ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

Improvement of Clustering Algorithms Based on Energy Distance and Its Application in Cancer Genes

Xiaoxia Deng[1, a], Shanru Chen[1] and Shanshan Li[2, *]

[1]College of school of Mathematics and Statistics, Yulin Normal University, Yulin, China

[2]College of Biology and Pharmacy, Yulin Normal University, Yulin, China

[a]xiaoxiaa0513@163.com, *Correspondence: lishanshan33@ylu.edu.cn

Abstract

The energy distance, compared to traditional distance measures in clustering algorithms, has the ability to differentiate between classes with small mean differences. In this study, we improved the measurement method of traditional clustering algorithms using the energy distance and empirically demonstrated its effectiveness in cancer genomics data. The improved clustering algorithm outperformed traditional methods in both numerical cancer gene expression data and non-numerical cancer gene survival data, showing superior classification performance and enhanced stability.

Keywords

Clustering algorithm; Energy distance; Cancer genes.

1. INTRODUCTION

In gene datasets, clustering analysis is a method used to uncover unknown gene information. This approach does not require labeled data, thus giving it an advantage in clustering analysis, particularly for the analysis of gene expression profiles data. Common clustering algorithms include partition-based algorithms (such as K-Means, Fuzzy C-Means), hierarchical algorithms (such as Clustering Using Representative, Chameleon), grid-based algorithms (such as MAFIA, Statistical Information Grid), density-based algorithms (such as Density-Based Spatial Clustering of Applications with Noise, Ordering Points to Identify the Clustering Structure), and model-based algorithms (such as Self-Organizing Maps, Expectation-Maximization Algorithm).[1-6] Emerging clustering methods include spectral clustering and probabilistic graphical model clustering.

In 2013, Liu Wenyuan et al. applied manifold learning-based clustering analysis methods of cancer gene expression data to gastric cancer gene expression datasets and leukemia gene expression datasets. They analyzed and evaluated the dimensionality reduction visualization effects of the Local Linear Embedding (LLE) algorithm and the Improved Distance and Multiple Weights Locally Linear Embedding (DMLLE) algorithm on cancer gene expression data. They also analyzed the clustering results after dimensionality reduction, obtained some biological explanations, and were able to identify genes with similar functions.[7]In 2013, Wang Xuehong demonstrated that traditional clustering algorithms are limited to capturing the global information in gene expression data. However, biclustering algorithms address this limitation and can obtain partial subsets of genes and condition sets, thereby capturing the local information in gene expression data more comprehensively.[8]In 2020, Wei Lixin proposed a Consensus Clustering with Davies-Bouldin Index (CC-DBI) algorithm that combines the Davies-Bouldin Index (DBI) to evaluate the clustering results and select the optimal number of clusters. It has been verified that the CC-DBI algorithm has the ability to select the best or suboptimal

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

number of clusters in cancer gene expression profile data. Additionally, this algorithm can also discover new cancer subtypes and exhibit characteristics of interactions between genes.[9]In 2022, Zhou Wengang et al. quantumbehaved Particle Swarm Optimization with Comprehensive Learning Strategy(CLQPSO) and Generalized Regression Neural Network (GRNN) are studied, A cancer gene clustering algorithm was generated based on CLQPSO. CLQPSO algorithm can make full use of each particle best position and particle swarm social cooperation information offered, avoiding premature convergence in local optimum value. Experiments show that the integrated use of GRNN and CLQPSO algorithm has better clustering performance and global convergence compared with K-Means, spectral clustering, discrete particle swarm algorithm in the aspect of cancer gene expressing data clustering.[10]

In 2011, Cai Xianfa et al. proposed an improved distance-based Locally Linear Embedding (LLE) algorithm and applied it to breast cancer and ovarian cancer gene expression profile datasets. Experimental results demonstrated that the improved distance-based LLE algorithm achieved higher classification accuracy compared to the traditional LLE algorithm.[11]In 2014, Rodriguez et al. proposed CFDP, a clustering algorithm that can rapidly search for and discover density peaks. CFDP is capable of automatically identifying and removing outliers, and its clustering results are independent of the spatial dimensions of the data.[12]Rashid Mahmoud et al. conducted clustering experiments on gene expression microarray data using the CFDPC algorithm. The results validated that the density peaks clustering algorithm can discover the underlying structure of the data and obtain accurate subtype clustering results.[13] Peng Wu et al. utilized the genomic, transcriptomic, and methylation HMK450 data from the cancer genome atlas network to study clear cell renal cell carcinoma (ccRCC). They applied a consensus clustering algorithm to the expression data and discovered three subtypes.[14]

