Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4465007, 13 pages
https://doi.org/10.1155/2022/4465007

Research Article

Clustering Ensemble Technology Based on Granular Computing to Extract Cervical Cancer Predictors

Ye-Cheng Wang,[1] Xu-Qing Tang ,[1] and Honglin Xu[2]

[1]*School of Science, Jiangnan University, Wuxi 214122, Jiangsu, China*
[2]*Wuxi Vocational Institute of Commerce, Wuxi 214122, Jiangsu, China*

Correspondence should be addressed to Xu-Qing Tang; txq5139@jiangnan.edu.cn

Received 7 January 2022; Revised 25 March 2022; Accepted 8 April 2022; Published 26 May 2022

Academic Editor: Ali Ahmadian

Background. Cervical cancer is the most common gynecological malignancy, and its incidence has tended to be younger in recent years. Through the analysis of high-throughput expression data, the identification of key genes in cancer and healthy individuals as predictors of cervical cancer is of great significance for the early detection and early treatment of cervical cancer. *Method*. Granular computing is a concept and computing paradigm to deal with problems through information granulation, and the process of granulation can be realized by means of clustering. Based on this, this paper proposes an AB method to obtain representative elements in a multiattribute data system. First, the evaluation index FHEI of the clustering structure is introduced, and Algorithm 1 is designed to obtain the optimal clustering structure of each attribute of the data system and use it as the base cluster. Secondly, based on the clustering ensemble technology of granular computing, Algorithm 2 is designed with the help of the concept of information entropy. The algorithm takes the base cluster as the input to obtain the optimal ensemble clustering structure. Finally, using the nearest center principle, the representative elements of each class in the optimal ensemble clustering structure are obtained. *Results*. In this paper, the differentially expressed genes (DEGs) are screened out by using the gene expression data of cervical cancer, and the scores of the four interaction relationships among the DEGs are used as a multiattribute data system and input into the AB method. The five representative elements obtained are RTTN, SAMD10, ZNF207, WAC, and METTL14, which are the predictors of cervical cancer. The classification accuracy of these predictors is as high as 98.82%. This paper also conducts a comparative study between the AB method and other classical methods on six independent gene expression datasets. The results show that the number of predictors obtained by the AB method is small but has a high classification accuracy in the classification of patient samples.

1. Introduction

In recent years, with the rapid development of high-throughput sequencing technology and the gradual reduction of costs, massive amounts of data have been accumulated in the field of biomedicine. Using machine learning, neural network, and other data analysis methods to apply biological data to disease research has become a research hotspot in recent years [1–4]. However, the rapid growth of data volume and data dimension makes it difficult for traditional tools and experimental methods to solve problems in complex biological systems. Therefore, simplifying the system and extracting key information in the data has become one of the means to solve this problem. Granular computing (GrC) is a computing paradigm that achieves dimensionality reduction through granular structures and is widely used in many fields [5, 6]. This paper uses the granularity point of view to study cervical cancer gene expression data in a coarse-grained and modularized manner, to find key genes that can be used for disease expression, and to provide support for the clinical diagnosis of cervical cancer.

Human beings have a recognized feature of problem solving; that is, people can observe and analyze the same problem from different levels and different angles, which is consistent with the inner thinking of granular computing.

Granular computing is an information processing concept and computing paradigm. Since Zadeh published the paper "Fuzzy Sets and Information Granularity" in 1979 [7], people have begun to pay attention to the research on information granularity. Zadeh believes that the concept of information granules exists in many fields in different forms, and it is an abstraction of reality. Information granulation is a way for humans to process and store information. In 1985, Hobss [8] published an article directly using "Granularity" as the title of the paper, discussing the decomposition and merging of particles, how to obtain particles of different sizes, and the models for generating particles of different sizes. Lin formally proposed the concept of granular computing in 1997 [9]. The earliest research granularity in our country is "Problem Solving Theory and Application," published by Zhang Bo and Zhang Ling [5] in 1990. The monograph extends the theory and method of quotient space to nonequivalent division, fuzzy equivalence relationship, and so on, studies the relationship between quotient space theory and rough set and fuzzy set theory, applies quotient space theory to the field of uncertainty, and further develops it into a granular computing theory based on quotient space theory, covering problems in many fields such as artificial intelligence. At present, the main theories of granular computing include quotient space, rough set, and fuzzy set and have been introduced into artificial intelligence, data mining, machine discovery, and other application fields [10–12]. The basic idea of granular computing is to use the basic principles, methods, techniques, and tools of granular computing on the basis of coarse-grained information so that computers can more effectively process uncertain, inaccurate, and incomplete massive data so as to solve complex problems. The basic principles and mechanisms of the problem are analyzed and solved.

In granular computing, the selection of granularity is closely related to research objectives and expert experience, but in practical engineering applications, the common method for obtaining granularity is clustering technology [13, 14]. Clustering is one of the most important tools in the field of pattern recognition and machine learning. Its purpose is to discover hidden and intrinsic relationships between patterns without supervision. In the clustering process, the Clustering Validity Index (CVI) is an important tool to measure the clustering effect and determine the optimal number of clusters [15]. CVIs mainly use mathematical knowledge to model and evaluate the effectiveness of clustering results. When the optimal value of the index is obtained, the corresponding clustering result is the optimal clustering of the dataset.

Some scholars have combined the idea of granular computing with clustering methods to carry out various researches. For example, in 2002, Bu Dongbo et al. [16] analyzed clustering and classification technology from the perspective of information granularity and tried to use the framework of information granularity principle to unify clustering and classifications. They point out that, from the point of view of information granularity, clustering is calculated under a unified granularity, while classification is calculated under different granularities, and a new classification algorithm is designed according to the principle of granularity. The application practice of large-scale Chinese text classification shows that this classification algorithm has a strong generalization ability. In recent years, Tang Xu-Qing and his team have also done a lot of work in the direction of granular computing and clustering. In 2013, Tang Xu-Qing et al. proposed several hierarchical clustering problems and analysis of fuzzy proximity relations based on granular space using strict mathematical descriptions [17]. On this basis, Li Yang et al. [18] proposed a method for constructing a coarse-grained viral protein evolutionary tree using influenza virus protein data. And on the basis of the granularity space theory, the research on the optimal clustering model is carried out [19]. In 2020, Tang Xu-Qing introduced the basic theory and model of granular space in detail in his book "Grain Size Space Theory and Its Application" and presented the application research of the basic theory, method, and model related to granular calculation in the analysis of ecosystem and biological network. This research is the work carried out on the basis of this book.

In the past few decades, scholars have used various technologies to develop a large number of clustering algorithms. Given a dataset, choosing different clustering algorithms and different parameters or even using different characteristics of the dataset may get different clustering results. In order to make full use of the complementarity and rich information in multiple clustering results, clustering ensembles technology as a powerful clustering tool has received more and more attention in recent years [20]. Clustering ensembles can obtain a more stable, accurate, and robust optimal clustering by combining multiple clustering results.

At present, scholars have developed a large number of successful clustering ensemble algorithms. For example, Dong Huang et al. proposed algorithms such as U-SPEC and U-SENC for high-dimensional data. Aiming at some limitations of the existing clustering ensemble methods, such as ignoring the problem of uncertain connections and lack of the ability to integrate global information to improve local links, the algorithm is more inclined to the integrated information at the object level and lacks the exploration ability at the high granularity level. Using the structural information of graphs, the team proposed a variety of clustering ensemble methods [21–24]. There are also many classic clustering ensemble algorithms as follows: (1) The method based on the coincidence matrix (CA) [25], which uses the CA matrix to measure the similarity between data points. (2) Voting method (Voting) [26], by considering the data partitions generated by different clusters, which conducts associated voting on samples in each independent run and compares standardized voting with fixed thresholds. (3) The method of information theory (InT) [27] that considers the cluster labels in the entire ensemble through the entropy criterion to estimate the uncertainty of each cluster, introduces a new ensemble-driven cluster validity measurement method, and proposes a locally weighted coincidence matrix to summarize the

integration of different clusters. Using the local diversity in the integration, two new consistency functions are further proposed. (4) Hypergraph method. A. Strehl and J. Ghosh proposed three hypergraph-based methods in the literature [13]: CSPA, HGPA, and MCLA. (5) The method of mixed model [14].

