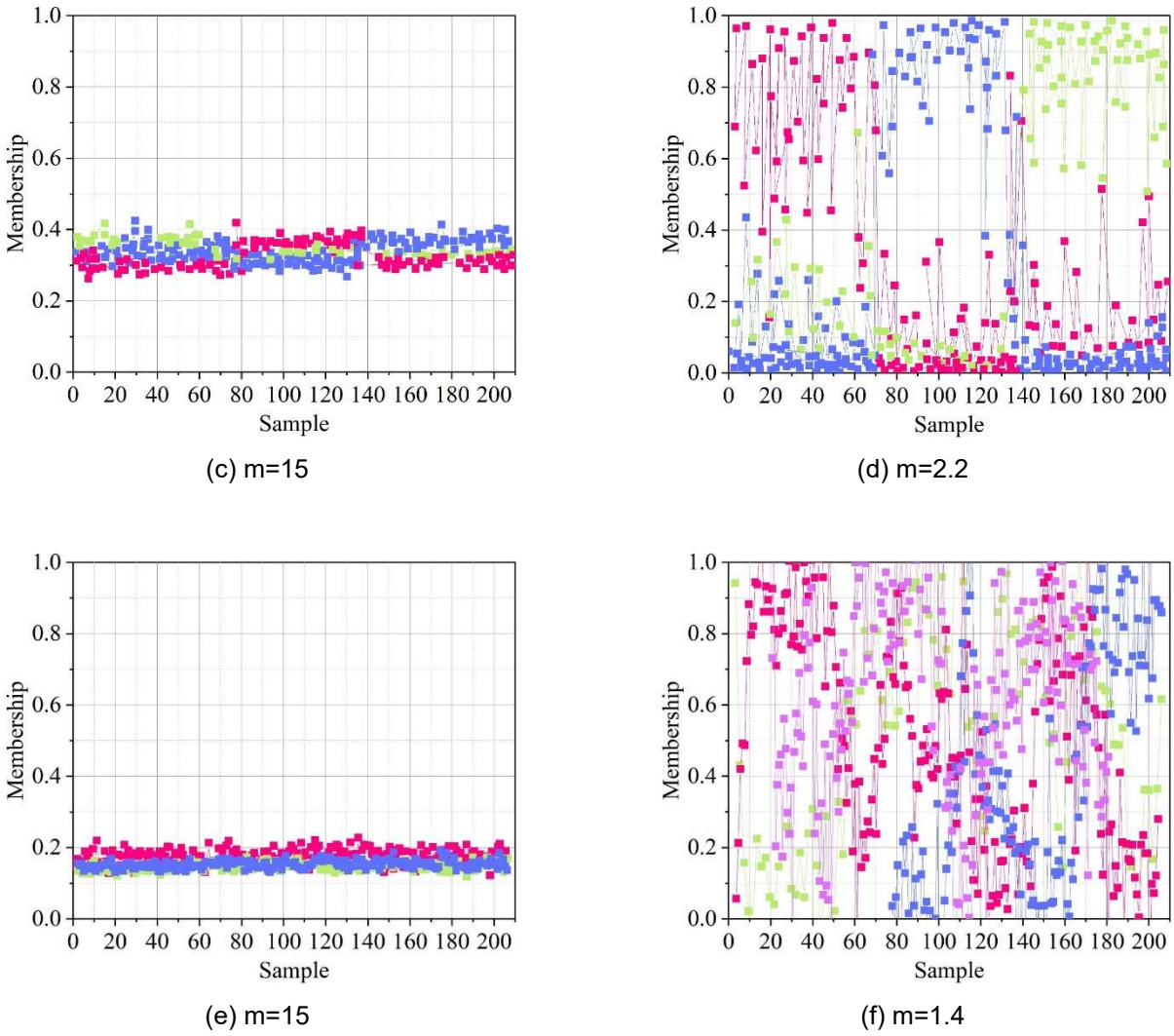
Figure 1: The process of ABC-SGFCM algorithm



(a) m=15



(b) m=2.2



(c) m=15 (d) m=2.2

(e) m=15 (f) m=1.4

Figure 2: Membership figure

FCM, the comparison of MIA values under different m values of the improved algorithm in this paper is shown in Table 1. From the table, we can see that the MIA values under the algorithm in this paper are smaller than those under the FCM algorithm. This indicates that the algorithm in this paper performs better than the FCM algorithm.

