Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9872849, 1 page
https://doi.org/10.1155/2023/9872849

Retraction

Retracted: Analysis of College Students' Ability to Improve Innovation and Entrepreneurship Based on Constrained Clustering Algorithm

Wireless Communications and Mobile Computing

Received 29 August 2023; Accepted 29 August 2023; Published 30 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] C. Shi, "Analysis of College Students' Ability to Improve Innovation and Entrepreneurship Based on Constrained Clustering Algorithm," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 8562194, 10 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 8562194, 10 pages
https://doi.org/10.1155/2022/8562194

Research Article

Analysis of College Students' Ability to Improve Innovation and Entrepreneurship Based on Constrained Clustering Algorithm

Chen Shi

Ningbo University of Finance & Economics, Ningbo, Zhejiang 315175, China

Correspondence should be addressed to Chen Shi; shichen@nbufe.edu.cn

Received 31 May 2022; Revised 13 July 2022; Accepted 20 July 2022; Published 30 July 2022

Academic Editor: Kalidoss Rajakani

With the expansion of college enrollment, college graduates have continued to expand, and the employment situation has become more and more severe. As a new form of employment, innovation and entrepreneurship are becoming more and more important in college teaching. Entrepreneurial success is crucial. This paper proposes an entropy-based active learning method (ALPCS), which is divided into three stages: selection, exploration, and consolidation. The main contents are as follows: in the selection stage, the fuzzy c-means algorithm is used to obtain the membership of all samples, then calculate their Shannon entropy, and finally, select the sample with large Shannon entropy to generate an information subset (the larger the Shannon entropy, the greater the uncertainty, and the more information it contains). The distance-first strategy actively selects samples from the information subset to construct a cluster skeleton cluster. If it is equal to the real number of clusters, it enters the consolidation phase; otherwise, the active learning method stops. In the consolidation phase, sequentially from the information, the nonskeleton set points with the largest uncertainty are selected in the subset to form queries with the points in the skeleton set until the must-link constraint is formed. In this stage, the principle of minimum symmetric relative entropy first is used to reduce the number of queries. The ALPCS algorithm is compared and evaluated, and the final experimental results show that the ALPCS algorithm has a good performance when the number of queries is large.

1. Introduction

The early research object of innovation and entrepreneurship ability is about entrepreneurs, but in recent years, scholars have begun to focus on the research of college students, so the research on this ability of college students has gradually been paid more attention. In 2010, the Ministry of Education put forward new requirements for it, requiring that more attention should be paid to it in colleges, and the education target has also changed from the original entrepreneurs to all college students. To the whole process of college students' talent training, at the same time, the core of it is to improve college students' awareness of it and it [1]. In 2016, the Ministry of Education once again focused on this, emphasizing the reform content of it, and should continuously enhance the spirit of independent innovation, entrepreneurial awareness and innovation, and entrepreneurship ability of students [2]. In 2017, the State Council issued the "Thirteenth Five-Year Plan" for this. In terms of improving this, it proposed to cultivate it, strengthen the construction of these teachers, and deepen the teaching reform and management of this education courses. System reform encourages students to actively participate in this and activities [3].

There are endless information in the massive text, pictures, audio, and video data in life, but how to effectively use this huge information is placed in the big problem in front of us. The 4V characteristics of big data are volume (large scale), variety (diversity), velocity (fast processing), and value (high value). There is an urgent need for a new automated tool to help us to extract valuable information from massive data; researchers put forward the concept of "data mining." "Data mining" is a process of extracting potentially useful information and knowledge that people do not know in advance from a large number of incomplete, noisy, fuzzy, and random actual data. Data mining can be simply understood as the process of discovering useful knowledge through the operation of a large amount of data. It is an interdisciplinary subject involving a wide range of fields, including machine learning, mathematical statistics, neural networks, databases, pattern recognition, rough sets, fuzzy mathematics, and other related technologies. It is a potentially useful complex process of valuable information, which is a wide-ranging interdisciplinary subject that integrates knowledge in machine learning and data fields [4].

Clustering [5] is the data mining—it divides the sample set into different clusters according to the similarity between samples, so that two samples in the same cluster will be as similar as possible, and the different two samples in the cluster are as different as possible. In natural science and social science, there are many classification problems. Cluster analysis, also known as group analysis, is a statistical analysis method to study the classification of (samples or indicators). Cluster analysis originated from taxonomy, but clustering is not equal to classification. The difference between clustering and classification is that the classification required for clustering is unknown. The content of cluster analysis is very rich, including system clustering method, ordered sample clustering method, dynamic clustering method, fuzzy clustering method, graph theory clustering method, and clustering prediction method. This phenomenon is also reflected in real life. Under unknown circumstances, we will subconsciously distinguish between people and things based on some potential characteristics. The ancients are even more vivid. The clustering is a traditional unsupervised learning method, which means that it does not need to use labeled samples for learning, so its learning effect is often not satisfactory. In contrast, our evaluation of unknown things and people is often not accurate enough [6].

2. State of the Art

The active learning algorithm has a wide range of research and applications in supervised learning problems, but it is relatively limited in semisupervised learning problems. Lopez-Martinez-Carrasco [7] and others first proposed an active learning algorithm for this; the algorithm consists of two parts: explore (explore) stage and consolidate (consolidate) stage. In the exploration stage, they use a farthest distance-first traversal method to get multiple disjoint neighborhoods (and each neighborhood is required to contain at least one point), and these disjoint neighborhoods constitute the basic skeleton set of the clustering. In the consolidation phase, the remaining points in the dataset are randomly selected, and the centroids of the skeleton set are sequentially. Construct the query until the must-link constraint is obtained, and then, add the point to the skeleton set to which the corresponding centroid belongs. Iterate repeatedly until the maximum number of queries. Based on the FFQS algorithm, Hassan et al. [8] proposed the Min-Max algorithm. Compared to this, the algorithm improves the consolidation stage. During each iteration, they use the Min-Max criterion to pick the point with the greatest uncertainty (instead of randomly picking it).

Xu et al. proposed to select pairwise constraints by examining the spectral eigenvectors of similarity matrices. The algorithm identifies boundary points (two classes) and sparse points by detecting eigenvectors and then forms these points into a query. Unfortunately, it shows limited applicability, because it requires constructing too many queries and assumes that errors in clustering results only occur at boundary points. Furthermore, this method can only be applied to problems with two clusters (classes).