In this paper, our purpose is to design two algorithms to obtain the optimal clustering structure based on the idea of clustering ensemble technology in granular computing: Algorithm 1 and Algorithm 2. The new method obtained by combining the two algorithms is used to identify predictors of cervical cancer. The method includes the following three steps. In the first step, the DEGs of cervical cancer are screened out, and four interaction scores are obtained through the differential gene interaction network. The score data is used as the input of Algorithm 1, and the output structure is called the base cluster. The second step is to input the base cluster into Algorithm 2 to obtain the optimal ensemble clustering structure that fuses the four interaction characteristics of DEGs and then use the nearest center principle to select and screen out the representative genes in each category in the structure as predictors; after calculation, the classification accuracy of the six predictors is 98.82%. In the final step, the predictive ability of the predictors is tested by applying 6 independent datasets; and the result is that, compared with several other classical algorithms, the classification accuracy is still higher under the premise of a small number of predictors.

2. Method Design

2.1. Granular Computing and Optimal Clustering Structure Algorithm. Given a distance d on the universe of X, if it satisfies that $\forall x, y \in X, 0 \le d(x, y) \le 1$ and no one value in the distance sequence $\{d(x, y), d(y, z), d(z, x)\}$ exceeds the maximum value of the other two, then d is called the isosceles normalized distance on X[5]. $\forall \lambda \in [0, 1]$, define the collection

$$\begin{aligned} [x]_\lambda &= \{y | d(x, y) \le \lambda, y \in X\}, \\ X(\lambda) &= \{[x]_\lambda | x \in X\}. \end{aligned} \tag{1}$$

Call $X(\lambda)$ the granularity of isosceles normalized distance d on X with respect to λ, and $[x]_\lambda$ is the particle in $X(\lambda)$. For the two granularities $X(\lambda_1)$ and $X(\lambda_2)$ on X, $\forall x \in X$ has $[x]_{\lambda_1} \subseteq [x]_{\lambda_2}$, then the granularity $X(\lambda_2)$ is said to be no finer than $X(\lambda_1)$, which is recorded as $X(\lambda_2) \le X(\lambda_1)$ [6].

The set $\{X(\lambda) | \lambda \in [0, 1]\}$ of all possible granularities on the universe X, which is called the granular space of X guided by d, is denoted as $\aleph_d(X)$. In other words, if an isosceles normalized distance on X is given, then a granular space containing the finest granularity on X is given (i.e., the smallest element $\forall x \in X$ exists as a particle in $\aleph_d(X)$) [17].

Clustering is the embodiment of granularity space in practical applications, and the clustering process is the process of changing from fine-grained to coarse-grained. That is to say, granular computing is an abstraction of the idea of clustering. Granular computing can be mapped to nouns in clustering:

$$\begin{aligned} &\text{particles} \longleftrightarrow \text{category}, \\ &\text{granularity} \longleftrightarrow \text{clustering structure}. \end{aligned} \tag{2}$$

Therefore, the granularity of different thicknesses contained in the granularity space can be regarded as including clustering structures of different thicknesses. In order to compare the gap between particles inside and between particles, an index to measure the gap between particles and objects within particles is introduced: interclass difference S_{inter} and intraclass difference S_{intra}, which are calculated as follows:

$$\begin{aligned} S_{\text{inter}}(X(\lambda)) &= \frac{1}{N} \sum_{i=1}^{c_\lambda} J_i \left\| \overline{a}_i - \overline{a} \right\|_2^2, \\ S_{\text{intra}}(X(\lambda)) &= \frac{1}{N} \sum_{i=1}^{c_\lambda} \sum_{j=1}^{J_i} \left\| x_{ij} - \overline{a}_i \right\|_2^2. \end{aligned} \tag{3}$$

In (3), N is the total number of elements in the universe, $X(\lambda) = \{a_1, a_2, \ldots, a_{c_\lambda}\}$, where c_λ is the number of particles contained in the granularity $X(\lambda)$, $a_i = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{J_i}}\}$ represents the i-th particle in $X(\lambda)$, J_i is the number of elements in the particle a_i, $\overline{a}_i = 1/J_i \sum_{j=1}^{J_k} x_{ij}$ represents the center of particle a_i, $\overline{a} = 1/N \sum_{i=1}^{c_\lambda} \sum_{j=1}^{J_i} x_{ij}$ represents the center of set X, and $\| \cdot \|_2$ represents the 2-norm in K-dimensional space.

In the granular space $\aleph_d(X)$, selecting an appropriate clustering structure so that $\aleph_d(X)$ loses the least information and reflects the structural information of the complex system to the greatest extent is the key issue of granular computing. In the granulation process, as λ becomes larger, the particles gradually become finer, S_{inter} becomes larger, and S_{intra} becomes smaller, but the sum is always the same [28]. According to this property, this paper introduces the evaluation index *FHEI* to measure the clustering results:

$$\text{FHEI} = \left| S_{\text{inter}} - S_{\text{intra}} \right|, \tag{4}$$

and when the FHEI value reaches the minimum, it is the optimal granularity.

Hierarchical clustering algorithm is a classic clustering algorithm, which is mainly divided into two types: top-down and bottom-up. The top-down clustering algorithm treats all data points as a whole and then divides them continuously until the structure that best meets the needs is reached, just like a big tree that keeps branching out. The bottom-up clustering algorithm first regards each data point as a class and combines different subclasses to form a new large class until the conditions for terminating the merging are reached. In this paper, a bottom-up hierarchical clustering algorithm is adopted, (4) is used as the stopping criterion, and an optimal clustering structure extraction algorithm,

Algorithm 1, based on complex systems is proposed, which can obtain the optimal clustering structure of a single attribute of the data system. In Algorithm 1, Matrix_A is the matrix formed by the Euclidean distance between the input particles.

Step 1. **Initialize**: $i \leftarrow 0$; $\lambda_i \leftarrow 1$;

Input: $X(\lambda_i) = C = \{a_1, a_2, \ldots, a_n\}(a_i = x_i)$;

Calculate: $S_{\text{intra}}(X(\lambda_i)), S_{\text{inter}}(X(\lambda_i)), \text{FHEI}(X(\lambda_i))$;

Step 2. $i \leftarrow i + 1$; $A \leftarrow C$; $C \leftarrow \varnothing$; $\lambda_i \leftarrow \max R(a_i, a_j)$;

Step 3. $B \leftarrow \varnothing$; $a_j \in A$, $B \leftarrow B \cup a_j$, $A \leftarrow A \backslash a_j$;

Step 4. $\forall a_k \in A (k \neq j)$,

If $R(a_k, a_j) = \lambda_i$,

Then $B \leftarrow B \cup a_k$, $A \leftarrow A \backslash a_k$, $C \leftarrow C \cup B$;

Step 5If $A \neq \varnothing$

Then go to Step 3;

Else $X(\lambda_i) \leftarrow C$;

Step 6. If $X(\lambda_i) \neq X(\lambda_{i-1})$

Then calculate: $S_{\text{intra}}(X(\lambda_i))$, $S_{\text{inter}}(X(\lambda_i))$, $\text{FHEI}(X(\lambda_i))$;

Step 7. **If** $\text{FHEI}(X(\lambda_i)) < \text{FHEI}(X(\lambda_{i-1}))$

Then go to Step 2;

Step 8. **Output** $X(\lambda_i)$, $S_{\text{intra}}(X(\lambda_i))$, $S_{\text{inter}}(X(\lambda_i))$, $\text{FHEI}(X(\lambda_i))$;

Step 9. **End**.