Table 1: The MIA value of FCM algorithm and our algorithm with different m value

m value	Iris database		Seeds database		Glass database	
	FCM	Ours	FCM	Ours	FCM	Ours
1.2	4.8596	4.8065	14.1108	14.0267	11.8297	9.4049
1.4	4.8318	4.8219	14.0573	13.9462	10.6221	6.0195
1.6	4.805	4.8737	14.0032	13.9734	11.9402	7.0119
1.8	4.8483	4.7231	13.93	13.9685	11.98	8.9079
2.0	4.7559	4.9604	13.9898	13.7423	11.9207	9.4095
2.2	4.9753	4.7581	13.8581	13.8373	12.4408	7.6458
2.4	4.8883	4.9134	13.9114	14.0465	12.5639	6.3151
2.6	4.9641	4.9002	14.1093	13.7444	12.5125	9.8743
2.8	4.9819	4.9812	14.1168	14.0236	12.9046	8.744
3.0	4.9103	4.8763	14.1714	14.0785	13.1123	9.4629
3.2	5.0012	4.8578	14.1307	13.9338	13.1911	7.8172

3.4	4.7993	4.885	14.0287	14.0882	13.3287	11.431
3.6	4.9215	4.7628	14.0802	14.1092	13.5941	13.0965
3.8	5.0667	4.7646	14.1074	14.0783	13.5916	8.8189
4.0	5.0075	4.9657	14.0045	14.1389	13.7315	11.4366
4.2	4.9244	5.0206	14.0705	14.0367	13.8892	10.6091
4.4	4.9898	4.9692	14.0432	14.1332	14.0729	10.8602
4.6	4.9221	5.0121	14.1404	14.1691	13.8238	12.2262
4.8	4.996	4.7856	14.2054	14.102	14.0389	8.5309
5.0	5.1449	5.0291	14.1582	14.1966	14.374	9.6067
5.2	4.9536	5.145	14.037	14.0858	14.2305	9.1513
5.4	5.0738	5.0592	14.0326	14.0411	14.4689	12.1593
5.6	5.0517	4.9161	14.3115	14.217	16.4033	10.6767
5.8	5.0422	5.0132	14.2672	14.1602	15.8568	11.3717
6.0	5.0233	5.1861	14.1057	14.2292	15.2053	13.7851

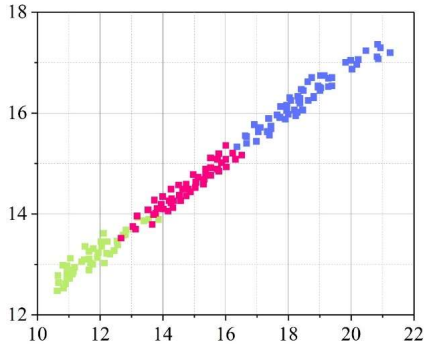
II. C. 2) Clustering accuracy verification

In this section, we evaluate the clustering performance of three clustering algorithms based on the clustering accuracy and clustering effectiveness evaluation metric XB. The XB metric is a ratio-type fuzzy clustering effectiveness metric, where intra-cluster compactness is represented by the sum of the distances between each sample and the cluster center, and inter-cluster separability is measured by the minimum distance between all cluster centers. The XB metric is defined as: under the same number of clusters, the smaller the XB metric value, the better the clustering effect. The evaluation of clustering effects is shown in Table 2.