In summary, both domestic and international researchers have made improvements to clustering algorithms and applied them to the analysis of cancer genes. These improved algorithms can assist researchers in conducting more accurate analyses using gene expression data, thereby revealing the biological characteristics of cancer and the relationships between different subtypes. These findings provide important evidence for early prediction, diagnosis, and treatment of cancer.

2. LIMITATIONS OF TRADITIONAL CLUSTERING ALGORITHMS

In traditional clustering algorithms, the limitations of metric methods can result in less accurate clustering results. For instance, Euclidean distance is often used to measure the similarity between samples. However, for high-dimensional data, the use of Euclidean distance may lead to high computational complexity and overlook non-linear correlations within the data. Therefore, to overcome the limitations of metric methods, more advanced and complex clustering algorithms are required to achieve more accurate and effective clustering results.

3. IMPROVEMENTS BASED ON ENERGY DISTANCE CLUSTERING ALGORITHM

3.1. Definition of energy distance

Energy distance is a generalized distance metric that better represents the distance between data points by considering the distribution characteristics of points in the manifold space. In clustering analysis, traditional Euclidean distance is a simple linear distance metric that fails to fully consider the nonlinear relationships between data points. This may result in clustering centers being biased towards a certain direction in linear space, while ignoring important manifold information.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

Table 1. The limitations analysis of clustering algorithm

Method of Application	Measuring Method	Limitations
K-means	Euclid Distance	1.The determination of the radius Eps and MinPts is required. 2.Difficulty in parameter selection when dealing with uneven spatial clustering densities. 3.Significant increase in computational complexity with larger datasets.
Mean-shift	Euclid Distance	1.Sensitivity of algorithm classification to noise points (outliers). 2.Inability to identify complex, non-spherical clusters with high mixing degrees.
DBSCAN	Euclid Distance/ Manhattan Distance	1.The determination of the radius Eps and MinPts is required. 2.Difficulty in parameter selection when dealing with uneven spatial clustering densities. 3.Significant increase in computational complexity with larger datasets.
HDBSCAN	Euclid Distance/ Manhattan Distance	1.Difficulty in MinPts and Eps selection when spatial clustering densities are uneven and cluster separations vary significantly. 2.Longer convergence time for clustering when the dataset is large.
Birch	Euclid Distance/ Manhattan Distance	1.CF Tree algorithm's restriction on the number of CFs per node may lead to clustering results deviating from the true class distribution.[15] 2.Poor performance of clustering algorithms when the distribution of the dataset is not similar to a hypersphere.

Let X and Y be two d-dimensional random variables that are mutually independent. The energy distance between X and Y is defined as follows:[16]

$$\varepsilon(X,Y) = 2E|X-Y|_d - E|X-X'|_d - E|Y-Y'|_d$$

Here, $E|X|_d < \infty, E|Y|_d < \infty$, X' and Y' are independently and identically distributed from X, Y ,respectively.

Energy distance is more effective in capturing the intrinsic characteristics of each data point, and its calculation method is better suited for handling high-dimensional data, thus avoiding the common problem of dimensionality curse in high-dimensional data processing. [17] Therefore, the application of energy distance in clustering algorithms has significant advantages and potential. In the current era of big data and high-dimensional data processing, using energy distance can enhance the accuracy and feasibility of clustering algorithms, further promoting the development and application of clustering algorithms in the field of data mining.