2.2. Construction of Optimal Ensemble Clustering Structure Method. There are several isosceles normalized distances in the universe X. Based on the granular space, if two isosceles normalized distances d_1 and d_2 on the universe X are given, then two structural clusters X_1 and X_2 are given, and two granularity spaces $\aleph_{d_1}(X_1)$ and $\aleph_{d_2}(X_2)$ are further obtained. In order to synthesize the information of the two granular spaces to obtain a more refined and accurate ensemble granular space on the universe of X, define

$$X_1 \cap X_2 = \{a_i \cap b_j | a_i \in X_1, b_j \in X_2\}. \tag{5}$$

Define the distance $d(a,b) = \max\{d_1(a_1,b_1), d_2(a_2, b_2)\}$ on the granularity $X_1 \cap X_2$, where $a, b \in X_1 \cap X_2$, $a \subseteq a_i \in X_i$, and $b \subseteq b_i \in X_i$, $i = 1, 2$. Note that the granularity space guided by d on $X_1 \cap X_2$ is $\aleph_d(X_1 \cap X_2)$. It can be seen from Section 2.1 that there must also be an optimal granularity in the granularity space $\aleph_d(X_1 \cap X_2)$; this granularity integrates the information of the optimal granularity among $\aleph_{d_1}(X_1)$ and $\aleph_{d_2}(X_2)$, which can more accurately reflect the internal structure of the universe of discourse.

From the perspective of clustering, for a data system with a single characteristic, Algorithm 1 can be used to obtain its optimal clustering structure, while for a multicharacteristic data system, it is necessary to first obtain the optimal clustering structure corresponding to each attribute as the base cluster, then fuse the base clusters through the ensemble algorithm, and finally obtain the optimal ensemble clustering structure, as shown in Figure 1. Next, we will design multiple optimal clustering structure ensemble algorithms based on Algorithm 1 to obtain the optimal ensemble clustering structure of the multicharacteristic data system.

2.2.1. AA Method. For the dataset $X = \{x_1, x_2, \ldots, x_n\}$, x_i is the i-th object in it. For the M characteristics of X, using Algorithm 1, respectively, M optimal results can be obtained, denoted as set $\text{X} = \{X_1, X_2, \ldots, X_M\}$, and X_i is called the i-th base cluster. Combine the cluster structures in M base clusters according to (5) and denote it as $\text{E} = \{E_1, E_2, \ldots, E_K\}, K > M$.

Taking the set E as the initial object of the clustering and continuing the clustering according to the validity index, the clustering ensemble result can be obtained. If FHEI is used as the effectiveness index and combined with Algorithm 1 for clustering ensemble, it can be called the AA method. Olatz Arbelaitz et al. have pointed out in the literature [15] that different algorithms, even different configurations of the same algorithm, have not been proven to show the best clustering results in all situations. In order to avoid clustering errors that may be caused by the same validity index, a validity index based on information entropy is proposed, and Algorithm 2 is designed. The overall algorithm that obtains the base cluster from Algorithm 1 and the ensemble clustering from Algorithm 2 is called the AB method.

2.2.2. AB Method. In information theory, entropy [29] is a tool used to measure the average uncertainty of random variables. For a set of discrete random variables X, the calculation formula of entropy $H(X)$ is shown in (6):

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x), \tag{6}$$

where $p(x)$ is the probability function of X.

Joint entropy is used to measure the average uncertainty of multiple interrelated random variables. For a pair of discrete random variables (X, Y), the calculation formula of joint entropy $H(X, Y)$ is shown in (7):

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X), \\ &= -\sum_{x \in X} \sum_{y \in Y} p(xy) \log_2 p(y|x). \end{aligned} \tag{7}$$

If and only if the random variables $X_1, X_2, \ldots, X_n$ are independent of each other, the joint entropy of $X_1, X_2, \ldots, X_n$ is equal to the sum of the respective entropies, which is shown in (8):

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i). \tag{8}$$

In clustering ensemble, without considering the original data, in order to evaluate the reliability of each clustering ensemble result, we use the concept of entropy to mark each set with a cluster label.

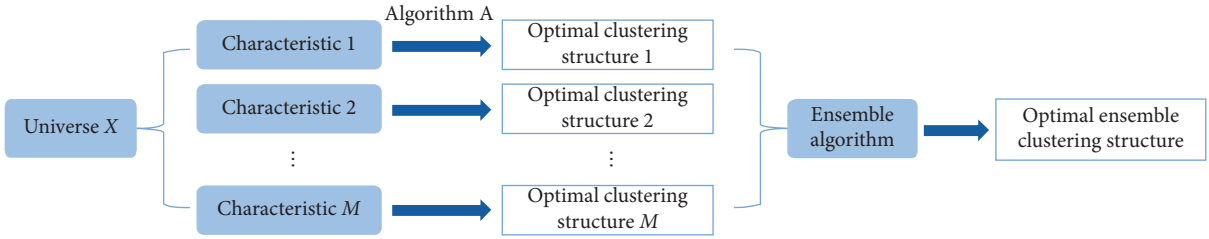

Figure 1: The construction process of optimal ensemble clustering structure.

Given the initial cluster structure $\mathrm{E} = \{E_1, E_2, \ldots, E_K\}$, for $E_i \in \mathrm{E}$, $X_m \in \mathrm{X}$, if $E_i \notin X_m$, then the object in E_i may belong to multiple clusters in X_m[27]. We use the concept of entropy to measure the uncertainty of cluster E_i relative to base cluster X_m, as shown in (9):

$$\begin{cases} H^m(E_i) = -\sum_{j=1}^{k(m)} p\left(E_i, X_j^m\right) \log_2 p\left(E_i, X_j^m\right), \\ p\left(E_i, X_j^m\right) = \dfrac{\left|E_i \cap X_j^m\right|}{\left|E_i\right|}, \end{cases} \tag{9}$$

where $k(m)$ and $|E_i|$ represent the number of objects in X_m and E_i, respectively.

Equation (9) gives the uncertainty of calculating cluster E_i relative to base cluster X_m, and it is easy to know that $p(E_i, X_j^m) \in [0, 1]$, $H^m(E_i) \in [0, +\infty)$. When cluster E_i completely belongs to the base cluster X_m, $|E_i \cap X_j^m| = |E_i|$; that is, $p(E_i, X_j^m) = 1$, $H^m(E_i) = 0$; when cluster E_i belongs to multiple clusters in the base cluster X_m, the value of $H^m(E_i)$ will increase, which means that, in the base cluster X_m, cluster E_i tends to belong to different clusters.

Without loss of generality, assuming that the base clusters in the set are independent of each other [30], then from (8), we can see that the uncertainty of cluster E_i relative to the entire set can be determined according to the sum of the uncertainties of cluster E_i relative to M base clusters, as shown in (10):

$$H^X(E_i) = \sum_{m=1}^{M} H^m(E_i). \tag{10}$$

It can be seen from (10) that the smaller the value of $H^X(E_i)$, the greater the probability that the objects in cluster E_i will be clustered into one category. Therefore, theoretically, the optimal cluster structure E_{opt} should be the cluster structure corresponding to the minimum $H^X(E_i)$ value, as shown in (11):

$$\begin{aligned} \mathrm{E}_{opt} &= \arg\min_{\mathrm{E}} H(\mathrm{E}, X), \\ &= \arg\min_{\mathrm{E}} \sum_{i=1}^{K} H^X(E_i). \end{aligned} \tag{11}$$

Take the initial cluster structure $\mathrm{E}_0 = \{X_1 \cap X_2, X_1 \cap X_3 \ldots X_{M-1} \cap X_M\}$ as input, merge the clusters with the largest cluster spacing, and calculate $H^X(E_i)$. As the number of clusters is merged more and more, the value of $H^X(E_i)$ gradually decreases. When a certain class is merged, the $H^X(E_i)$ value increases significantly, indicating that the merger has a greater disturbance to the cluster structure and is not conducive to the formation of the optimal cluster structure. Therefore, the previous cluster structure with a significant increase in $H^X(E_i)$ value is considered to be the optimal cluster structure. Based on this, an optimal clustering structure algorithm, Algorithm 2, is constructed, where $Matrix\_H(i, j) = H^j(E_i)$.