Rahim et al. [9] constructed an active learning framework for the document clustering task. The method performed clustering with it to the cluster assignments. The documents calculate the probability that they belong to cluster. In the process of pairwise constraint selection, the method selects the most uncertain pair among them [10]. If must-link is returned after the query, it will stop and go to the next iteration. Otherwise, the existing neighborhood will be queried for unassigned points until must-link is returned. Propose a neighborhood-based approach, a framework that focuses on samples. The uncertainty of the point in its neighborhood (rather than the pairwise uncertainty) selects the pairwise constraint with the largest amount of information. However, this method applies random forest, so it has a huge amount of computation [11].

Cheng and Liu [12] proposed an entropy-based method (AIPC). AIPC consists of two stages: preclustering and labeling. In this, the membership matrix of the dataset is obtained by using the FCM clustering algorithm and cluster centers. In the marking stage, the membership entropy is used to measure the samples. According to this, the samples are divided into strong samples (the uncertainty is small and close to the cluster centers obtained by preclustering) and weak samples (they also propose the second smallest symmetric relative entropy priority principle to improve query efficiency).

3. Methodology

3.1. Introduction to Semisupervised Learning. The former refers to the use of labeled data for learning and finally a model that outputs a predicted label, representing the tasks of classification and regression [13]. The latter refers to it for learning, and the representative task is clustering. Although supervised learning can achieve better results, in practical problems, the data we collect is often unlabeled, and adding labels to the data may consume a lot of manpower and material resources [14]. Therefore, scholars have proposed whether it is possible to use less-labeled data and a large amount of unlabeled data for learning and achieve the best possible learning effect. This is the idea of semisupervised learning [15].

Semisupervised learning has three basic assumptions to construct the relationship between the samples to be predicted and the model [16]: the manifold assumption, the cluster assumption, and the smoothness assumption [17].

(1) Manifold hypothesis

High-dimensional data is distributed on a low-dimensional manifold structure, and two samples that are close enough to each other have the same label.

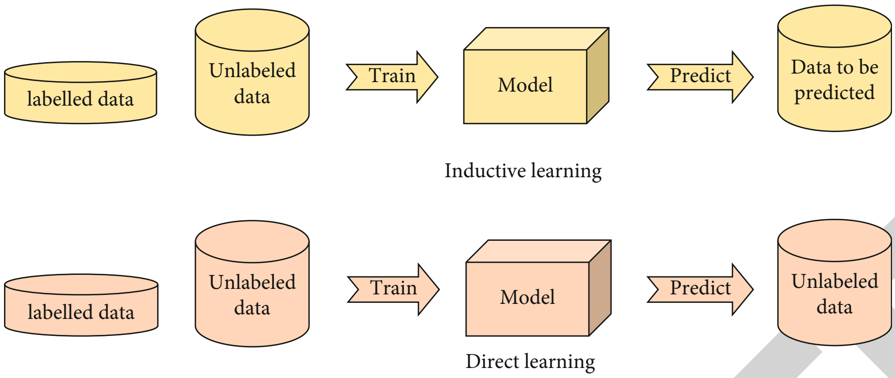

FIGURE 1: Inductive learning and direct inference learning.

(2) Clustering assumption

Assuming that two samples belong to the same cluster, they have the same label, a given decision region (if it is in a high-density boundary, a cluster may be divided into two different classes, violating the assumption).

(3) Smooth assumption

Two samples in a high-density space have the same label if they are close enough to each other.

The essence of these three types of hypotheses is that similar samples have the same label. Among them, the manifold hypothesis pays more attention to the local characteristics of the model, while the clustering hypothesis pays more attention to the overall model [18].

While the latter are exactly the samples to be predicted and the performance on these samples [19], in general, the difference between induction and direct inference is whether the samples to be predicted appear in the training. In the data, Figure 1 visually shows the difference between them.

By definition, we can get the following two properties:

Symmetry:

$$\begin{aligned}(x_i, x_j) &\in M \Rightarrow (x_j, x_i) \in M,\\ (x_i, x_j) &\in C \Rightarrow (x_j, x_i) \in C.\end{aligned} \tag{1}$$

Transitivity:

$$\begin{aligned}(x_j, x_k) &\in M \Rightarrow (x_i, x_k) \in M,\\ (x_j, x_k) &\in C \Rightarrow (x_i, x_k) \in C.\end{aligned} \tag{2}$$

Compared with class label information, pairwise constraint information is a weaker prior information, because pairwise constraint information only gives the relationship between two samples, so their specific labels cannot be inferred [20]. But pairwise constraint information is often easier to obtain. Using the above transitivity, more pairwise constraint information can be obtained, and pairwise constraint information can be converted from class label information, but not vice versa [21]. Supervised clustering has a wider range of applications.

Traditional clustering algorithms cannot directly utilize prior information, so these algorithms need to be improved.

A natural method is to use these pairwise constraints directly in the clustering process. The representative of this kind of algorithm is the pckmeans algorithm [22], which adds must-link constraints and cannot-link constraints on the basis of the K-means algorithm; its objective function is as follows:

$$J_{\text{pckmeans}} = \sum_{x_i \in X} \left\| x_i - \mu_{l_i} \right\|^2 + \sum_{(x_i, x_j) \in M} \omega_{ij} \Gamma(l_i \neq l_j) + \sum_{(x_i, x_j) \in C} \bar{\omega}_{ij} \Gamma(l_i = l_j). \tag{3}$$

The steps of this algorithm are similar to K-means. The objective function J_{pckmeans} is minimized, and finally, the cluster ilN corresponding to each sample *ix* is obtained.

Another method is to use pairwise constraint information to adjust the distance metric so as to minimize the distance between samples constituting the must-link constraint and maximize the distance between samples constituting the cannot-link constraint. The representative of this type of algorithm is mkmeans algorithm, which first trains a distance metric to satisfy, and then uses the distance metric for clustering. Its objective function is as follows:

$$J_{\text{mkmeans}} = \sum_{x_i \in X} \left(\left\| x_i - \mu_{l_i} \right\|_{A_{l_i}}^2 - \log \left(\det \left(A_{l_i} \right) \right) \right). \tag{4}$$

A is a symmetric positive definite matrix. ilA is the weight matrix corresponding to each cluster ilN. If it is a diagonal matrix, it will scale each dimension feature of the sample by weight. Otherwise, it is called Mahalanobis distance [23, 24] for the parametric representation.

