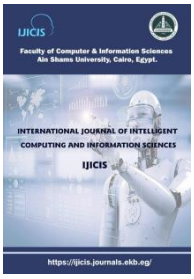
IJICIS, Vol.22, No.3, 138-151
DOI: 10.21608/ijicis.2022.146718.1197

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



A REVIEW OF CLUSTERING ALGORITHMS FOR DETERMINATION OF CANCER SIGNATURES

Hassan Sayed Ramadan*

Department of Information systems,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
hassanramadan@cis.asu.edu.eg

Khaled El-Bahnasy

Department of Information systems,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
khaled.bahnasy@cis.asu.edu.eg

Received 2022-06-24; Accepted 2022-07-02

Abstract: *Important information needed to comprehend the biological processes that happen in a specific organism, and for sure with a relevance to its environment. Gene expression data is responsible to hide that. We can improve our understanding of functional genomics, and this is possible if we understood the underlying trends in gene expression data. The difficulty of understanding and interpreting the resulting deluge of data is exacerbated by the complexity of biological networks. These issues need to be resolved, so clustering algorithms is used as a start for that. Also, they are needed in many files like the data mining. They can find the natural structures. They are able to extract the most effective patterns. It has been demonstrated that clustering gene expression data is effective for discovering the gene expression data's natural structure, comprehending cellular processes, gene functions, and cell subtypes, mining usable information from comprehending gene regulation, and noisy data. This review examines the various clustering algorithms that could be applied to the gene expression data, this is aiming to identify the signature genes of biological diseases, which is one the most significant applications of clustering techniques.*

Keywords: *Signature Genes, Clustering, Gene Expression Data, Prognosis, Biological Process*

1. Introduction

Various fields of study, including image analysis, pattern recognition, data mining, machine learning, and bioinformatics, have extensively used the clustering of an unsupervised learning technique. From noisy gene expression data, we can find meaningful information, and this is in order to develop new

* Corresponding author: Hassan Sayed Ramadan
Department of Information Systems, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.
E-mail address: hassanramadan@cis.asu.edu.eg

hypotheses. Grouping gene expression data into clusters, which differ from one another and have comparable expression patterns could be the start. Distance is a typical way to quantify similarity in data; if two or more genes are tightly related, and this relation depends on a given distance, they are objects of a specific cluster.
For specific experimental datasets, it remains challenging to find a good clustering method despite the availability of numerous clustering algorithms.
Samples are considered as the features in gene-based clustering. Also, genes are considered as the objects. Samples are viewed as objects in sample-based clustering. Also, genes are viewed as features, the samples can be divided into identical groups. The nature of the clustering tasks for gene expression data is considered as the difference between sample-based, and gene-based clustering [1].
Clustering can be partial or complete; whereas assigning each gene to a cluster is done by complete clustering, partial clustering does not. Frequently, there are some irrelevant genes or samples in the gene expression data, partial clustering tends to be more appropriate for gene expressions.
Because of representing noise by most often genes in the expression data, partial clustering in gene expression permits certain genes are not part of well-defined clusters, allowing their influence to be proportionately smaller on the conclusion. Additionally, it helps to overlook a significant number of unnecessary contributions by preventing some genes that are part of the expression data from becoming a part of clearly defined clusters.
Thus, by not requiring the participation of unrelated genes, partial clustering aids in preventing situations when an important subset in a cluster is kept [2]. Hard or overlapping clustering are two types of clustering [2]. While overlapping clusters assign each input gene varying degrees of membership in many clusters, each gene is assigned to a single cluster by hard clustering as an output, and also both during operation. If each gene got assigned to a cluster that have the highest degree of participation, that will convert an overlapped clustering into a hard clustering.
Identification of gene signatures is one the most significant application of clustering algorithms. This review aims to discuss many clustering algorithms, and demonstrate their challenges, and their applications. Also, it will focus on identifying the signature genes with various clustering algorithms.
The paper is structured as next. In Section 2, we discuss various types of clustering algorithms with details as advantages, and disadvantages of each algorithm. In Section 3, challenges of different clustering algorithms in different types of clustering are presented. The authors present various applications of many clustering algorithms in Section 4. In Section 5, various studies are discussed analyzing how signature genes are identified using different clustering algorithms. In Section 6, conclusions are drawn.