3.2. Ideas for improving distance algorithms

Based on the analysis above, one major advantage of energy distance compared to common systematic clustering algorithms is its ability to distinguish classes with small mean differences and describe the inherent similarities between data points more effectively. Therefore, the improvement of traditional clustering algorithms can be approached as shown in the following table:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

Table 2. Improved clustering algorithm based on energy distance

Traditional Distance-based Clustering Algorithm	Measuring Method	Improved Measuring Method	Improved Clustering Algorithm Based on Energy Distance
K-means	Euclid Distance	Energy distance	E_K-means
Mean-shift	Euclid Distance	Energy distance	E_Mean-shift
DBSCAN	Euclid Distance/ Manhattan Distance	Energy distance	E_DBSCAN
HDBSCAN	Euclid Distance/ Manhattan Distance	Energy distance	E_HDBSCAN
Birch	Euclid Distance/ Manhattan Distance	Energy distance	E_Birch

4. EMPIRICAL ANALYSIS

4.1. Experimental results and analysis based on gene expression data

Based on the research above, the improved clustering algorithm is more suitable for high-dimensional and large-scale samples. In modern biological sciences, the analysis of gene expression data is a crucial research area, characterized by high dimensionality, sparsity, and noise. Therefore, this study aims to validate whether the energy distance-based clustering algorithm can handle numerical data clustering problems more efficiently and accurately, specifically focusing on gene expression data.

4.1.1 Acquisition and preprocessing of gene expression data

This study obtained RNA-seq transcriptome data along with corresponding clinical and prognostic information of 1217 breast cancer patients from the TCGA database (https://cancergenome.nih.gov/). The study was conducted in accordance with the Helsinki Declaration (revised in 2013). Preprocessing of gene expression data was performed using outlier detection for denoising, direct deletion of missing values, data normalization, and principal component analysis for dimensionality reduction.

4.1.2 Clustering experimental results of gene expression data

This study applies the improved clustering algorithm before and after the improvement to breast cancer gene expression data, for different sample sizes. The data is clustered into three categories based on cancer stages (early stage (Ⅰ), mid-stage (Ⅱ), late stage (Ⅲ)). The consistency of the clustering results is measured using the Adjusted Rand Index (ARI), as shown in the following table:

Table 3. ARI comparison table of clustering results for cancer gene expression data before and after the improvement of clustering algorithm

Sample Size	10	30	50	100	300	600	900	ARI Mean	ARI Variance
K-means	0.418	0.734	0.812	0.907	0.942	0.953	0.962	0.818	0.038
E_K_means	0.982	0.986	0.990	0.994	0.991	0.993	0.993	0.990	1.91E-05
Mean-shift	0.456	0.598	0.648	0.778	0.862	0.890	0.891	0.732	0.028
E_Mean-shift	0.945	0.966	0.975	0.984	0.992	0.994	0.992	0.978	0.000
DBSCAN	0.567	0.666	0.698	0.756	0.846	0.903	0.914	0.764	0.017
E_DBSCAN	0.889	0.967	0.966	0.983	0.994	0.998	0.996	0.970	0.001
BIRCH	0.430	0.732	0.802	0.869	0.916	0.940	0.947	0.805	0.033
E_BIRCH	0.982	0.986	0.991	0.996	0.993	0.991	0.992	0.990	2.18E-05
HDBSCAN	0.567	0.666	0.698	0.756	0.846	0.903	0.914	0.764	0.017
E_HDBSCAN	0.874	0.966	0.966	0.982	0.992	0.998	0.995	0.968	0.002

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

From a horizontal perspective, both the traditional clustering algorithms and the improved ones show an increase in clustering accuracy as the sample size increases. Furthermore, the clustering effectiveness of the energy distance-based algorithm is generally superior to that of traditional clustering algorithms. From a vertical perspective, E_BIRCH and E_kmeans are more suitable for small-sample gene expression data, while E_DBSCAN is more suitable for large-sample gene expression data.

In addition, when examining the mean and variance of ARI, the improved clustering algorithms consistently achieve ARI values above 0.96 for gene expression data clustering, which is significantly better than traditional clustering algorithms. Moreover, the variance is controlled below 0.2%, indicating better stability compared to traditional clustering algorithms.