Combining the objective functions of the two algorithms above, we get:

$$\begin{aligned}J_{\text{combine}} = {} & \sum_{x_i \in X} \left(\left\| x_i - \mu_{l_i} \right\|_{A_{l_i}}^2 - \log \left(\det \left(A_{l_i} \right) \right) \right)\\ & + \sum_{(x_i, x_j) \in M} \omega_{ij} \Gamma(l_i \neq l_j) + \sum_{(x_i, x_j) \in C} \bar{\omega}_{ij} \Gamma(l_i = l_j).\end{aligned} \tag{5}$$

We assume that the violation penalty coefficients ij and ij are the same, which means that all violations of the constraint will be treated equally. This means that under the current metric, the two points that constitute the must-link constraint are far away, so the metric needs to be modified. Correspondingly, the clusters corresponding to these two points also have this problem, so the penalty should affect the two clusters. Metric: the penalty function is as follows:

$$f_M(x_i, x_j) = \frac{1}{2}\|x_i - x_j\|^2_{A_{t_i}} + \frac{1}{2}\|x_i - x_j\|^2_{A_{l_j}}. \tag{6}$$

Similarly, the penalty for violating the cannot-link constraint between close points should be higher than for far distances. The penalty function is as follows:

$$f_C(x_i, x_j) = \left\|x_i' - x_j''\right\|^2_{A_{A_i}} - \|x_i - x_j\|^2_{A_i}, \tag{7}$$

where (x, x'') is the farthest distance under the first metric, f, and has nonnegativity, because its second item is always less than or equal to the first item.

On the basis of the above, Basu et al. proposed the MPCK-means algorithm, and its objective function is as follows:

$$\begin{aligned} J_{\mathrm{mpckmeans}} = & \sum_{x_i \in X}\left(\left\|x_i - \mu_{l_i}\right\|^2_{A_{l_i}} - \log\left(\det\left(A_{l_i}\right)\right)\right) \\ & + \sum_{(x_i, x_j)\in M} \omega_{ij} f_M \Gamma(l_i \neq l_j) + \sum_{(x_i, x_j)\in C} \bar{\omega}_{ij} f_C \Gamma(l_i = l_j). \end{aligned} \tag{8}$$

MPCK-means algorithm first uses the transitivity of pairwise constraints to expand must-link set M and cannot-link set C and uses these pairwise constraints to generate a neighborhood. If input = c, directly use the centroid of each neighborhood as the initial cluster center; if input < c, use the random perturbation of the centroid of the overall dataset to generate the remaining c-input centroids, and if $a > c$, select c among them; the principle is to make the c clusters mutually as far as possible.

3.2. Active Learning Algorithms. First, some mathematical symbols and abbreviations used below are introduced, as shown in Table 1.

The APLCS algorithm consists of three stages: the selection stage, the exploration stage, and the consolidation stage. In the selection stage, we use the FCM algorithm to select data with information from the original dataset to generate an information subset and make the latter two stages work. On this subset, in the exploration phase, we take the sample with the largest uncertainty in the information subset obtained in the selection phase as the initial point and then use our proposed farthest distance first strategy to build the cluster skeleton while collecting pairwise constraints. The algorithm enters the third stage; otherwise, the algorithm ends. In the consolidation stage, the nonskeleton set points in the selected information subsets and the representative points in each skeleton set are selected to form a query and collected at the same time. For constraints, this strengthens the skeleton. To reduce the number of queries, we propose the principle of minimum symmetric relative entropy first.

Table 1: Definitions of symbols in active learning algorithms.

Symbol	Meaning
X	Dataset
x_j	jth sample
C	Number of clusters
P	The ratio of the information subset to the dataset
u_{ij}	Membership
V_i	Cluster center
ε_0	Threshold
N_t	Skeleton set
M	Must-link constraint set
C	Cannot-link constraint set
D_{SKL}	Symmetric relative entropy

Figure 2 shows the algorithm flow chart; the left side is the active learning method proposed by us, and the right side is any underlying semisupervised clustering algorithm.

Let the dataset x be divided into c clusters, $\{4;,4., \cdots, H4,)$, $j = \{1, 2, \cdots, n\}$ is the membership vector of x, where represents the sample x, which belongs to the probability of the ith cluster. On this basis, we define the Shannon direct descendant of x, as:

$$E(x_j) = -\sum_{i=1}^{c} \mu_{ij} \ln \mu_{i_j}. \tag{9}$$

We first give a definition of the distance between points and sets used in this phase.

We define the distance between a point x and a set $Y = \{y, y, \cdots, ym\}$ as:

$$d(x, Y) = \frac{\sum_{j=1}^{w}\left\|y_j - x\right\|}{w}. \tag{10}$$

On the basis of this definition, we propose a farthest-distance-first strategy to select and construct skeleton sets.

The goal of this stage is to obtain a basic clustering skeleton (c disjoint skeleton sets) with the help of pairwise constraint information, which can be quickly achieved by using our proposed farthest distance first strategy. We select the sample xma with the largest uncertainty (Shannon's descendant) in the information subset S as the initial point in the skeleton set. Compared with the FFQS algorithm that randomly selects the initial point, such an operation can reduce randomness on the one hand, and on the other hand, pairwise constraints composed of points with large uncertainty have higher potential value. Next, we use the farthest distance first strategy to select a point x from the information subset S, and a point in the skeleton set N, to construct a query.