2. Types of Clustering Algorithms

Clustering is frequently used in a variety of academic disciplines, like image analysis, bioinformatics, pattern recognition, data mining, and machine learning. Depending on the type of dataset, clustering can be done using genes, samples, or time variables. To cluster both genes and samples, you shouldn't disregard the value of clustering in gene expression data of both.

2.1. Hierarchical clustering:

An algorithm called hierarchical clustering, commonly referred to as hierarchical cluster analysis, divides objects into clusters based on how similar they are. The result is a collection of clusters, each of which differs from the others while having things that are generally similar to one another.

Examples for Hierarchical methods:

- We have an algorithm that uses hierarchical agglomerative approach, it is Agglomerative nesting (AGNES) [3]. It is known as bottom-up approach. In data mining, it is a very popular hierarchical clustering algorithm. Based on the comparability of various objects, AGNES gather them using the concept of various leveled clustering.

  Advantages:
  - The number of clusters is not required.
  - Its usage, and implementation are very easy.

  Disadvantages:
  - It is not possible for taking a step back in AGNES.
  - It has a high time complexity. At least it is $O(n^2 \log n)$.

- The second method here, we have an algorithm that uses hierarchical divisive approach, it is Divisive Analysis (DIANA) [3]. It is known as a top-down approach. It starts with one cluster contains the whole data, so it needs a splitting method for that cluster, and it keeps splitting clusters recursively until having singleton clusters of individual data.

  Advantages:
  - Same as AGNES the number of clusters is not required.
  - It is more accurate, and efficient than AGNES.

  Disadvantages:
  - It is not possible for taking a step back in DIANA.
  - It is more complex comparing to AGNES, and it is hard to implement, because of the clustering method that is required to split clusters, which makes it less widely used.

- Also, the third method here, we have an algorithm that is a hierarchical clustering (HC) algorithm, which is CHAMELEON [4]. For the decision of the similarity among pairs of groups, it uses dynamic modeling. Also, for the existing models, it overcomes the limitations. For the clusters being merged, CHAMELEON depends on their internal characteristics, and it does require the supplied information of the user.

  Advantages:
  - The ease of clustering two dimensional, and three-Dimensional data sets.

  Disadvantages:
  - From the database corruptions, it cannot recover.

- Between all the data objects, CHAMELEON performs inter-closeness, and inter-connectivity, so it has a highly complex computation.

- We have an algorithm that the clustering feature's concept. It is Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [5]. It is one of the scalable clustering algorithms. Also, it is designed for only one scan of data, and can perform clustering over very large data sets.

Advantages:

- Only one single scan is enough to find a good clustering.
- It is heuristic that it doesn't need to scan all data points to make a decision.

Disadvantages:

- It can handle only numeric data.

2.2. Partitioning clustering:

The iterative relocation techniques are the most well-liked category of clustering algorithms that we have. These algorithms aim to achieve a (locally) optimal partition by repeatedly moving data points across clusters to minimize a particular clustering criterion. Examples for Partitioning methods:

- We have an algorithm that is considered as a hybridized K-means. It is K-means' extension. This method is with Cluster Centre Initialization Algorithm (CCIA). It identifies the closest pair of data points, then it is able to find the appropriate centroid.

Advantages:

- The number of clusters is not required that it can overcome this.

Disadvantages:

- It is biased to spherically shaped clusters, so it doesn't have a good performance for clustering gene expression data.
- It cannot handle both highly connected clusters, and high-dimensional dataset.

- Also, the second algorithm here is considered as K-means' extension too. It is Intelligent Kernel K-means (IKKM) [6]. It uses the benefit of kernel K-means, and intelligent K-means to avoid the K-means algorithm's drawbacks.

Advantages:

- The number of clusters is not required.
- It has a good performance for clustering gene expression data.

Disadvantages:

- It cannot handle the issue of high dimensionality.