Step 1. **Input** $\mathrm{E} = \{E_1, E_2, \ldots, E_K\}$;

Step 2. **Calculate** $Matrix\_H$;

Step 3. **Calculate** $S\_\mathrm{inter}(E_i, E_j)$;

Step 4**For** all $S\_\mathrm{inter}(E_i, E_j) = \min S\_\mathrm{inter}(E_i, E_j)$, $\mathrm{E}_1 = (\mathrm{E}_0 \backslash E_i) \backslash E_j$, $B = E_i \cup E_j$, $\mathrm{E}_1 = \mathrm{E}_1 \cup B$;

Step 5. **Update** Matrix_H;

Step 6. **If** $H(\mathrm{E}_1, X) < H(\mathrm{E}_0, X)$

Then go to Step 3, $\mathrm{E}_0 = \mathrm{E}_1$, $\mathrm{E}_1 = \phi$;

Step 7. **Output** $\mathrm{E}_0 = \mathrm{E}_{opt}$;

Step 8. **End**.

3. Applications

Cervical cancer is one of the malignant tumors that seriously threaten women's health, and it is the fourth most common cancer among women in the world. According to global tumor epidemiology research reports, there are about 570,000 newly diagnosed cases of cervical cancer each year, of which about 311,000 cases of cervical cancer cause death [31]. The occurrence and development of cancer are often accompanied by complex interactions between genes and changes in their products. This complexity may be one of the main obstacles hindering clinical diagnosis [32]. Nowadays, with the rapid development of high-throughput sequencing technology and the reduction of costs, the biomedical field has accumulated massive amounts of data. Using data analysis methods to apply biological big data to disease research has become a research hotspot in recent years [33–35]. For example, from gene expression data, identifying key genes as predictors for inferring the classification of tumors and normal samples is of great significance for clinical diagnosis. In this paper, we use gene expression data to find DEGs and further find out the four characteristic scores between DEGs. Taking the score matrixes as the input of the AA method and the AB method, the optimal clustering ensemble structure is obtained, the representative of

Table 1: Basic information of cervical cancer dataset.

Data type	Data name	Genes number	Cases number	Primary tumor sample	Normal sample
Analysis set	TCGA-DNA	13125	307	304	3
	TCGA-miRNA	1881	310	307	3
Validation set	GSE6791	54675	28	20	8
	GSE7803	22283	31	21	10
	GSE9750	22284	57	33	24
	GSE63514	54675	52	28	24
	GSE52903	25294	72	55	17
	GSE29570	25294	62	45	17

each class in the clustering structure is selected as the predictors, and the biological analysis is carried out.

3.1. Data Source and Processing. Download the DNA and miRNA data of cervical cancer samples and normal samples from the TCGA database. Download six independent gene expression datasets GSE6791 [36], GSE7803 [37], GSE9750 [38], GSE63514 [39], GSE52903 [40], and GSE29570 [41] in the GEO database to verify the final key gene expression classification ability. The specific information of the data is shown in Table 1.

Use RStudio to process and filter the analysis set data. Firstly, the samples are filtered using the TCGAtumor_purity package to screen out samples with a tumor purity greater than 60%. Among them, 289 tumor tissue samples and 3 normal tissue samples are obtained after filtering the gene samples. 292 tumor tissue samples and 3 normal tissue samples are obtained after filtering miRNA samples. Furthermore, DEG screening is performed using the limma package. In this paper, $\log FC \geq 1$ and P value ≤ 0.05 are selected as the criteria, and finally, 3933 DEGs and 35 differentially expressed miRNAs (DE-miRNAs) are obtained. Finally, in the GeneMANIA database, 35 DE-miRNA target genes totalling 698 are found. The two datasets of DEG and DE-miRNA target genes are intersected, and finally, 4450 diff-genes are obtained.

Use the obtained diff-gene to construct a gene interaction network. In this network, nodes represent diff-genes, edges represent certain types of interactions between nodes, and the weights of edges are represented by interaction scores. Four interactions of 4450 differential genes are founded in the GeneMANIA database: coexpression interaction (CoExp), colocation interaction (CoLoc), gene interaction (GInc), and physical interaction (PhyInc). The data information is shown in Table 2 and Figure 2. It can be seen from Figure 2 that the score data of the four characteristics are evenly distributed, which is beneficial to the accuracy of clustering.

3.2. Experimental Results and Analysis. The scores of the four characteristics are brought into Algorithm 1, and four optimal clustering structures containing 6, 6, 109, and 13 classes are obtained as base clusters, respectively. The changes of S_{inter}, S_{intra}, and the evaluation index $FHEI$ with the clustering process are shown in Figure 3.

Table 2: Basic information of GeneMANIA score dataset.

Character	Number of genes involved	Minimum score	Maximum score
CoExp	3230	8.4e-04	0.064
CoLoc	2623	9.7e-04	1
GInc	2036	1.2e-04	1
PhyInc	3864	1.0e-04	1

Mark the obtained 134 initial base clusters as set X, and merge the cluster structures in X according to (5) to obtain an ensemble cluster structure E containing 1630 objects. For the AA method, using E as the initial input of Algorithm 1 again, the optimal clustering structure containing 45 categories is obtained. For the AB method, using E as the initial input of Algorithm 2, draw a graph of the change of $H(\mathrm{E}, \mathrm{X})$ with the clustering process, as shown in Figure 4(a). In order to find the point with the largest change in $H(\mathrm{E}, \mathrm{X})$ value more clearly, Figure 4(b) shows the absolute value change diagram of the difference between two adjacent points of $H(\mathrm{E}, \mathrm{X})$.

It can be seen from Figure 4 that, in the initial stage of clustering, the entropy value remains high. As the number of clusters increases, the cluster structure tends to the optimal solution, and the entropy value also drops to a low level and remains stable for a long time. However, when over-clustering, the cluster structure is far from the optimal solution, the entropy value increases again, and the increase is larger. After searching, the 746th point is the point before the entropy increase, and the corresponding cluster structure is 5 categories, which is the optimal clustering structure of the AB method.

From the perspective of hierarchical clustering, the particle signature can reflect the characteristics of this class to the greatest extent. Therefore, the particle signature can be used as a key gene extraction method. The particle signature is based on the principle of nearest to the center, and the object with the greatest similarity to other objects in each particle is selected as the characteristic representative of the particle. Therefore, the particle signature corresponding to each class in the optimal clustering structure is the predictor. The predictors obtained by the AA method are shown in Figure S1 in the Supplementary Material, and the five predictors obtained by the AB method are RTTN, SAMD10, ZNF207, WAC, and METTL14.

In order to verify the accuracy of the particle signature selection, on the basis of the signature set P, the particles in

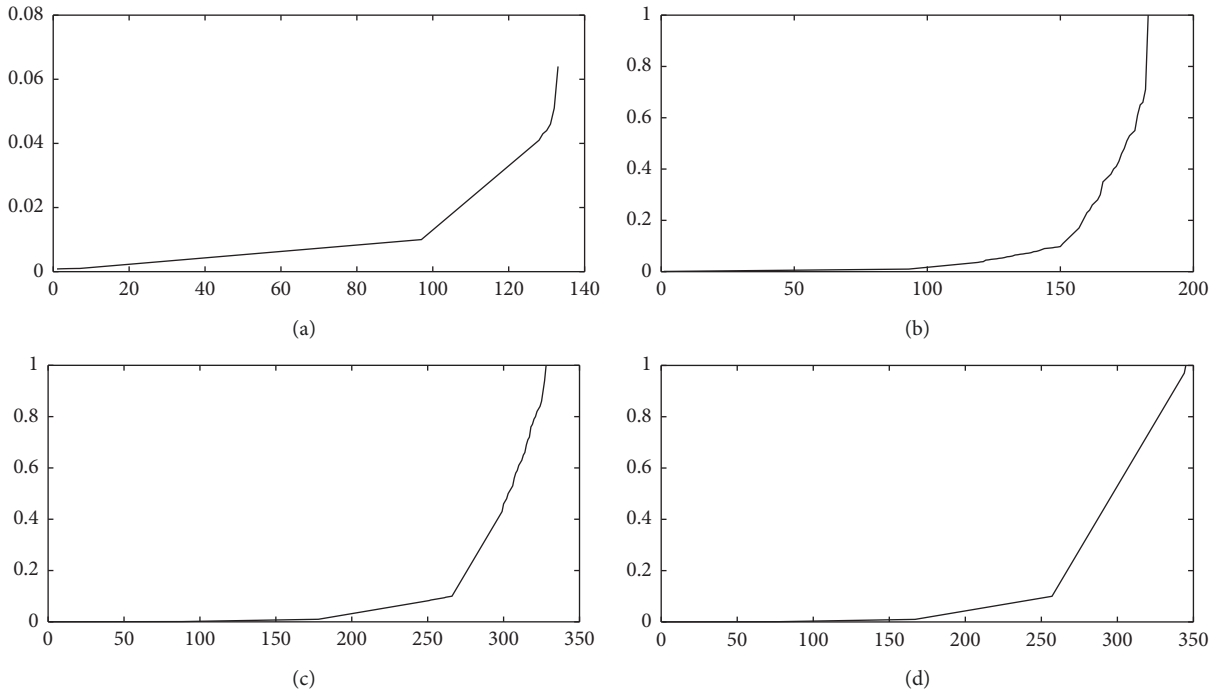

FIGURE 2: Score distribution diagrams of the four characteristics, of which (a), (b), (c), and (d) are CoExp, CoLoc, GInc, and PhyInc, respectively.