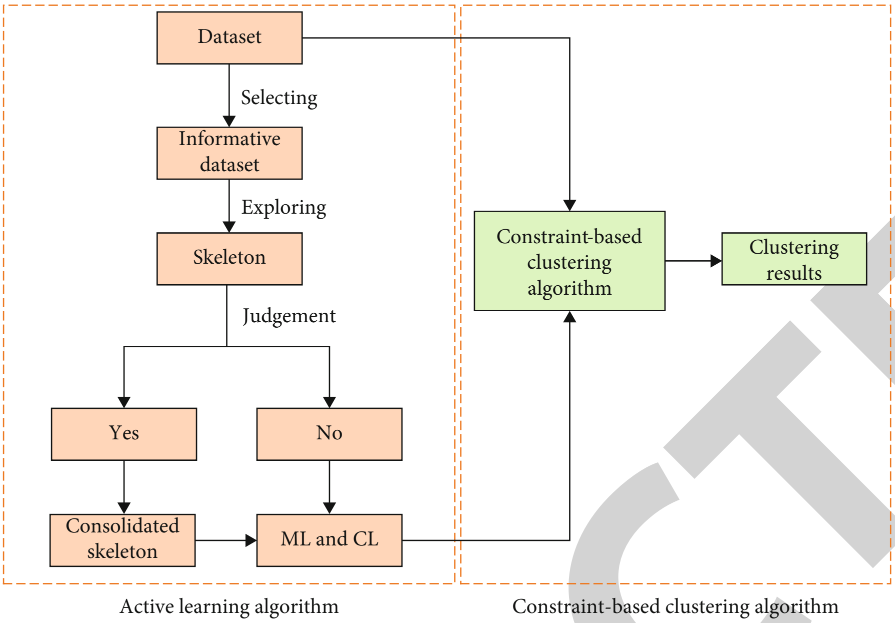

Figure 2: Algorithm flow chart.

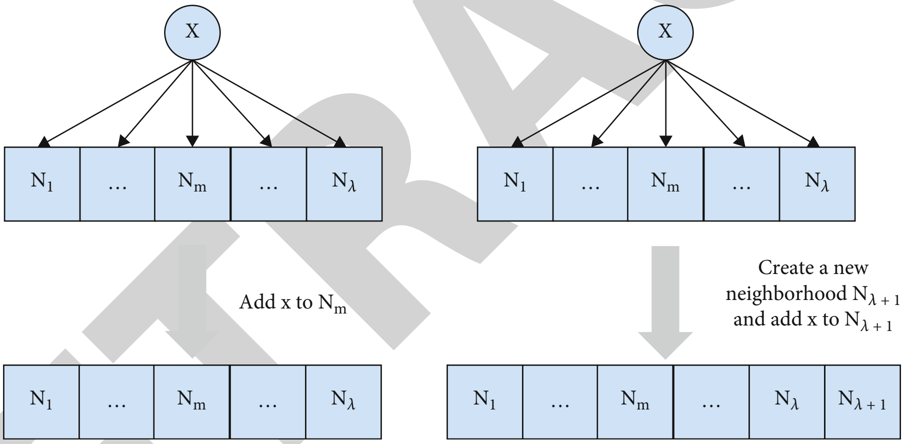
X
N_1
…
N_m
…
N_λ
Add x to N_m
Create a new neighborhood $N_{\lambda+1}$ and add x to $N_{\lambda+1}$
$N_{\lambda+1}$

Figure 3: Overview of the exploration phase.

The key steps of the algorithm are shown in Figure 3.

3.3. ALPCS Algorithm. If the ALPCS algorithm obtains c (number of clusters) skeleton sets in the exploration stage, the algorithm will enter the consolidation stage; otherwise, the algorithm ends. The consolidation stage, as the name suggests, is to consolidate the obtained c skeleton sets, so continue from the updated selected samples from the information subset 2S to add to the skeleton set.

At this stage, we need a distance to measure the similarity of the two samples; according to (11), we define the relative entropy between the two samples j_x and k_x as:

$$D_{KL}\left(x_j \| x_k\right) = \sum_{i=1}^{c} \mu_{ij} \ln \frac{\mu_{ij}}{\mu_{ik}} (1 \le j, k \le n). \quad (11)$$

Among them, u and u represent the membership degrees of the jth and kth samples corresponding to the ith cluster, respectively. The larger Dx, $(x, ll\, x)$, the greater the difference between x and x.

Considering that relative descendants do not have symmetry, we introduce:

$$D_{SKL}\left(x_j \| x_k\right) = \frac{1}{2} D_{KL}\left(x_j \| x_k\right) + \frac{1}{2} D_{KL}\left(x_k \| x_j\right). \quad (12)$$

The key steps of the algorithm are shown in Figure 4, the arrows represent the construction of the query between the two, and the solid lines represent the must-link constraints.

3.4. Strategies for Improving the Innovation and Entrepreneurship Ability of College Students. The improvement of the ability needs to start from multiple aspects, integrate all resources related to improving it, and need coordination and cooperation at all levels to jointly realize this. Based on the analysis of the reasons in the previous chapter, this paper proposes strategies for improving college students

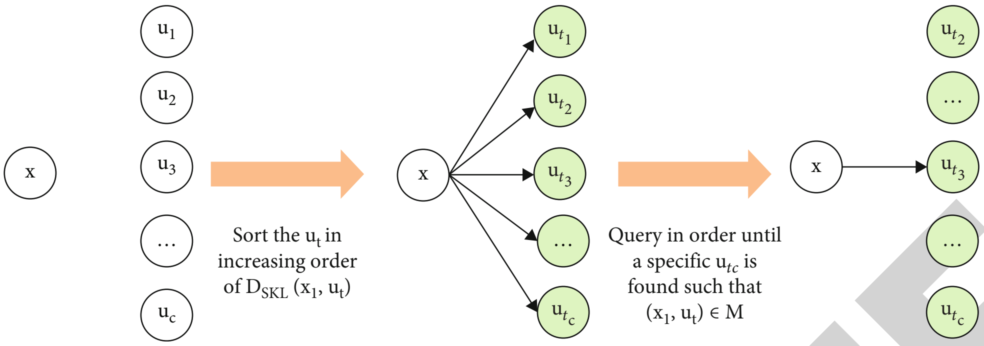
x
u_1
u_2
u_3
…
u_c
Sort the u_t in increasing order of D_{SKL} (x_1, u_t)
x
u_{t_1}
u_{t_2}
u_{t_3}
…
u_{t_c}
Query in order until a specific u_{tc} is found such that $(x_1, u_t) \in M$
x
u_{t_2}
…
u_{t_3}
…
u_{t_c}

Figure 4: Overview of the consolidation stage.

from five levels: individual, family, university, government, and colleges and universities. The individual aspects include establishing a correct concept of it, correcting the attitude of entrepreneurship and innovation, cultivating interest in it, and participating more in it. Activities cultivate the ability of independent learning. The family level includes changing family concepts and increasing entrepreneurial support. The college level includes creating a strong atmosphere of it, opening a variety of this education courses, cultivating entrepreneurial teachers, and enriching it to practice platform for college students. The government level includes increasing the publicity of the policies and increasing investment in entrepreneurship education. The colleges and universities strengthen the guidance of thought and practice and strengthens the cultivation of students' innovation and entrepreneurship awareness and ability; improve the innovation and entrepreneurship teaching system and strengthen the construction of entrepreneurship tutors; establish an innovation and entrepreneurship incubation center and jointly build an internship and training base through school enterprise cooperation; and organize and carry out second class activities to encourage college students to participate in innovation and entrepreneurship competitions.