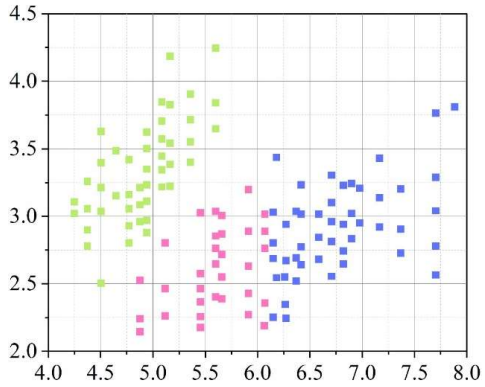
Table 2: Clustering effect evaluation

	FCM		KFCM		Ours	
	XB value	Correct rate (%)	XB value	Correct rate (%)	XB value	Correct rate (%)
Iris database	13.98	91.66	12.32	91.81	4.39	94.93
Seeds database	9.89	89.37	5.85	90.81	5.83	93.41
Glass database	25.82	75.86	23.59	79.92	19.37	83.82

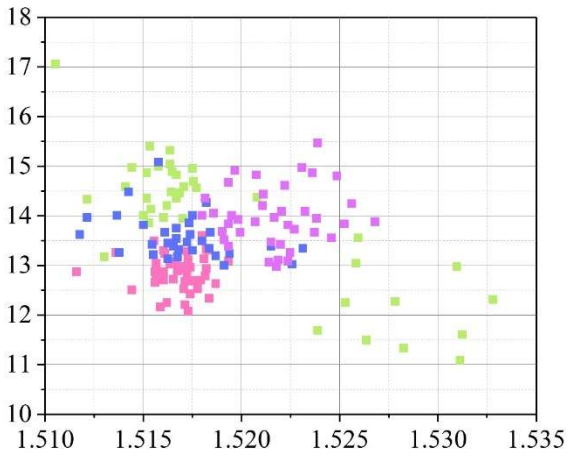
The clustering results are shown in Figure 3 (Figures a–c show the clustering results for the classic test datasets Seeds, Iris, and Glass, respectively). From the figure, we can clearly see that the XB metric of the algorithm proposed in this paper is lower than that of the FCM and KFCM algorithms in the three different datasets, and the accuracy has also improved. The clustering results obtained using the algorithm proposed in this paper are satisfactory. Additionally, the FCM algorithm is a hill-climbing algorithm that is highly sensitive to initial values and prone to getting stuck in local minima. The selection of its fuzzy parameters also directly affects the performance of the FCM algorithm. The XB metric further indicates that the algorithm proposed in this paper can efficiently perform effective clustering on the data.



(a) Seeds database



(b) Iris database



(c) Glass database

Figure 3: Clustering graph

III. Building a psychological profile of college students

III. A. Basic Process of User Profiling

III. A. 1) Overview of the User Profile Creation Process

Based on current research and practical applications, the basic process of user profiling is as follows: data collection, data description, data classification and label extraction, and data modeling. The specific process is as follows:

(1) Determine the basic direction of user profiling. Determine the purpose of the research and clarify the type of data required. This paper aims to study the cognitive attitudes toward mental health among working professionals, so the data type is questionnaire survey data on mental health.

(2) Data collection. Based on the research objectives, various data collection methods can be employed, such as surveys, observations, experiments, and online data collection. This study obtained questionnaire survey data through an online platform.

(3) Data preprocessing, primarily involving data cleaning, including verification, screening, filtering, handling of missing and outlier values, and data transformation.

(4) User tag modeling. This involves data description, data classification, extraction of category tags, feature summarization for various population groups, and user profiling.

(5) Through association analysis or classification regression, a profiling model is established for recommendation, prediction, or warning purposes. User profiles are formed, and unknown information is classified and predicted based on user feature tag data.

III. A. 2) Data Collection and Organization

(1) Questionnaire Data Validation

To ensure that the questionnaire has high validity and reliability, validity and reliability analysis must be conducted. The questionnaire is validated for validity and reliability.