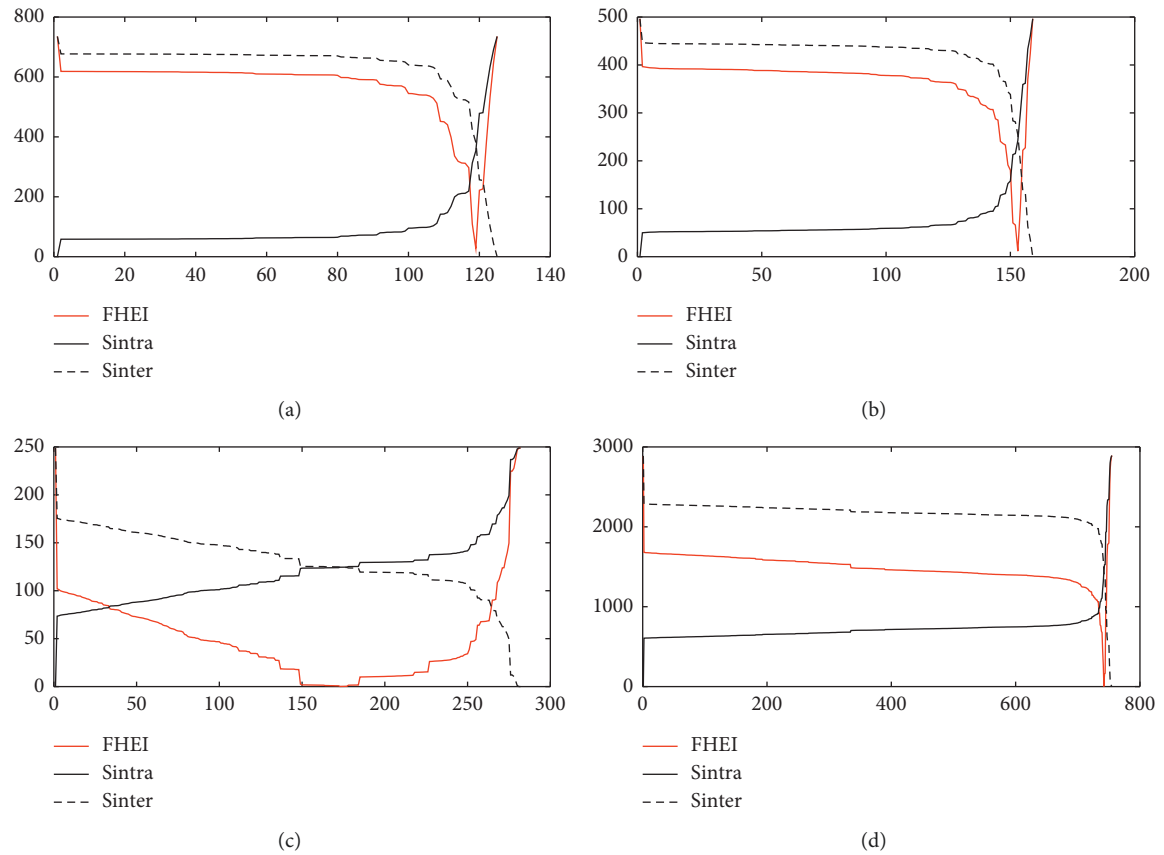

FIGURE 3: The variation of the S_{inter}, S_{intra}, and *FHEI* changes with clustering process of the four characteristics, of which (a), (b), (c), and (d) are CoExp, CoLoc, GInc, and PhyInc, respectively.

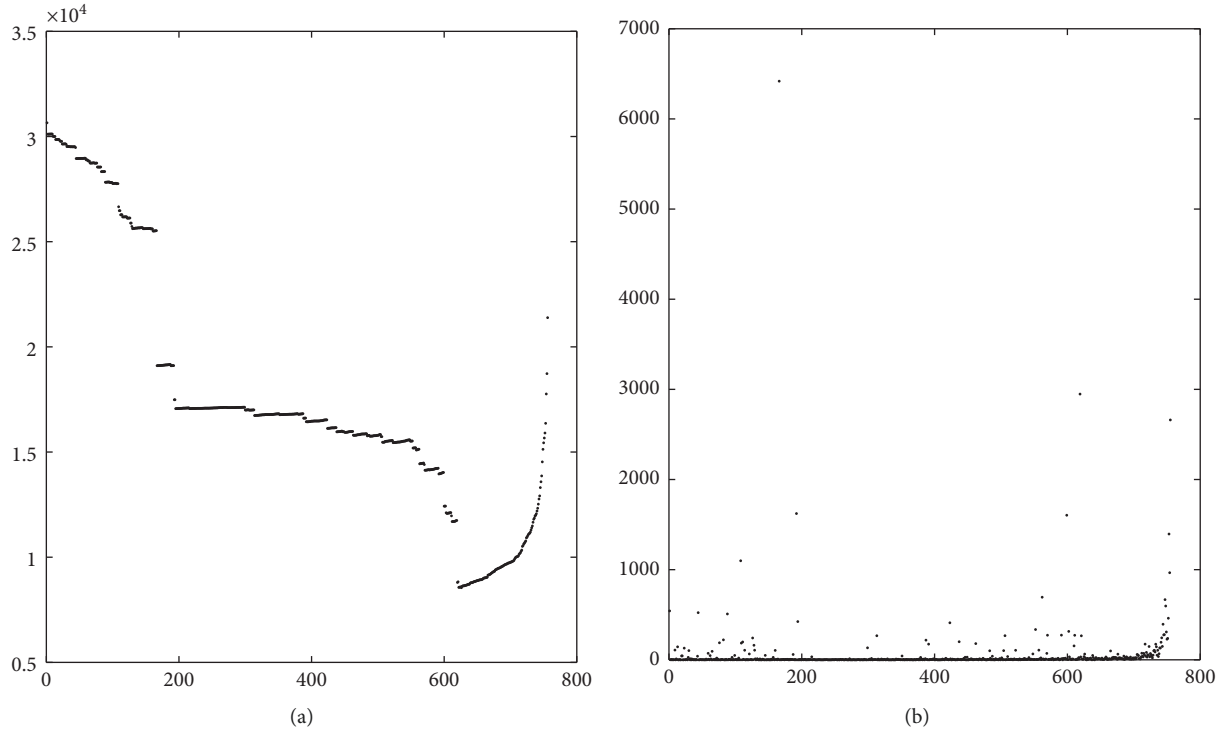

Figure 4: (a) Changes of H (E, X) with the clustering process. (b) Changes in the absolute value of difference between two adjacent points of H (E, X).

Table 3: Particle signature verification based on single clustering and ensemble clustering.

Cluster structure	Single clustering				Ensemble clustering
	CoExp	CoLnc	GInc	PhyInc	
Optimal number of clusters	6	6	109	13	5
r	72.8%	59.12%	18.69%	87.04%	98.82%

$X\backslash P$ are assigned to each particle signature according to the principle of nearest relationship, so as to construct a new granularity $\{b_1, b_2, \ldots b_{|p|}\}$. Compare the two particle sizes of $\{a_1, a_2, \ldots a_{|p|}\}$ and $\{b_1, b_2, \ldots b_{|p|}\}$ to find the particles with the same classification, and define the r value as the classification accuracy. The calculation of r is as (12):

$$r = \frac{\sum_{i=1}^{|P|} |a_i \cap b_i|}{|X\backslash P|}. \tag{12}$$

In (12), $r \in [0, 1]$, and the larger the value of r, the more accurate the selection of particle signatures and the higher the accuracy of the clustering results.