- The third one is a partitional clustering technique, and it is called PAM. It is also called as K-Medoids. It can work efficiently with dissimilarity matrix as an input data.

  Advantages:
  - It is fast, easy to implement, and simple to understand.
  - Other partitioning algorithms are very sensitive to outliers, but it is less sensitive to them.

  Disadvantages:
  - For non-spherical groups, it is not suitable.
  - Because of choosing the first k medoids randomly, results are changed with different runs on the same dataset.

2.3. Model-based clustering:

This is a statistical method of clustering data. The observed (multivariate) data was produced from a finite mixture of component models. The probability distribution is considered as every part of the model, also the parametric multivariate distribution is usually. Examples for Model-based methods:

- We have an algorithm that is a Self-organizing maps (SOMs) [7]. In gene clustering, it is widely used. Also, based on neural network methods, it was developed. It is designed to have any size of dimensions.

  Advantages:
  - Large, and complex data sets can be clustered by SOMs.
  - Providing intelligible, interactive, and useful summary of the data, which helps solving various types of challenges.

  Disadvantages:
  - To develop meaningful clusters, data should be sufficient, and necessary.
  - It cannot perform the perfect mapping always, especially when groupings are unique.

- The second algorithm depends on model-based Bayesian clustering approach. It is an improved version of it. It is Chinese restaurant clustering (CRC) [8]. It checks each gene, and its probability to join the existing clusters.

  Advantages:
  - With high accuracy, it can derive the number of clusters, and also it can cluster genes simultaneously.
  - It is able to group strong correlated genes.

2.4. Soft clustering:

In this type of clustering, an item has the opportunity to exist in multiple clusters after grouping the data items. Examples for Soft clustering:

- Fuzzy Analysis (FANNY) [3]. Hard decisions are not used by FANNY. It does not use them to assign degree of membership to each element to cluster it.

  Advantages:
  - For each object, means of membership coefficients quantifies its relationship's degrees to different clusters, so FANNY does not force it to be clustered into a specific group.
  - For the spherical cluster assumption, it is more robust.

  Disadvantages:
  - It has complex computations, and it is not easy to implement.

- Fuzzy C-Means (FCM) Clustering Algorithm [9]. It is a widely used soft clustering approach that determines the degree of membership of each sample point by evaluating its weight.

  Advantages:
  - For overlapped data, and objects that have a high correlation, the performance of FCM is superior.
  - Its rate of convergence is high comparing to other algorithms.

  Disadvantages:
  - In case of the existence of outliers, and noise in data, FCM will have a poor performance.
  - It is not suitable for large dataset, and its process is very slow.
  - The number of clusters is required.

- The third algorithm is also a fuzzy clustering, but it depends on Local Approximation of MEmbership (FLAME) [10]. Among objects there are neighborhood relationships, which FLAME uses to cluster objects. Also. In the dataset’s dense parts, FLAME defines clusters.

  Advantages:
  - It identifies cluster outliers.
  - It can find nonglobular clusters, and nonlinear relationships.

2.5. Grid-based clustering:

It uses a specific type of data structure, which is the multi-resolution grid. To create a grid structure, there is a division of the object areas, and as a result we have a limited number of cells. In this case, all clustering procedures are carried out. Examples for Grid-based methods:

- The first algorithm is called STING, which stands for Statistical information grid-based algorithm. Operations for clustering are developed on the grid structure, which is formed based on a finite number of cells that are considered as the object areas, which are quantized by STING. Results of query are approximately expected because STING uses the statistical information.

  Advantages:
  - It has a low time complexity.
  - It is efficient for large datasets.

  Disadvantages:
  - Handling outliers is hard for high dimensional datasets.
  - According to the number of grids, the quality level of the clustering result is determined.

- The second algorithm, which also is a grid-based clustering algorithm, called OptiGrid. For each dimension, it calculates the best partitioning hyperplanes to determine the optimal grid-partitioning.

  Advantages:
  - It handles the drawback of dimensionality.
  - With high dimensionality large datasets, it is very efficient.
  - Strong to outliers.