To provide a clearer observation of the changes in accuracy before and after each clustering algorithm improvement, this study presented a corresponding lines graph comparing the accuracies, as shown below:

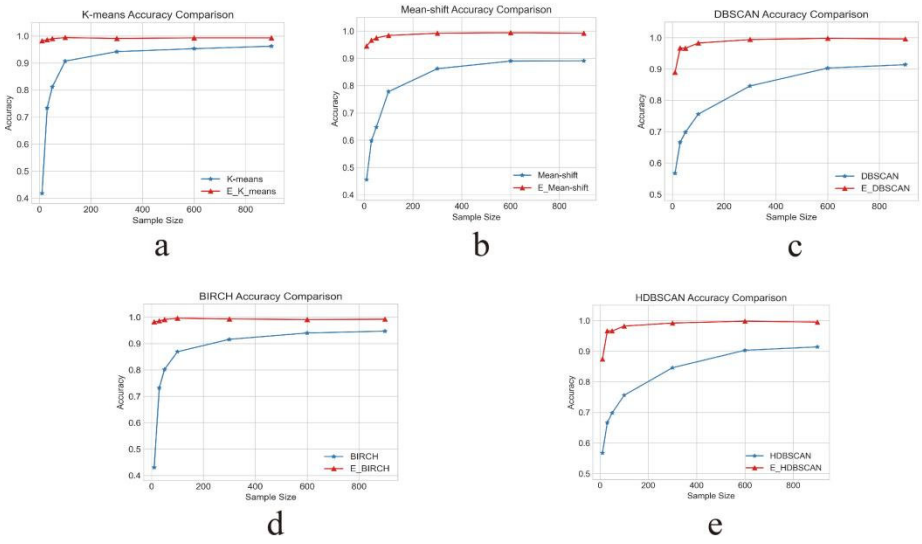

Figure 1. Accuracy Comparison Line Chart of clustering results for cancer gene expression data before and after the improvement of clustering algorithm

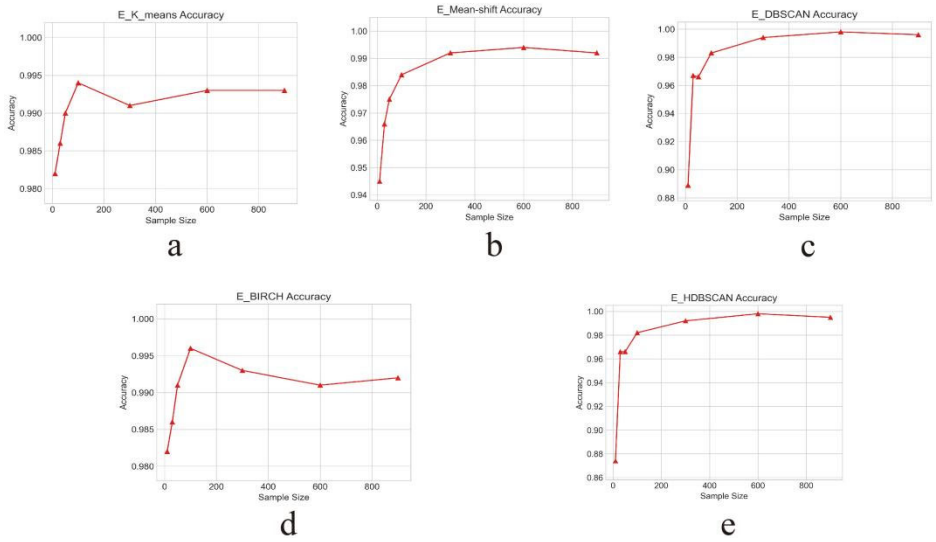

Figure 2. Accuracy Line Chart of clustering results for cancer gene expression data after the improvement of clustering algorithm

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

Figure 1 shows the accuracy comparison of the five clustering algorithms before and after improvement. It can be observed that all five energy distance-based clustering algorithms outperform traditional clustering algorithms, regardless of sample size.Figure 2 presents the accuracy comparison of the five improved clustering algorithms under different sample sizes. It can be seen that E_BIRCH and E_Kmeans algorithms are more suitable for small sample data, while E_DBSCAN performs the best in large sample data.

4.2. Experimental results and analysis based on gene survival data

Unlike typical numerical data, gene survival data is not numeric but non-numerical nominal data.[18,19] Therefore, effective handling of this data requires different methods. Hence, this study utilizes the improved algorithms to analyze gene survival data and employs clustering analysis to uncover the relationship between survival analysis and gene expression, thereby enhancing the depth and breadth of research in the field of gene survival analysis.