Table 3 shows the r value of the particle signatures extracted by the four initial base clusters and the optimal ensemble clustering obtained by the AB method relative to the 4450 differential gene sets. It can be seen from the table that the r value of a single cluster is significantly lower than that of the ensemble cluster. This is because the ensemble cluster makes full use of the information of multiple single clusters to obtain more accurate and superior clustering results.

Among the 5 key genes extracted by the AB method, RTTN, SAMD10, and ZNF207 are downregulated, and WAC and METTL14 are upregulated. The RTTN gene encodes a large protein. In view of the intracellular location of the protein and the phenotypic effect of mutations, this gene is suspected of playing a role in maintaining normal cilia structure, which in turn affects the development of left and right organs, axial rotation, and perhaps notochord development [42]. Experiments have shown that ZNF207 and ILF3 are the target genes for differential expression of miR-298 and miR-4261, and they are also transcription factors for the core gene EZH2 of the Polycomb family protein (PcG protein). The PcG protein participates in the regulation of embryonic development, has the ability to maintain cell self-renewal, participates in multiple cellular processes such as tumor occurrence and development and cell cycle regulation, and is abnormally expressed in a variety of tumor tissues [43]. Studies have found that the ubiquitination of histone H2B is very important for the assembly of chromatin during gene transcription, and the reduction of WAC expression level will destroy the ubiquitination level of H2B [44]. Epstein–Barr virus (EBV) is a ubiquitous

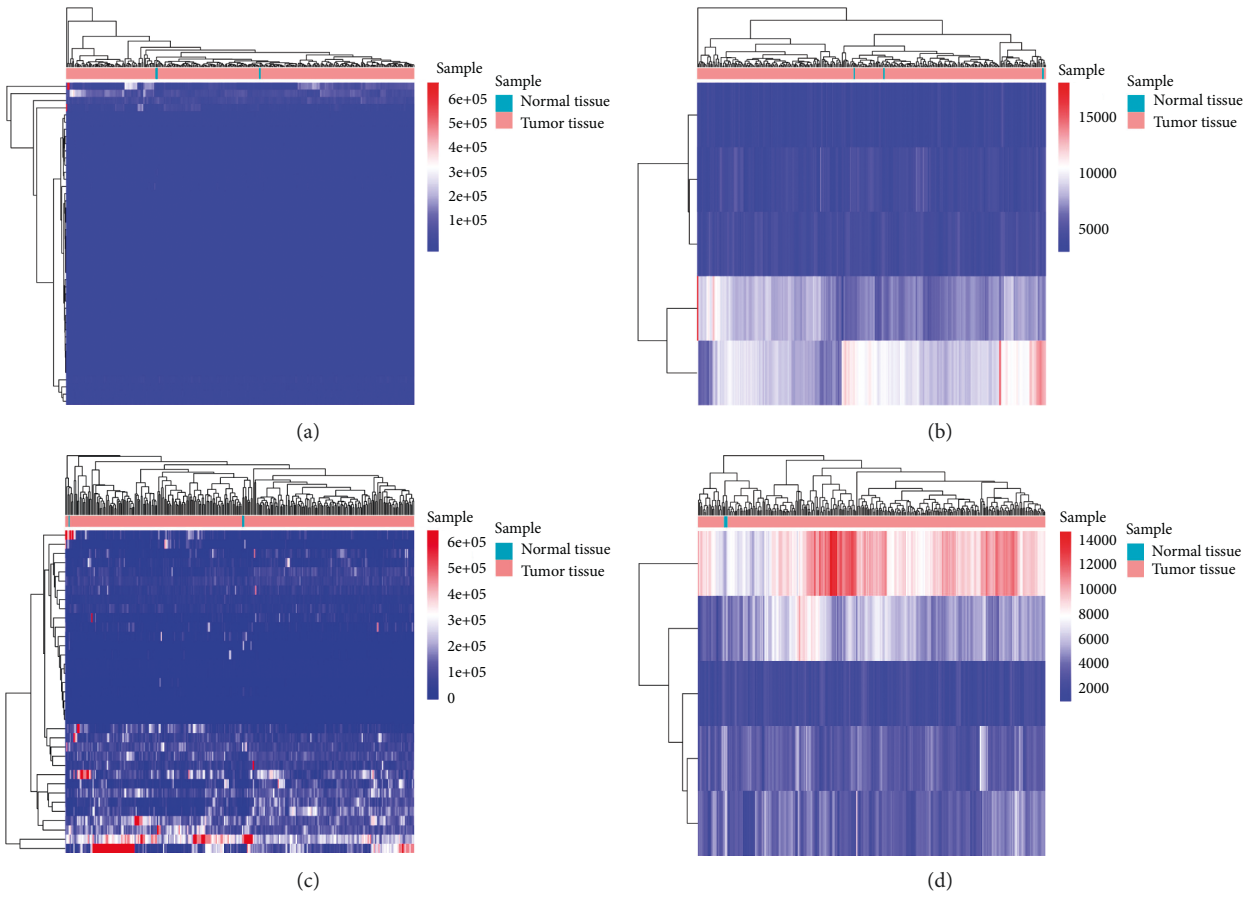

Figure 5: The expression heatmaps of key genes screened by four methods, of which (a), (b), (c), and (d) are AA method, AB method, CA method, and CSPA method, respectively.

carcinogenic virus that can induce a variety of cancers. And METTL14 can significantly induce EBV-positive tumors and promote the growth of EBV-transformed cells and tumors in xenograft animal models [45].

4. Method Comparison

In order to test the superiority and robustness of the method, compare the classification accuracy of the AA method and AB method proposed in this paper with the following two classic methods: (1) CA method: perform the K-means algorithm 10 times on the initial dataset, use the result as the base cluster, calculate the corresponding CA matrix, take 0.5 as the threshold, and the point pairs greater than 0.5 in the CA matrix are the same class in the final clustering result; (2) CSPA method: like the CA method, obtain the base cluster and calculate the CA matrix. The CA matrix generates a graph with vertices as data points, the CA value between the data points is the weight of the edges, then use the clustering algorithm METIS algorithm based on graph theory to cluster, and get the final ensemble clustering result.

Apply the above method to 4450 diff-genes; 35 (CA method) and 5 (CSPA method) optimal clustering structures are obtained, respectively. Then the key genes screened by different methods are obtained; see Figure S1 in the Supplementary Materials. Figure 5 shows the heatmap of the expression of key genes screened by the four methods, which proves the effectiveness of the four methods.

Further, in order to demonstrate the superiority and robustness of the method, the key genes screened are calculated and compared in the classification accuracy of cancer samples and healthy samples in six independent gene expression datasets downloaded from GEO. Different classifiers have their own advantages and disadvantages. In order to avoid the classification accuracy deviation caused by the classifiers, this paper uses three commonly used classifiers: decision tree (DTree), support vector machine (SVM), and random forest (RF). The classification accuracy is calculated through the rpart package, the e1071 package, and the randomForest package in RStudio. 80% of the dataset is randomly selected as the training set and 20% as the test set. The three classifiers are repeated 200 times for the key genes screened by each method, and the average precision is calculated.

Table 4 and Figure 6 show the classification accuracy of different clustering methods applied to different datasets under the three classifiers. It can be seen from Table 4 and Figure 6 that although the classification accuracy of the four methods is affected by the classifier and the test

Table 4: Classification accuracy of four clustering methods using three classifiers on six datasets.