Reliability analysis refers to the reliability of the collected questionnaires. It refers to the degree of consistency of the results obtained by having the same respondents repeatedly fill out the same questionnaire using the same survey methods. The α reliability coefficient method is commonly used to measure the reliability of a questionnaire, and its formula is:

$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum \sigma_i^2}{\sigma^2}\right) \tag{13}$$

where k is the number of questions in the questionnaire, σ_i^2 is the variance of the survey results for question i, and σ^2 is the variance of all survey results. It is generally considered that reliability coefficients between 0.65 and 0.7 are the minimum acceptable range, while those between 0.8 and 0.9 are very good.

(2) Outlier and Missing Data Handling

Missing values refer to data that could not be obtained during data collection due to various reasons. Missing types generally include completely random missing, random missing, and non-random missing. Based on the

missing mechanism of the data, the main methods for handling missing values are direct exclusion and data interpolation.

(3) Data conversion

Raw data often has inconsistent scales between different indicators due to their respective units of measurement, making them incomparable. If the raw data is used without any processing, it will affect the analysis results. Therefore, it is necessary to standardize the data. For qualitative data, since most algorithms use distances in vector spaces for correlation calculations, it is necessary to convert qualitative data so that discrete feature values can be calculated in Euclidean space. A common method is to convert them into dummy variables, with the basic idea being to encode qualitative variables into binary form. When converting qualitative variables, dummy encoding and one-hot encoding can be used.

Virtual coding refers to the process of generating $m-1$ binary codes from m categorical values of a qualitative variable. For example, under the feature of region (D), there are three options (central, eastern, and western regions). Only two virtual variables need to be set, and the coding format is D_1D_2. An example is shown below:

$$D_1=\begin{cases}1,\text{Central}\\0,\text{Non-central}\end{cases},D_2=\begin{cases}1,\text{Eastern}\\0,\text{Non-eastern}\end{cases} \tag{14}$$

When $D_1=1$ and $D_2=0$, it represents the central region, i.e., 10 represents the central region. When $D_1=0$ and $D_2=1$, it represents the eastern region, i.e., 01 represents the eastern region. When $D_1=0$ and $D_2=0$, it represents the western region, i.e., 00 represents the western region.

III. A. 3) Data Description and Exploratory Analysis

(1) Descriptive statistics calculation

Descriptive statistics include: the mean, which describes the average trend of numerical data, and the mode, which describes the frequency of categorical data. Measures of dispersion include quantiles, standard deviation, variance, and coefficient of variation. Descriptive statistics that describe the shape of the data distribution include kurtosis and skewness, both of which are based on the normal distribution as a reference standard.

(2) Analysis of the correlation between variables

The trend of change between variables can be represented by correlation coefficients, including Pearson's correlation coefficient, Spearman's correlation coefficient, and Kendall's correlation coefficient. Different types of variables typically require different correlation coefficients. Since the data type in this paper is qualitative data, Kendall's correlation coefficient is introduced.

The Kendall correlation coefficient is a rank correlation coefficient suitable for qualitative variables. Its value ranges from -1 to 1. When the Kendall correlation coefficient equals 1, it indicates a strong correlation between the two variables, representing a perfect positive correlation. When it equals -1, it indicates a perfect negative correlation between the two variables. When it equals 0, it means the two variables are independent of each other. There are three formulas, each applicable to different situations. Formula 1, denoted as Tau-a, is applicable when there are no common elements between the two variables, as follows:

$$Tau-a=\frac{A-B}{\frac{1}{2}N(N-1)} \tag{15}$$

where A represents the number of identical pairs of elements in the two variables. B represents the number of non-identical pairs of elements in the two variables. N is the total sample size.

(3) Data visualization

Using statistical tables and statistical graphs to present data has the advantage of being intuitive and vivid.

III. A. 4) Classification and Feature Label Extraction

When modeling user tags, clustering methods are typically used to classify samples based on their attitudes toward mental health. Clustering data is done to maximize the separation between classes, minimizing the similarity between classes and ensuring that samples with high similarity are grouped together. After classifying users, it is necessary to describe the category characteristics of the profiling model. Each actual category of people should be described, summarizing and generalizing their characteristics and analyzing the differences between different categories of people.