4.2.1 Acquisition and preprocessing of gene survival data

The breast cancer gene expression dataset was obtained from the UCI Machine Learning Repository, which includes gene data for 569 breast cancer patients along with their corresponding clinical and prognostic information. Each sample consists of 30 gene features. Each data point has 30 features, where one feature is the categorical label (M for malignant, B for benign), and the other features capture various biological characteristics of the tumor, such as size, shape, and texture.

Similarly, for the gene survival data, outlier elimination, direct deletion of missing values, and data normalization preprocessing were performed.

4.2.2 Clustering experimental results and analysis of gene survival data

Similarly, in this study, the breast cancer gene survival data was clustered using the improved clustering algorithms before and after improvement, based on different sample sizes. The data was clustered into two categories according to the survival data performance (M for malignant, B for benign), and the Adjusted Rand Index (ARI) was calculated as a measure of clustering consistency. The results are presented in the following table:

Table 4. ARI comparison table of clustering results for cancer gene survival data before and after the improvement of clustering algorithm

Sample Size	10	30	50	100	200	300	400	500	569	ARI Mean	ARI Variance
K-means	0.500	0.767	0.720	0.870	0.885	0.863	0.883	0.874	0.873	0.804	0.016
E_K-means	0.500	0.700	0.600	0.960	0.945	0.953	0.943	0.948	0.948	0.833	0.033
Mean-shift	0.769	0.900	0.920	0.900	0.925	0.927	0.928	0.928	0.922	0.902	0.003
E_Mean-shift	0.800	0.933	0.940	0.960	0.950	0.963	0.963	0.966	0.983	0.940	0.003
DBSCAN	0.600	0.630	0.700	0.760	0.870	0.920	0.910	0.910	0.930	0.803	0.018
E_DBSCAN	0.700	0.730	0.680	0.730	0.950	0.970	0.960	0.930	0.960	0.846	0.017
BIRCH	0.500	0.867	0.920	0.900	0.925	0.946	0.952	0.950	0.947	0.879	0.021
E_BIRCH	0.650	0.933	0.960	0.940	0.960	0.947	0.962	0.958	0.962	0.919	0.010
HDBSCAN	0.920	0.910	0.900	0.910	0.800	0.810	0.760	0.710	0.700	0.824	0.008
E_HDBSCAN	0.950	0.950	0.940	0.910	0.900	0.850	0.800	0.800	0.740	0.871	0.006

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

From a cross-sectional perspective, all clustering algorithms (except HDBSCAN) showed improvements in clustering accuracy with an increase in sample size, whether before or after improvement. Additionally, clustering algorithms based on energy distance are almost superior to traditional clustering algorithms.

From a longitudinal perspective, E_BIRCH and E_HDBSCAN were found to be more suitable for small sample gene expression data, while E_BIRCH and E_Meanshift were more suitable for large sample gene expression data.

Similarly, based on the mean and variance of the ARI, the improved clustering algorithms achieved ARI above 0.833 for gene survival data clustering, which is significantly better than traditional clustering algorithms. Furthermore, the variance was controlled below 0.3%, indicating greater stability compared to traditional clustering algorithms.

To visually observe the changes in accuracy before and after improvements for each clustering algorithm, this study presented a corresponding lines graph comparing the accuracies. The results are shown below:

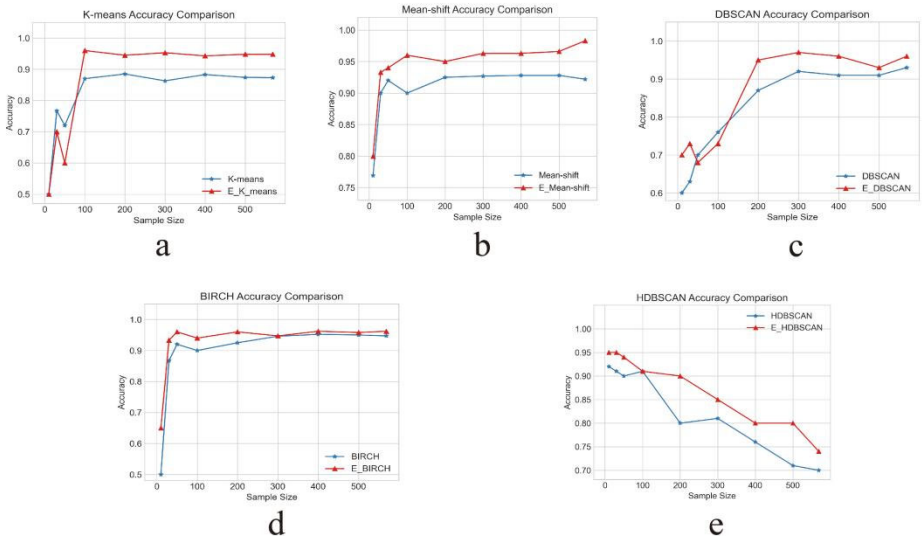

Figure 3. Accuracy Comparison Line Chart of clustering results for cancer gene survival data before and after the improvement of clustering algorithm

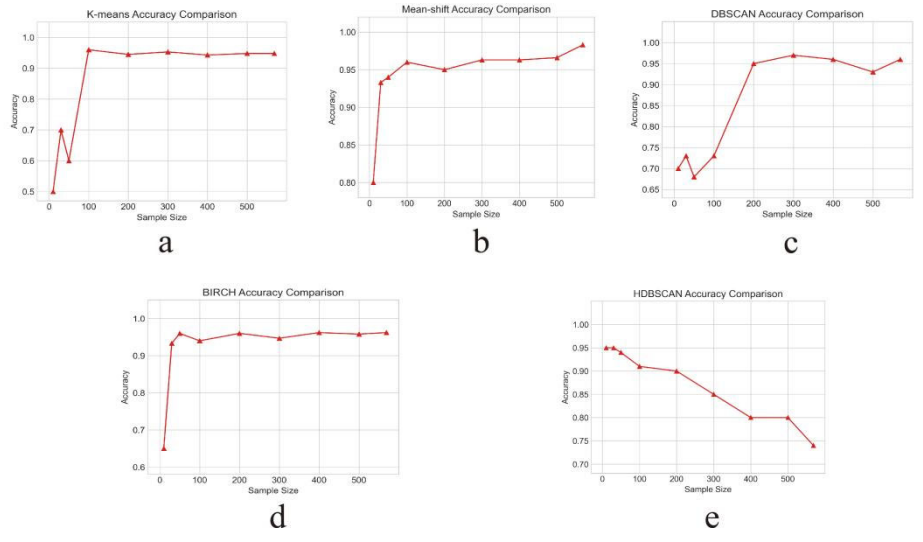

Figure 4. Accuracy Line Chart of clustering results for cancer gene survival data after the improvement of clustering algorithm

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

Figure 3 presents a comparison of the accuracy of the five clustering algorithms before and after improvement. It can be observed that E_Meanshift and E_BIRCH consistently outperform traditional clustering algorithms, regardless of sample size. Additionally, Figure 4 illustrates a comparison of the accuracy of the five improved clustering algorithms at different sample sizes. It can be seen that E_BIRCH and E_HDBSCAN are more suitable for small sample data, while E_Meanshift performs best in large sample data.

5. CONCLUSION

This study aimed to address the issues faced by traditional clustering algorithms (including k-means, Birch, DBSCAN, HDBSCAN, and mean-shift) when dealing with numerical and non-numerical data. It optimized these algorithms using an improved method based on energy distance and validated the effectiveness of the improved clustering algorithms using cancer data. The results showed that the improved clustering algorithms performed better for both numerical and non-numerical data.

Firstly, in numerical data, compared to traditional methods, the improved clustering algorithms were found to better preserve the relevant features among the data and differentiate between different data clusters more clearly. They also improved the accuracy and stability of clustering.

Secondly, in terms of non-numerical data, traditional clustering algorithms may not adapt well to the unique characteristics of text data. However, the improved energy distance algorithm has shown better performance in handling text data and has been widely applied in the field of data analysis. Its effectiveness is more pronounced compared to traditional algorithms.

In summary, the application of improved algorithms based on energy distance has shown positive effects in both numerical and non-numerical data clustering. There is still ample room for exploration and improvement in the future. For instance, this study incorporated both inter-class heterogeneity and intra-class homogeneity into the clustering objective function, aiming to achieve a unified approach. However, there still exists a certain error rate and probability of missing detections. Future research can consider introducing more precise and reliable feature selection algorithms for optimization. Furthermore, this study only applied the algorithm to cancer gene data analysis, leaving room for expansion and application to other types of gene data. By extending the application of the algorithm to a wider range of gene data analysis, we aim to better identify important gene features in bioinformatics, providing more reliable guidance and support for cancer treatment and prevention in clinical medicine.