Dataset / Classifier	Method	GSE6791	GSE7803	GSE9750	GSE63514	GSE52903	GSE29570
DTree	AA	0.67	0.86	0.83	0.73	0.8	0.77
	AB	1	0.71	0.83	0.73	0.8	0.46
	CA	1	0.86	0.83	0.91	0.93	0.85
	CAPA	0.83	1	0.5	0.64	0.73	0.92
Forest	AA	0.83	0.86	0.75	0.82	0.8	0.85
	AB	1	0.71	0.83	0.64	0.8	0.69
	CA	1	1	1	0.91	0.93	0.92
	CAPA	0.83	0.71	0.92	0.82	0.8	0.69
SVM	AA	0.83	1	0.92	0.82	1	0.85
	AB	1	0.71	0.83	0.45	0.8	0.69
	CA	1	1	1	0.82	0.8	0.85
	CAPA	0.83	0.71	0.83	0.73	0.8	0.69

The number "1" corresponding to the AB method in each classifier should be marked in italics in the GSE6791 column of data.

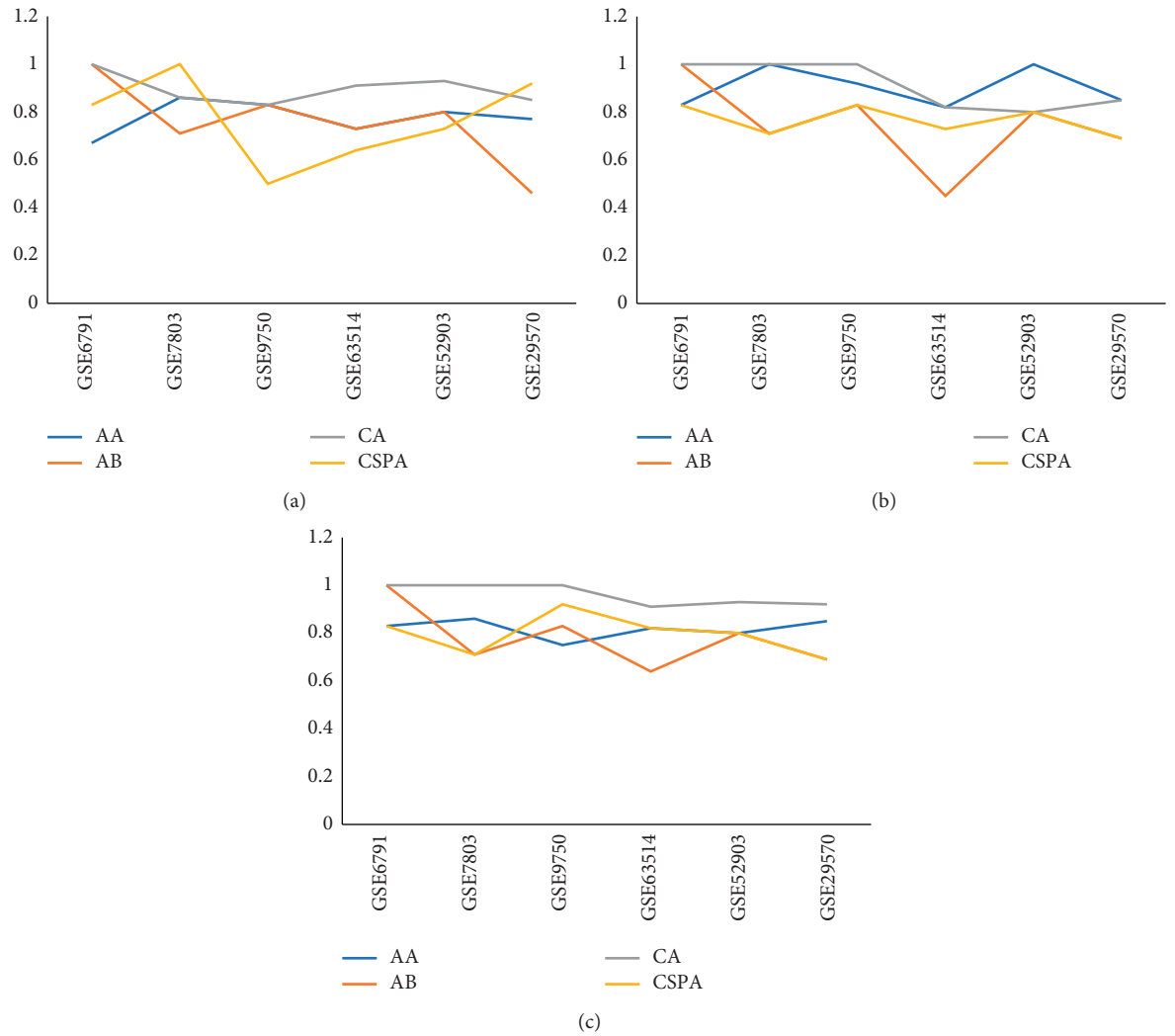

Figure 6: The classification accuracy rates of the classifiers, of which (a), (b), and (c) are DTree, SVM, and RF, respectively.

dataset, they all show better classification results. In particular, the AB method proposed in this paper has achieved 100% classification accuracy under different classifiers in the GSE6791 dataset (data marked in italics in Table 4). Due to the difference in the number of key genes obtained by the four methods, the number of key genes

screened by the AB method is the least, but it is still comparable to the classification accuracy of the key genes screened by the other three methods, which shows the superiority of the AB method.

5. Conclusions

Granular computing is a young research field. There is no clear definition and scope of research so far. Different researchers have different understandings of this field, but it can be roughly divided into two categories: one focuses on uncertainty processing, and the other type focuses on multigranularity computing. Among them, the idea of multigranularity computing is to use abstraction and layering to deal with problems, thereby reducing the complexity of dealing with complex problems. Data clustering is an ancient and active research field, and it has a very important meaning and function for studying the laws between discrete data. The clustering ensemble technology can increase the stability of the clustering process, that is, reduce the dependence on algorithm parameters, thereby improving the quality of clustering, enable different algorithms to collaborate when searching for consensus partitions, and consider problems from "multiple perspectives." Effectively solving the advantages of mixed numbers and classification features, missing values, and noisy clustering tasks has also become one of the hot research fields in recent years [46].

Cervical cancer is a malignant tumor that seriously threatens women's health. Although its occurrence and development process are complicated, disease predictors can be extracted through the analysis of gene expression data, thereby increasing the basis for clinical diagnosis. In this paper, with the help of concepts such as granular computing, clustering ensemble, and entropy, a new method is designed to identify predictors of cervical cancer. And a prediction accuracy of 98.82% is obtained under the premise of fewer predictors. Comparing the method proposed in this paper with other classical methods shows the superiority and robustness of the method in this paper and provides an effective method and basis for the analysis of biological data and the clinical diagnosis of diseases.

The research work in this paper focuses on the idea of granular computing and uses clustering methods to simplify the complex system. Through the identification of cervical cancer predictors, it provides some support and contributions to the clinical diagnosis of the disease, but there are also some problems: (1) The method has certain limitations. At present, only the hierarchical clustering algorithm is used as the construction method of granularity space, and future work will consider the comparison of different clustering algorithms. (2) Life is a dynamic development process; the work of this paper only focuses on static data. Using data at different time nodes to complete the process of discretized data points approaching dynamic changes can more effectively reflect the overall process of disease occurrence and development and have a deeper understanding of the disease.

Data Availability

The datasets generated during and analyzed during the current study are available in the TCGA repository: https://portal.gdc.cancer.gov.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (Grant no. 11371174).

Supplementary Materials

Figure S1 shows the key genes screened by the four clustering methods mentioned in Section 4. (*Supplementary Materials*)

References

[1] G. Senthilkumar, J. Ramakrishnan, J. Frnda et al., "Incorporating artificial fish swarm in ensemble classification framework for recurrence prediction of cervical cancer," *IEEE Access*, vol. 9, pp. 83876–83886, 2021.

[2] N. Benameur, M. Abed Mohammed, R. Mahmoudi et al., "Parametric methods for the regional assessment of cardiac wall motion abnormalities: comparison study," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 1233–1252, 2021.

[3] V. Lahoura, H. Singh, A. Aggarwal, M. A. Mohammed, and R. Damaševičius, "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, no. 2, Article ID 241, 2021.