III. B. Analysis of college students' psychological profiles based on ABC-SGFCM clustering

III. B. 1) Correlation results

By integrating, analyzing, and modeling students' psychological and behavioral data, we can create a “precise profile” of their psychological state. The correlation matrix of college students' on-campus performance indicators is shown

in Table 3. The positive or negative values indicate positive or negative correlations, respectively, with larger absolute values indicating stronger correlations. We focus on the last row, which examines the correlation between 2024 psychological data and the other 13 categories of characteristics. We found that college students' mental health during their time at school is negatively correlated with seven indicators: first-class performance, second-class performance, participation in ideological and political, scientific and technological, and cultural and sports activities, employment status, and student leadership roles. That is, the more severe the mental health issues of the study subjects, the poorer their performance in first-class activities, the lower their participation in second-class activities, the lower their participation in ideological and political, scientific and technological, and cultural and sports activities, the poorer their employment status, and the less likely they are to hold student leadership roles. This finding aligns with the practical work of counselors, as some students with mental health issues, influenced by their emotions, are prone to developing academic aversion, unwillingness to participate in school activities, and reluctance to interact with others. At the same time, we found that psychological issues during school were positively correlated with five indicators: social practice, volunteer service, skill development, family economic situation, and psychological status at the time of enrollment (i.e., the 2020 psychological assessment). That is, students with more severe psychological issues in this research center performed better in social practice, volunteer service, and skill development (such as obtaining socially recognized driver's licenses or technical certifications). Meanwhile, students from economically disadvantaged families and those with more severe psychological issues at enrollment tend to have more severe psychological issues during their school years. In the actual work of counselors, we have found that some students with psychological issues, although unwilling to participate in on-campus activities, are willing to engage with off-campus matters, willing to step out of campus to participate in practical or volunteer activities, and expand their social skills. At the same time, students from economically disadvantaged families and those who showed psychological issues at enrollment require more focused attention.

Table 3: Correlation matrix of college students' performance indicators at school

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Second class(1)	1													
Thinking(2)	0.07	1												
Social practice(3)	0.07	0.02	1											
Academic technology(4)	0.97	0.01	0.02	1										
Stylistic development(5)	0.07	0.26	-0.11	-0.03	1									
Skill development(6)	0.2	0.15	-0.03	0.13	0.08	1								
Volunteer service(7)	0.03	0.09	0.01	0	0.2	0.11	1							
Employment implementation(8)	-0.06	0.2	-0.15	-0.07	0	-0.07	0.03	1						
First class(9)	0.27	0.2	-0.29	0.26	0.24	0.14	0.06	0.22	1					
Political appearance(10)	0.56	0.19	0.07	0.54	0.12	0.19	0.11	-0.03	0.27	1				
Family poverty(11)	0.09	0.07	0.01	0.08	0.15	0.04	0.05	0.01	0.13	0.03	1			
Student cadre(12)	0.24	0.38	0.05	0.18	0.18	0.15	0.11	-0.07	0.22	0.24	0.09	1		
2020 assessment(13)	-0.05	0.06	-0.02	-0.01	0.04	-0.05	0.15	0.04	0.09	-0.07	-0.07	-0.04	1	
2024 assessment(14)	-0.13	0.15	0.39	-0.13	-0.13	0.19	0.18	-0.1	-0.19	0.11	0.15	-0.08		1