Now is the arrival of the "Internet +" era, which provides a broader space for the teaching of innovation and entrepreneurship and brings more abundant learning resources into the classroom of independent learning innovation and entrepreneurship. Abundant teaching resources are the carrier of teaching activities and the medium for college students to carry out autonomous learning. Therefore, the construction of rich and diverse teaching resources is not only conducive to better teaching activities but also to college students' independent learning, innovation, and entrepreneurship. How to build rich teaching resources requires innovative resource media, timely update of learning platforms, and improvement of resource search functions, so that college students can obtain the latest and most useful knowledge in time and improve the efficiency of independent learning; the second is to develop more and more high-quality teaching resources, while selecting educational resources, actively select some personalized, innovative, and real-time resources, combined with the characteristics of current college students, and select and classify high-quality resources that are more conducive to the development of it of college students, to facilitate better service for college students.

Table 2: Dataset.

Dataset	Number of samples	Number of features	Number of clusters
Iris	150	4	3
Wine	178	13	3
Breast	683	9	2
Heart	270	12	2
Parkinsons	195	22	2
Ecoli	327	7	5

4. Result Analysis and Discussion

4.1. Measured Analysis. In this section, we evaluate the performance of the ALPCS algorithm by comparing it with five representative active learning methods. The MPCK-means algorithm introduced in Section 3 combines the ideas of constraints and metrics and is currently the best semioptimal algorithm, supervised clustering algorithm, so we directly choose it as the underlying clustering method. The detailed experimental settings and final experimental results will be given below.

In the experiments to evaluate the performance of constraint-based clustering algorithms, six groups of UCI datasets are widely used, so we also select these six datasets for experiments. For the dataset Ecoli, we remove the smallest three classes, which contain only 2 and 5 sample points, respectively. Table 2 presents the characteristics of these six datasets.

In order to evaluate the performance of ALPCS algorithm, we introduce five representative active learning methods for comparison. Below, we briefly introduce their basic principles.

(1) Random policy: randomly select points to construct pairwise constraints, which is often used as a benchmark for active learning research

(2) FFQS: contains two phases (exploration and consolidation). In the exploration phase, some is used to construct disjoint skeleton sets, and each skeleton set contains at least one point. In the consolidation phase, nonskeletons are randomly selected. Set points construct queries with representative points

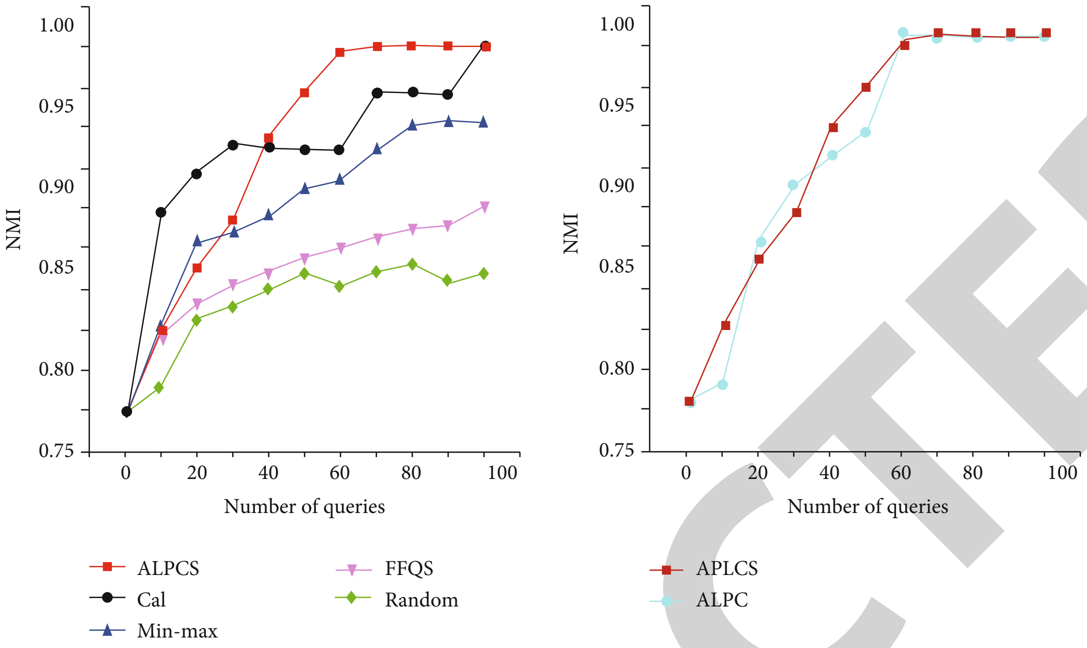

Figure 5: NMI values for different number of queries on the Iris dataset.