ACKNOWLEDGMENTS

This paper was supported by China's College Student Innovation and Entrepreneurship Training Program (202310606001).

REFERENCES

[1] Sheng Hua and Zhang Guizhu: A Clustering Method Combining Kmeans and Fast Search Algorithm of Density Peaks, Computer Applications and Software, Vol. 33(2016) No.10, p.260-264.

[2] Zhang Jianpei, Yang Yue, Yang Jing, et al: Algorithm for Initialisation of K-Means Clustering Centre Based on Optimised-Division, Journal of System Simulation, Vol. 21(2009) No.9, p.2586-2590.

[3] Murtagh F and Contreras P: Algorithms for Hierarchical Clustering: an Overview, Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, Vol. 2(2012) No.1, p.86-97.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202311_9(11).0003

[4] Lior R: A Survey of Clustering Algorithms, Data Mining & Knowledge Discovery, Vol. 2(2012) No.8, p.269-298.

[5] Joerg S: *In Encyclopedia of Machine Learning* (Springer Press, USA 2011).

[6] Alias U F, Ahmad N B, Hasan S: Mining of e-learning behavior using SOM clustering, *Ict International Student Project Conference*(Skudai Malaysia IEEE Press, 2017). p.535-538.

[7] Liu Wenyuan, Wang Chunlei, Wang Baowen, et al: Application of Improved Locally Linear Embedding Algorithm in Dimensionality, Journal of Biomedical Engineering, Vol. 31(2014) No.1, p.85-90.

[8] Xuehong Wang: *The research on clustering algorithm applied to gene expression data*(MS., Shanghai Normal University, China 2013).

[9] LiXin Wei: *Research on cancer subtype clustering algorithm of gene expression profile data*(MS., Lanzhou JiaoTong University, China 2020).

[10] Zhou Wengang, Zhao Yu, Wang Feng, et al: A Cancer Gene Clustering Algorithm Based on Quantum-Behaved Particle Swarm with Comprehensive Learning Strategy, Journal of Beijing University of Posts and Telecommunications, Vol. 37(2014) No.4, p.59-63.

[11] Cai Xianfa, We Jia, Wen Guihua, et al: A Classification Method of Gene Expression Profile Based on a Locally Linear Embedding Algorism with Improved Distance, Journal of Biomedical Engineering, Vol. 28(2011) No.6, p.1213-1216.

[12] Alex R and Alessandro L: Clustering by Fast Search and Find of Density Peaks, Science, Vol. 344(2014) No.6191, p.1492-1496.

[13] Mehmood R, El-ashram S, Bie R, et al: Effective Cancer Subtyping by Employing Density Peaks Clustering by Using Gene Expression Microarray, Personal & Ubiquitous Computing, Vol. 22(2018) No.11, p.615-619.

[14] Wu P, Liu J L, Pei S M, et al: Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Renal Clear Cell Carcinoma, BMC Cancer, Vol. 18(2018) No.1, p.287-296.

[15] Shen Hua, Song Wei, Tang Chuanzhang, et al: BIRCH Clustering of Near-Surface Structural Characteristics, Geophysical Prospecting for Petroleum, Vol. 60(2021) No.2, p.283-294.

[16] Chen Xingrong and Yao Ningning: Research on Ward Clustering Method Based on Energy Distance Extending, Statistics and Decision, Vol. 22(2017) No.4, p.21-25.

[17] Tao Yonghui and Wang Yong: Improved K-means Algorithm Based on The Selection of Initial Clustering Centers, Foreign Electronic Measurement Technology, Vol. 41(2022) No.9, p.54-59.

[18] Peng Donghai and Zhang Liuwei: Testing The Distributional Difference of Two Multi-Dimensional Categorical Yariables Based on The Energy Distance, Journal of Hubei Normal University: Natural Science, Vol. 42(2022) No.3, p.10-17.

[19] Zhongbo Cao: *Application of improved biclustering method to cancer gene expression data*(MS., Jilin University, China 2009).