III. B. 2) Clustering results

Precise profiling will play a crucial role in student affairs work. On the one hand, by conducting in-depth exploration, uncovering data and the relationships between data, and extracting key information, we can promptly obtain insights into students' psychological states and ideological conditions, thereby achieving dynamic management of psychological monitoring. On the other hand, by extracting effective data information and establishing a psychological crisis early warning model, we can make reasonable assessments and interpretations of students' psychological states, addressing the shortcomings of psychological assessment scales. For example, by analyzing academic performance and attendance records, we can understand students' attitudes toward learning and the pressures they face. By analyzing library borrowing and browsing information, we can uncover students' interests and hobbies. Through dormitory access control systems and campus card consumption data, we can understand students' daily routines and living habits. By analyzing WeChat, Weibo, and QQ Space, we can assess students' interpersonal relationships and emotional changes. After obtaining indicators of college students' performance on campus, we integrate, analyze, and model their psychological behavior data. Through multiple clustering

experiments, we found that clustering into three categories yields the best results. The clustering results are shown in Figure 4, where A, B, and C represent the three categories of groups generated by clustering. By examining these three groups and creating precise profiles of the individuals in each, we found that these three categories of students have distinct characteristics.

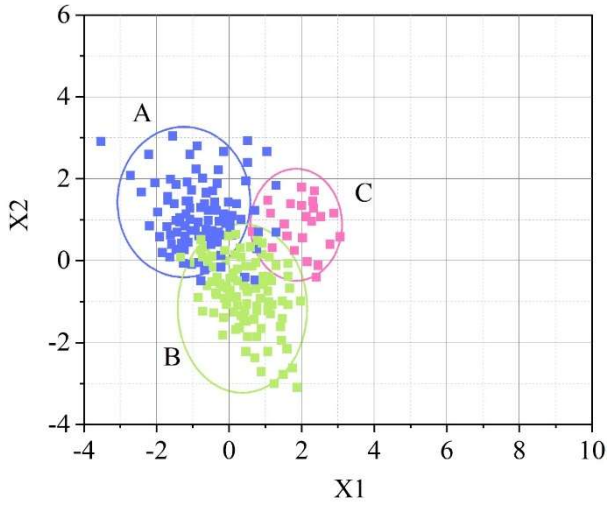


Figure 4: Clustering results

After clustering different groups, we found that Group A consists of the largest number of Party members, excels in both formal and informal classroom performance, and actively participates in science and technology innovation competitions. This group includes students who have been recommended for graduate school admission. No students with psychological abnormalities were identified in this group, making it a psychologically healthy group that does not require psychological crisis intervention. Group B includes a higher proportion of student leaders, who are more active in participating in ideological and political activities and have average academic performance. This group includes some students with psychological issues, who are classified as a group requiring ongoing monitoring. Group C has an advantage in social practice scores but also includes a higher proportion of students with psychological issues, making it a group requiring vigilance. In our work, we can identify students requiring vigilance based on psychological assessment results, but we cannot determine which groups may potentially transition into requiring vigilance. Therefore, through this study, we can identify non-monitored students in Group C who share common characteristics with monitored students based on psychological assessment results. These students are a group that may develop into monitored students and are the focus of this study. Through the research, we have preliminarily concluded that students with different levels of mental health exhibit clustering tendencies. We can use this study to identify groups of students who may develop into monitored students and implement targeted measures based on the characteristics of each group.

IV. Application of the methods described in this paper to mental health intervention among college students

IV. A. Research subjects and methods

IV. A. 1) Research subjects

A total of 200 second-year male and female college students from a certain university were selected as the experimental group, and another 200 male and female college students from the same grade were randomly selected as the control group.

IV. A. 2) Research Methods

Reviewing relevant research literature in psychology both domestically and internationally to determine the indicators for this experimental study. The indicators used in this study include: the Brief Mood State Scale (POMS), the Self-Rating Symptom Checklist (SCL-90), and the College Student Personality Health Survey, also known as the College Student Personality Questionnaire. Participants in both the experimental and control groups underwent psychological assessments using indicators such as body self-esteem, mood state, and mental health levels before the experiment began and after it concluded. Finally, statistical software SPSS 16.0 was used to statistically analyze and compare the experimental data.