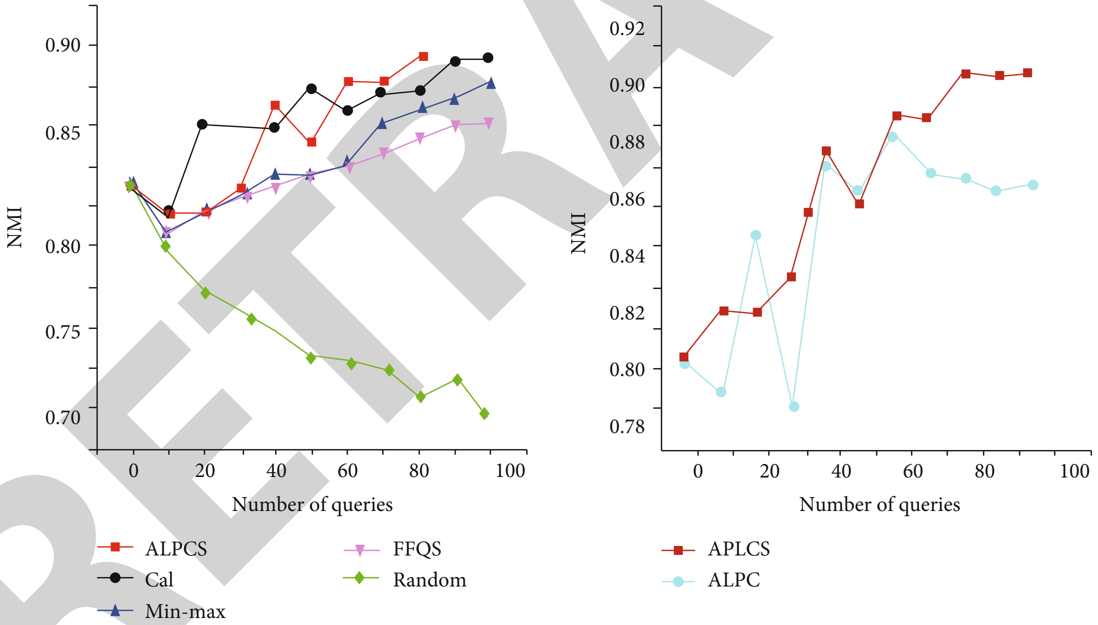

Figure 6: NMI values for different number of queries on the Wine dataset.

in each skeleton set until a must-link constraint is obtained

(3) Min-Max: FFQS, which is embodied in the consolidation phase using the maximum and minimum criteria to select nonskeleton set points and representative points in each skeleton set to construct a query

(4) Cai's method: an improved version of Min-Max, which is reflected in two aspects. One is to propose a method to select an information dataset from the original dataset and to make the next stage of exploration and consolidation act on the set (instead of the initial dataset); the second is to select the point with the largest uncertainty in the information dataset as the initial point of the exploration stage (rather than random selection)

(5) AIPC: it contains two stages of preclustering and labeling. First, weak samples and strong samples use a direct-based method. Then, a query is

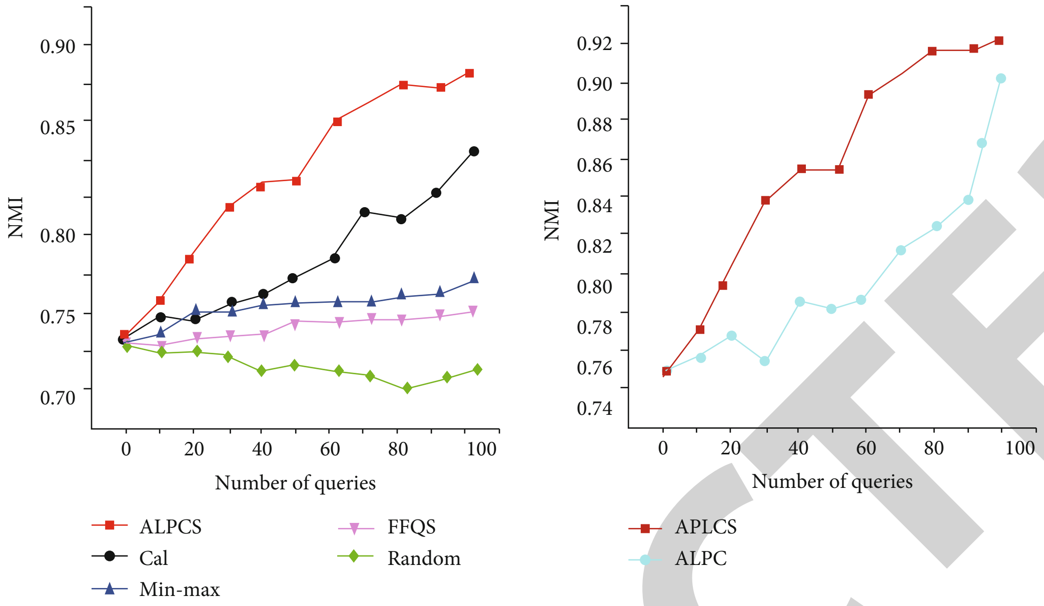

Figure 7: NMI values for different number of queries on the Breast dataset.

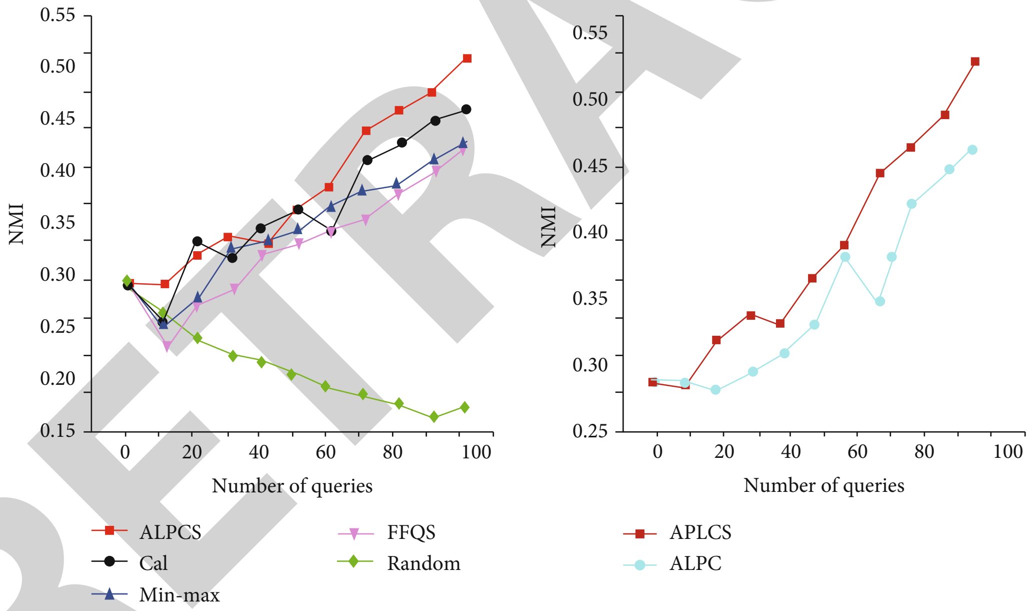

Figure 8: NMI values for different numbers of queries on the Heart dataset.