[4] M. Aziz, S. A. Mostafa, C. Foozy, and A. ZaidAbualkishikc, "Integrating elman recurrent neural network with particle swarm optimization algorithms for an improved hybrid training of multidisciplinary datasets," *Expert Systems with Applications*, vol. 183, no. 12-2, Article ID 115441, 2021.

[5] L. Zhang and B. Zhang, *Problem Solving Theory and Application: Theory and Application of Granular Computing in Quotient Space*, Tsinghua University Press, Beijing, China, 2007.

[6] X. Q. Tang, *Granular Space Theory and its Applications*, Science Press, Beijing, China, 2020.

[7] L. A. Zadeh, "Fuzzy sets and information granulation," *Advances in Fuzzy Set Theory and Applications*, North-Holland Publishing, Amsterdam, Netherlands, 1979.

[8] J. R. Hobbs, "Granularity," in *Proceedings of the International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc, Los Angeles, CA, USA, August 1985.

[9] T. Y. Lin, "Granular computing: from rough sets and neighborhood systems to information granulation and computing in words," *European Congress on Intelligent Techniques and Soft Computing*, vol. 8-12, pp. 1602–1606, 1997.

[10] W. Pedrycz, *Granular Computing: Analysis and Design of Intelligent Systems*, Taylor & Francis Group, Oxfordshire UK, 2013.

[11] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, no. 2, pp. 111–127, 1997.

[12] X.-Q. Tang, P. Zhu, and J.-X. Cheng, "The structural clustering and analysis of metric based on granular space," *Pattern Recognition*, vol. 43, no. 11, pp. 3768–3786, 2010.

[13] A. Strehl and J. Ghosh, "Cluster ensembles-A knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 583–617, 2002.

[14] A. Topchy, A. K. Jain, and W. Punch, "A mixture model of clustering ensembles," in *Proceedings of the SIAM Intl. Conf. on Data Mining*, pp. 379–390, San Francisco, CA, USA, May 2003.

[15] O. Arbelaitz, I. Gurrutxaga, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[16] D. B. Bu, S. Bai, and G. J. Li, "Principle of granularity in clustering and classification," *CHINESE J. COMPUTERS*, vol. 25, no. 8, pp. 810–816, 2002.

[17] X.-Q. Tang and P. Zhu, "Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 5, pp. 814–824, 2013.

[18] Y. Li and X. Q. Tang, "Construction of phylogenetic tree of flu virus proteins based on coarse graining," *Pattern Recognition and Artificial Intelligence*, vol. 29, no. 10, pp. 936–942, 2016.

[19] X. Q. Tang, Q. H. Liang, and Y. Li, "Study on optimal clustering index based on granular space," *Systems Engineering - Theory & Practice*, vol. 38, no. 3, pp. 755–764, 2018.

[20] B. Minaei-Bidgoli, A. Topchy, and W. F. Punch, "A comparison of resampling methods for clustering ensembles," in *Proceedings of the Conf. on Machine Learning, Models, Technologies and Applications (MLMTA 2004)*, pp. 939–945, Las Vegas, Nevada, June 2004.

[21] D. Huang, C. D. Wang, H. Peng, J. Lai, and C. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 1–13, 2021.

[22] D. Huang, J. H. Lai, and C. D. Wang, "Robust ensemble clustering using probability trajectories," in *Proceedings of the Closer-international Conference on Cloud Computing & Services Science. DBLP*, Porto, Portuga, April 2012.

[23] D. Huang, C. D. Wang, J. H. Lai, and C. Keong Kwoh, "Toward multidiversified ensemble clustering of high-dimensional data: from subspaces to metrics and beyond," *IEEE Transactions on Cybernetics*, 2021.

[24] D. Huang, C. D. Wang, J. Wu, J. Huang Lai, and C. Keong Kwoh, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, 2019.

[25] A. L. N. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Proceedings of the 16th International Conference on Pattern Recognition, (ICPR 2002)*, pp. 276–280, Quebec, Canada, August 2002.

[26] A. Fred, "Finding consistent clusters in data partitions," in *Proceedings of the International Workshop on Multiple Classifier Systems, MCS 2001*, vol. 2096, pp. 309–318, Cambridge, UK, July 2001.

[27] D. Huang, C. D. Wang, and J. H. Lai, "Locally weighted ensemble clustering," *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1460–1473, 2016.

[28] X. Q. Tang, Y. Li, W. W. Li, and W. Shen, "A novel method for constructing the optimal hierarchical structure based on fuzzy granular space," *Applied Soft Computing*, vol. 87, Article ID 105962, 2020.

[29] Z. Y. Fu, *Information Theory: Principles and Applications*, Publishing House of Electronics Industry, Beijing, China, 4th edition, 2015.

[30] S. Vega-Pons and J. Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 3, pp. 337–372, 2011.

[31] M. Arbyn, E. Weiderpass, L. Bruni, J. Ferlay, and F. Bray, "Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis," *Lancet Global Health*, vol. 8, no. 2, 2019.

[32] X. M. Fan, Y. L. Wang, and X. Q. Tang, "Extracting predictors for lung adenocarcinoma based on granger causality test and stepwise character selection," *BMC Bioinformatics*, vol. 20, no. 7, pp. 83–96, 2019.

[33] I. W. Taylor, R. Linding, D. Warde-Farley et al., "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature Biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[34] R. Ben-Hamo, M. Gidoni, and S. Efroni, "PhenoNet: identification of key networks associated with disease phenotype," *Bioinformatics*, vol. 30, no. 17, pp. 2399–2405, 2014.

[35] X. Fan, P. Zhu, and X.-Q. Tang, "VD-analysis: a dynamic network framework for analyzing disease progressions," *IEEE Access*, vol. 8, pp. 153202–153214, 2020.

[36] D. Pyeon, M. A. Newton, P. F. Lambert et al., "Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers," *Cancer Research*, vol. 67, no. 10, pp. 4605–4619, 2007.

[37] Y. Zhai, R. Kuick, B. Nan et al., "Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion," *Cancer Research*, vol. 67, no. 21, pp. 10163–10172, 2007.

[38] L. Scotto, G. Narayan, S. V. Nandula et al., "Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression," *Genes, Chromosomes and Cancer*, vol. 47, no. 9, pp. 755–765, 2008.

[39] J. A. den Boon, D. Pyeon, S. S. Wang et al., "Molecular transitions from papillomavirus infection to cervical precancer and cancer: role of stromal estrogen receptor signaling," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 25, pp. 3255–3264, 2015.

[40] I. Medina-Martinez, V. Barrón, E. Roman-Bassaure et al., "Impact of gene dosage on gene expression, biological processes and survival in cervical cancer: a genome-wide follow-up study," *PLoS One*, vol. 9, no. 5, Article ID e97842, 2014.

[41] M. Guardado-Estrada, I. Medina-Martínez, E. Juárez-Torres et al., "The amerindian mtDNA haplogroup B2 enhances the risk of HPV for cervical cancer: de-regulation of mitochondrial genes may be involved," *Journal of Human Genetics*, vol. 57, no. 4, pp. 269–276, 2012.

[42] M. Zakaria, A. Fatima, J. Klar et al., "Primary microcephaly, primordial dwarfism, and brachydactyly in adult cases with biallelic skipping of RTTN exon 42," *Human Mutation*, vol. 40, no. 7, pp. 899–903, 2019.

[43] T. Gui, *Tumor Heterogeneity in Recurrent Ovarian Cancer as Demonstrated by Polycomb Group Proteins Expression*, Peking Union Medical College Hospital, Beijing, China, 2013.
[44] F. Zhang and X. Yu, "WAC, A functional partner of RNF20/40, regulates histone H2B ubiquitination and gene transcription," *Molecular Cell*, vol. 41, no. 4, pp. 384–397, 2011.
[45] F. Lang, R. K. Singh, Y. Pei, S Zhang, K Sun, and E. S Robertson, "EBV epitranscriptome reprogramming by METTL14 is critical for viral-associated tumorigenesis," *PLoS Pathogens*, vol. 15, no. 6, Article ID e1007796, 2019.
[46] V. Berikov, "Weighted ensemble of algorithms for complex data clustering," *Pattern Recognition Letters*, vol. 38, pp. 99–106, 2014.