  Disadvantages:
  - Three input parameters are required.
  - It is sensitive to parameters choice.

- CLIQUE (for Clustering In Quest) clustering technique [11]. CLIQUE uses the dimensionality reduction method principal component analysis, which get lower dimensional space after transforming the original dataset space optimally.

  Advantages:
  - It has interpretability of results, and considered as a simple method.

- For high dimensional datasets, it can find any number of clusters of any shape.
- In both accuracy, and execution time, it outperforms many other clustering algorithms.

Disadvantages:

- It cannot find the correct cluster always, especially in case of the cell's size cannot fit for a set of very high values.

2.6. Density-based clustering:

There are adjacent regions of low point density as a separation between a cluster in a data space, and others of the same kind, it distinguishes unique groups/clusters in the data. Examples for Density-based methods:

- The first approach is called DENCLUE, which stands for Density-based Clustering approach. The group of density distribution functions are used for DENCLUE the clustering approach. Through a two-phase process, and for a very large multidimensional dataset, it can find natural clusters.

  Advantages:

  - It can handle datasets that have large amounts of noise.
  - It is significantly faster than many other Density-based algorithms.

  Disadvantages:

  - It requires careful selection of the input parameters that the quality of clustering could be influenced significantly.

- The second algorithm is considered as a DBSCAN's extension. It is called MDBSCAN [12], which stands for Prototype-Based Modified DBSCAN. First, it generates k number of subclusters depending on any squared error clustering algorithm by applying it on the input dataset.

  Advantages:

  - It handles noise in the dataset.
  - For arbitrary shapes, it performs clustering well.

  Disadvantages:

  - In case of the number of clusters is large, its performance is affected.

2.7. Multiobjective optimization methods:

- The first algorithm is a multiobjective clustering algorithm, it is called Multiobjective clustering (MOCK) [13]. Evaluation of the connectedness of items, and minimizing the deviation is the aim of MOCK to produce quality clusters.

  Advantages:
  - It has a low execution time.
  - It handles outliers.
  - Finding the optimal number of clusters.

- GenClust-MOO clustering algorithm [14]. It provides an initial spread of solutions, and this initial spread is good. This is based on the initialization procedure that does this spread, and it is partly random.

  Advantages:
  - With the presence of noise, it can extract the appropriate number of clusters.
  - It can detect the appropriate partitioning from datasets.
  - It has a superior computational performance.

- Mofuzzy is the third clustering algorithm, and it is a fuzzy multiobjective (MO) [15]. Its objective is to optimize its objective functions simultaneously.

  Advantages:
  - Automatic partitioning of various types of datasets, which have clusters in different sizes, and shapes.
  - Its ability to detect an appropriate number of clusters.

3. Challenges of Various Clustering Algorithms

The majority of clustering methods used today are based on distance. For gene expression data, there are many clustering algorithms, but the most popular clustering algorithms are K-means clustering, HC, and SOMs. The performances of these algorithms could be noise-sensitive [1], although these algorithms are visually appealing, and fairly straightforward.
Table 1 summarizes the challenges of the most popular clustering approaches.