constructed between the weak samples and the strong samples. They use a certain first-order principle to reduce the number of queries

4.2. Evaluation Indicators. In the case of knowing the true labels of the data, we choose NMI 521 to evaluate the quality of the clustering effect. This is a practical random variable. The amount of information about another random variable contained in normalized mutual information treats both the true label and the predicted label of the data now that the output value is normalized to a range of zero to one. Let E be a random variable representing the predicted label of the sample, F is a random variable representing the true label of the sample, and NMI is defined as follows:

$$\mathrm{NMI} = \frac{2I(E, F)}{H(E) + H(F)}, \tag{13}$$

where $I(E, F) = H(E)H(E \mid F)$ is the random variable E and random variable F, $H(E)$ is the entropy of random variable E, and $H(E \mid F)$ is the conditional entropy of random variable E given the conditions of random variable F. The closer the value of NMI is to 1, the better the clustering effect.

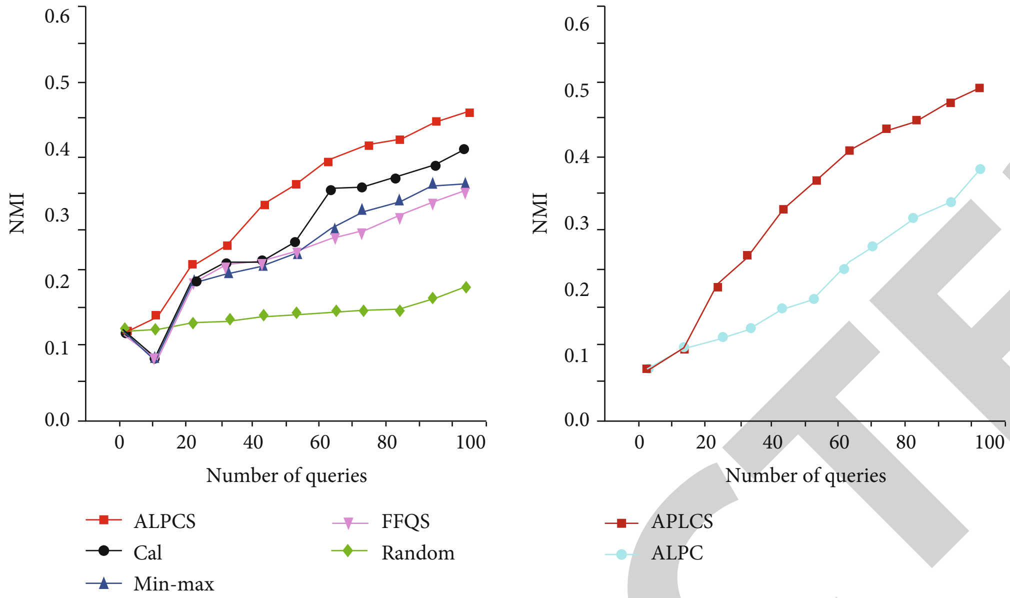

FIGURE 9: NMI values for different number of queries on Parkinsons dataset.

4.3. Experimental Results. Figures 5–9 present the experimental results on different datasets. The horizontal axis in the figure represents the number of pairwise constraints that constitute the query, and the vertical axis represents the normalized mutual information. The estimated value of NMI: as mentioned above, all experiment results are the averaged values after 50 independent runs. For better visual effects, we put the comparison between the neighborhood-based method and our proposed method on the left and put the comparison between the latest proposed descendant-based method AIPC with our proposed method on the right. It can be seen from Figure 5 that the closer the NMI values of ALPCS and ALPC are, when the number of queries reaches 60, the clustering effect basically remains unchanged with the increase of the number of queries, and the clustering effect of ALPCS on iris dataset is not significantly improved.

As can be seen from Figure 6, the clustering effect of ALPCS is significantly better than ALPC on the Wine dataset. When the number of queries is greater than 70, the NMI value of ALPCS exceeds 0.9.

From the Figures 5–9, we can see that our proposed active learning algorithm ALPCS (combined with MPCK-means algorithm) can usually achieve better clustering results. It is especially prominent on the Breast, Heart, Parkinsons, and Ecoli datasets, ALPCS The curves of, almost completely cover the other curves. On the Iris and Wine datasets, although our method works well with a small number of queries, the performance will gradually improve with the increase of the number of queries. With a large number of queries, our method significantly outperforms.

We also found some interesting phenomena: on the Wine, Breast, and Heart datasets, the performance of the random strategy degrades as the number of queries increases, and previous studies have shown similar results, which again shows that the selection of inappropriate pairwise constraint information may have a negative impact on the clustering results. Min-Max is slightly modified on the basis of FFQS, so the two have similar curve shapes. Compared with AIPC, the fluctuation of the ALPCS curve is smaller, which indicates that the ALPCS selected pairwise constraints are more appropriate. This phenomenon may stem from the fact that ALPCS constructs pairwise constraints on subsets of information while AIPC constructs pairwise constraints on the entire dataset.

5. Conclusion

Our active learning method, ALPCS, consists of three stages. In the selection stage, we use Shannon entropy to measure the uncertainty of the sample and use the FCM algorithm to obtain an informative subset. In the exploration stage, we quickly construct a clustering skeleton. In the consolidation phase, samples are selected from the information subset to strengthen the clustering skeleton. The ALPCS algorithm is inspired by Zhong's method and Cai's method, and its innovation lies in the use of entropy and its series of concepts and properties derived from it. In order to verify the performance of ALPCS, we selected five comparison methods and applied the obtained pairwise constraint information to the MPCK-means semisupervised clustering algorithm. The final experimental results show that with a small number of queries, the advantages of ALPCS are not obvious, but with the increase of the number of queries, the advantages of ALPCS are gradually reflected. In terms of comprehensive clustering effect and time complexity, ALPCS is even better.