IV. B. Analysis of the results of interventions in college students' mental health

The statistical changes in college students' mental health levels are shown in Table 4 (*P<0.05, **P<0.01). After 12 weeks of experimental intervention on the experimental group of college students, the results showed that the somatization factor, obsessive-compulsive symptom factor, and fear factor, after T-test, P<0.05, exhibited significant differences. The interpersonal sensitivity factor, depression factor, anxiety factor, and hostility factor showed highly significant differences after T-tests, with P < 0.01. The phobic factor showed significant differences after T-tests, with P < 0.05. The psychoticism factor did not show significant differences after T-tests, with P > 0.05. Thus, the 12-week mental health intervention for college students had a significant effect.

Table 4: Statistics on Changes in the Mental Health Level of College Students

Factor	Control group		Experimental group	
	Before the experiment	After the test	Before the experiment	After the test
Depression	1.54±0.4	1.56±0.44	1.24±0.25	1.35±0.26**
Anxiety	1.42±0.23	1.5±0.33	1.05±0.15	1.25±0.28**
Antagonism	1.52±0.56	1.46±0.57	1.13±0.16	1.37±0.26**
Horror	1.64±0.45	1.62±0.34	1.15±0.42	1.48±0.39*
Paranoia	1.59±0.47	1.56±0.51	0.99±0.39	1.38±0.43**
Somatization	1.56±0.43	1.51±0.46	1.15±0.17	1.4±0.4*
Compulsion	1.72±0.41	1.67±0.39	1.33±0.32	1.55±0.28*
Insanity	1.5±0.51	1.44±0.41	1.4±0.51	1.43±0.5
Interpersonal Relation	1.7±0.36	1.67±0.54	1.14±0.28	1.53±0.27**

V. Conclusion

In recent years, college students have faced increasing pressure in terms of academics, employment, relationships, and finances, which has seriously affected their mental health. Through an improved clustering analysis algorithm, this study constructed a mental health profile of college students and drew the following conclusions through precise psychological education practices:

After clustering different groups, three categories—A, B, and C—were identified. Group A consists of students with the highest number of party members, outstanding academic performance in both formal and informal education settings, and active participation in science and technology competitions. This group does not require psychological crisis intervention. Group B includes a higher proportion of student leaders and represents a group of students with psychological issues that require ongoing monitoring. Group C has an advantage in social practice scores but also includes a higher number of students with mental health issues, making it a group requiring vigilance.

After a 12-week experimental intervention on the experimental group of college students, the results showed significant differences between the experimental and control groups in somatization, obsessive-compulsive, and phobic factors (P < 0.05). Significant differences (P < 0.01) were observed in interpersonal sensitivity, depression, anxiety, hostility, and paranoia factors. No significant differences (P > 0.05) were observed in psychoticism factors. This indicates that the psychological health levels of the experimental group improved significantly after the 12-week intervention.

Funding

This work was supported by 2022 Annual Special Project Approval Topics of Philosophical and Social Sciences Research in Jiangsu Colleges and Universities (2022SJSZ1149).

References

[1] Hossain, M. M., Tasnim, S., Sultana, A., Faizah, F., Mazumder, H., Zou, L., ... & Ma, P. (2020). Epidemiology of mental health problems in COVID-19: a review. F1000Research, 9, 636.

[2] Fergusson, D. M., McLeod, G. F. H., Horwood, L. J., Swain, N. R., Chapple, S., & Poulton, R. (2015). Life satisfaction and mental health problems (18 to 35 years). Psychological medicine, 45(11), 2427-2436.

[3] Pedrelli, P., Nyer, M., Yeung, A., Zulauf, C., & Wilens, T. (2015). College students: mental health problems and treatment considerations. Academic psychiatry, 39(5), 503-511.

[4] Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden, J., & Gisle, L. (2017). Work organization and mental health problems in PhD students. Research policy, 46(4), 868-879.

[5] Bruffaerts, R., Mortier, P., Kiekens, G., Auerbach, R. P., Cuijpers, P., Demyttenaere, K., ... & Kessler, R. C. (2018). Mental health problems in college freshmen: Prevalence and academic functioning. Journal of affective disorders, 225, 97-103.

[6] Lattie, E. G., Adkins, E. C., Winquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. Journal of medical Internet research, 21(7), e12869.
[7] Eisenbarth, C. (2012). Coping profiles and psychological distress: A cluster analysis. North American Journal of Psychology, 14(3).
[8] Liu, F., Yang, D., Liu, Y., Zhang, Q., Chen, S., Li, W., ... & Wang, X. (2022). Use of latent profile analysis and k-means clustering to identify student anxiety profiles. BMC psychiatry, 22, 1-11.
[9] Lannoy, S., Mange, J., Leconte, P., Ritz, L., Gierski, F., Maurage, P., & Beaunieux, H. (2020). Distinct psychological profiles among college students with substance use: A cluster analytic approach. Addictive behaviors, 109, 106477.
[10] Lei, J. (2022). An analytical model of college students' mental health education based on the clustering algorithm. Mathematical Problems in Engineering, 2022(1), 1880214.
[11] Guo, R., Dong, R., Lu, N., Yu, L., Chen, C., Che, Y., ... & Yang, J. (2025). Physical Health Portrait and Intervention Strategy of College Students Based on Multivariate Cluster Analysis and Machine Learning. Applied Sciences, 15(9), 4940.
[12] Zhang, J. (2024, April). Research on Student Accurate Portrait Personalized Recommendation System Based on Improved KMeans Algorithm. In Proceedings of the International Conference on Algorithms, Software Engineering, and Network Security (pp. 428-434).
[13] Feng, Y., Zhao, T., Zhang, Y., Zu, Y., Tavares, A., Gomes, T., & Xu, H. (2024). Innovative Attempt at Enhancing Psychological Assessment: A Preliminary Investigative Study of Measuring College Students' Learning Motivation Levels through the Lens of Passive Sensing via Smartphones. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46).
[14] Wang, Y., & Teo, P. C. (2024). I know you: User profiling on social media usage of Chinese private university students. Jurnal Ilmiah Peuradeun, 12(1), 71-98.
[15] Sundhu, R., & Kittles, M. (2016). Precision teaching: does training by educational psychologist have an impact?. Educational Psychology in Practice, 32(1), 13-23.
[16] El Ansari, W., Ssewanyana, D., & Stock, C. (2018). Behavioral health risk profiles of undergraduate university students in England, Wales, and Northern Ireland: a cluster analysis. Frontiers in public health, 6, 120.
[17] Lin, Y. H. (2021). Clustering and Profile Analysis on Statistics Anxiety Styles of Undergraduate Students. International Journal of Intelligent Technologies & Applied Statistics, 14(4).
[18] Hart, P. D. (2017). Profiling physical fitness attributes in college students: A cluster analysis. International Journal of Physiology, Nutrition, and Physical Education, 2(2), 741-744.
[19] Scheidt, M., Godwin, A., Berger, E., Chen, J., Self, B. P., Widmann, J. M., & Gates, A. Q. (2021). Engineering students' noncognitive and affective factors: Group differences from cluster analysis. Journal of Engineering Education, 110(2), 343-370.
[20] Ng, M. Y., & Weisz, J. R. (2016). Annual research review: Building a science of personalized intervention for youth mental health. Journal of Child Psychology and Psychiatry, 57(3), 216-236.
[21] Toufik Arrif,Mawloud Guermoui,Abdelfetah Belaid & Badraddine Bezza. (2025). Pattern-free heliostat layout optimization using a modified Artificial Bee Colony Algorithms: Comparison between staggered and spiral layouts. Solar Energy,299,113694-113694.
[22] Ahmed A. Abdulhamed,Prabhat Ranjan Singh,Tanya Shakir Jarad & Shengwu Xiong. (2025). Artificial Bee Colony Algorithm with Multi-objective in Collaboration Edge Computing. International Journal of Cooperative Information Systems,(prepublish).