Under different datasets, the performance of random strategies decreases with the increase of the number of queries. Previous studies have also shown similar results, which again shows that the selection of inappropriate constraint information may have a negative impact on the clustering results. Innovation and entrepreneurship education must be integrated into the teaching of professional courses in colleges and universities. On the basis of mastering the teaching content, in practical teaching, teachers should be good at mining, developing, and penetrating the information technology content in the existing curriculum, actively guide students' innovative and entrepreneurial ideas, and take advantage of opportunities to stimulate students' enthusiasm for information technology, so as to provide this opportunity for college students.

Data Availability

The figures and tables used to support the findings of this study are included in the article.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

The authors would like to show sincere thanks to those techniques which have contributed to this research.

References

[1] M. C. Thrun, "Distance-based clustering challenges for unbiased benchmarking studies," *Scientific Reports*, vol. 11, no. 1, article 98126, pp. 111–119, 2021.

[2] C. Huiqin, "Belief functions clustering for epipole localization," *International Journal of Approximate Reasoning*, vol. 137, no. 3, pp. 146–165, 2021.

[3] X. Wu, H. Zhou, B. Wu, and T. Zhang, "A possibilistic fuzzy Gath-Geva clustering algorithm using the exponential distance," *Expert Systems With Applications*, vol. 184, no. 1, article 115550, 2021.

[4] D. Zhao, X. Hu, S. Xiong et al., "k-means clustering and kNN classification based on negative databases," *Applied Soft Computing Journal*, vol. 110, no. 6, article 107732, 2021.

[5] H. Xie and L. Peng, "A density-based evolutionary clustering algorithm for intelligent development," *Engineering Applications of Artificial Intelligence*, vol. 104, no. 4, article 104396, 2021.

[6] V. Ramakrishnan, P. Chenniappan, R. K. Dhanaraj, C.-H. Hsu, Y. Xiao, and F. Al-Turjman, "Bootstrap aggregative mean shift clustering for big data anti-pattern detection analytics in 5G/6G communication networks," *Computers and Electrical Engineering*, vol. 95, no. 10, article 107380, 2021.

[7] A. Lopez-Martinez-Carrasco, J. M. Juarez, M. Campos, and B. Canovas-Segura, "A methodology based on trace-based clustering for patient phenotyping," *Knowledge-Based Systems*, vol. 232, no. 2, article 107469, 2021.

[8] B. A. Hassan, T. A. Rashid, and H. K. Hamarashid, "A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star," *Computers in Biology and Medicine*, vol. 138, article 104866, 2021.

[9] M. S. Rahim, K. A. Nguyen, R. A. Stewart, T. Ahmed, D. Giurco, and M. Blumenstein, "A clustering solution for analyzing residential water consumption patterns," *Knowledge-Based Systems*, vol. 233, no. 1, article 107522, 2021.

[10] L. Yu and C. Fuyuan, "Weighted matrix-object data clustering guided by matrix-object distributions," *Engineering Applications of Artificial Intelligence*, vol. 109, no. 12, article 104612, 2022.

[11] N. P. Barnett, T. DiGuiseppi Graham, and E. A. Tesdahl, "Peer selection and influence for marijuana use in a complete network of first-year college students," *Addictive Behaviors*, vol. 124, no. 16, pp. 76–88, 2022.

[12] Z. Cheng and X. Liu, "Establishment of the model on the expression and guidance of contemporary college students' demands in the cyberspace environment," *Complex & Intelligent Systems*, vol. 3, article 559, pp. 41–56, 2021.

[13] R. Bryant and L. Dundes, "Fast food perceptions: a pilot study of college students in Spain and the United States," *Appetite*, vol. 51, no. 2, pp. 327–330, 2008.

[14] M. N. Laska, S. Fleischhacker, C. Petsoulis, M. Bruening, and M. J. Stebleton, "Food insecurity among college students: an analysis of US State legislation through 2020," *Journal of Nutrition Education and Behavior*, vol. 53, no. 3, pp. 261–266, 2021.

[15] B. Ellison, M. Bruening, D. J. Hruschka et al., "Viewpoint: Food insecurity among college students: a case for consistent and comparable measurement," *Food Policy*, vol. 101, pp. 102031–102108, 2021.

[16] H. Evans, M. Nicodemus, L. Irvin et al., "Volunteer impact in an equine-assisted activities and therapy program on confidence and knowledge in college students," *Journal of Equine Veterinary Science*, vol. 76, no. 4, pp. 144–155, 2019.

[17] K. Abula, P. Gröpel, K. Chen, and J. Beckmann, "Does knowledge of physical activity recommendations increase physical activity among Chinese college students? Empirical investigations based on the transtheoretical model," *Journal of Sport and Health Science*, vol. 7, no. 1, pp. 77–82, 2018.

[18] J. Maloof and K. B. Vanessa, "White team study training in the college biology laboratory," *Journal of Biological Education*, vol. 39, no. 3, pp. 14–26, 2005.

[19] K. Angelini, M. A. Sutherland, and H. C. Fantasia, "College health center utilization among a sample of senior college women," *The Journal for Nurse Practitioners*, vol. 13, no. 10, article S1555415517306335, pp. e477–e480, 2017.

[20] F. Xingming and Q. Zitang, "Public attitude and policy selection of future energy sustainability in China: evidence of the survey of the college students," *Energy Policy*, vol. 165, no. 5, article S0301421522001860, pp. 112961–112967, 2022.

[21] V.-F. Barbara, "Being unprepared: a grounded theory of the transition of asthma self-care in college students," *Journal of Pediatric Nursing*, vol. 61, no. 1, article S0882596321002529, pp. 305–311, 2021.

[22] L. Qiao and Q. Tiaolan, "The development and management work innovation of university students by artificial intelligence and deep learning in the wireless network era," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 7774185, 10 pages, 2022.

[23] J. Park and O. S. Hong, "Factors affecting adjustment of first-year nursing students to college life: a descriptive correlational study," *Nurse Education Today*, vol. 102, no. 7, article S0260691721001684, pp. 104911–104989, 2021.

[24] J. Han, H. Cheng, Y. Shi, L. Wang, Y. Song, and W. Zhnag, "Connectivity analysis and application of fracture cave carbonate reservoir in Tazhong," *Science Technology and Engineering*, vol. 16, no. 5, pp. 147–152, 2016.