Table 1 Challenges of Most Popular Clustering Approaches

Clustering Approach	Challenges
Hierarchical Clustering	- The algorithm is resistant to missing data, and its performance is noise-sensitive. Also, it finds it challenging to convey details like the required number of clusters and the confidence measures for each cluster. - Clustering bigger data can be challenging for HC. - Additionally, based on local decisions HC can group points into clusters due to its deterministic characteristics, and this is all without allowing for the possibility of reexamining the grouping. - There are limitations in the hierarchical agglomerative clustering that the patterns' structure is fixed according to a binary tree [16], which affects finding all the signature genes especially with large datasets. - With large datasets, reevaluating the results cannot be processed by HAC, because it is hard for it to interpret patterns [16]. In this case, based on local decisions are some clusters of patterns. This makes it not a candidate for identifying signature genes in large datasets.
Partition Clustering	- These clustering techniques' fundamental drawback is that they consistently generate subpar results because data points overlap whenever one point is close to another cluster's center. - The first seed point of preferred clusters is chosen at random by the K-Means method; - The K-means clustering's results are particularly susceptible to noise and outliers. - The convergence to a local minimum is guaranteed by K-means. the global one is not guaranteed. Being an iterative algorithm, in each run K-means does the convergence to a different local minimum. In many applications, this is acceptable, but the results of extracting cancer signatures are not even close [17].
Model-based Clustering	- The dataset has the opportunity to fit a particular distribution may occasionally be made by model-based clustering algorithms. - SOM is frequently used to cluster gene expression data; however, attempts to combine various patterns into a cluster may render SOM unproductive [1]. To distinguish clear clustering limits from the SOM result every time it provides an unstable solution is a difficult challenge. - SOM is not effective when different patterns are merged into a cluster [16]. This is common in early stages of the process of identifying the signatures, so in this case SOM is not a candidate for identifying signature genes.
Grid Clustering	- Grid clustering techniques require a certain minimum value for the grid's size. - For high-dimensional datasets, most of Grid-based clustering algorithms are not suitable. Because of large number of genes [16], which are considered as the main dimensions of the biological datasets that makes identifying signature genes is not an application that could be processed using most of Grid-based clustering algorithms.
Density-Based Clustering	- When the data structure gets complex, and the datasets are quite huge, these approaches make it very challenging to determine the attraction tree's structure [1]. - The Euclidean distance, which is used to determine object closeness when the data is high dimensional, likewise does not work well. - DBSCAN is a very efficient algorithm to identify signature genes that it is robust against noise, also works fine for identifying clusters of arbitrary shapes, however it strongly hinges on choosing the right parameters.

4. Applications of Various Clustering Algorithms

Clustering is the process of grouping items in groups, and this process depends on criteria for similarity, and therefore the items in one group are remarkably similar to one another and stand out from those in other groups. Gene patterns have been discovered using a variety of clustering algorithms, making it simpler to identify genes with related functions.

Using clustering to find unknown disease-fighting pathways is another potential application. Genes that are the main targets of pathogen attack can be extracted by clustering gene expression data, providing chemists with a clear direction for therapeutic development.

In order to determine for studying lung cancer, which type of algorithm and dataset would be most effective, in 2015 the datasets of lung cancer were examined. There is a format called ARFF stands for Attribute Relation File Format, and the dataset of lung cancer with ARFF was found to be efficiently clustered using the K-Means algorithm [18].

To ascertain the invasive breast cancer's stage, Karmilasari et al. have used images from MIAS Which stands Mammography Image Analysis Society, and applied the K-means algorithm to these images.

From the genome-wide transcriptional patterns, Bochkov et al. [19] sets of differentially expressed genes were identified by using HC, and they were relevant to epithelial repair and inflammatory mechanisms, which obviously distinguished the normal and asthma groups.

In order to look into the mechanisms governing the genes implicated in parasite transmission, Heard et al. grouping genes has been done using the Bayesian model-based HC algorithm, and this was with comparable expression.

The identification of potential genetic markers that would reveal efficient genes in the treatment of HIV/AIDS and also pattern-based clustering was employed to do the regulation, HC, and research by Raman and Domeniconi. We will explain more studies about the signature genes in the next section.

5. Analysis and Discussion of Clustering Algorithm usage in Identifying Signature genes

A single gene or group of genes in a cell with a distinctively distinctive pattern of gene expression is known as a gene signature or gene expression signature. Clustering analysis of gene expression data identifies these signatures. Arora et al. [20] discovering cancer subtypes that are not only prognostically significant but also molecularly distinct, and this was done using the survClust algorithm. This algorithm was used as a supervised learning approach to in order to get around the existing restriction of molecular clustering analysis.

It is a multidimensional omics-profiling data-based for survival stratification using outcome-weighted integrative clustering approach. In order to down weight molecular features that are irrelevant to the desired outcome, the algorithm learns a weighted distance matrix.

The Cancer Genome Atlas study was used to extract data that they examined more than 6000 tumors, across 18 different cancer types, using six different molecular data types, including the integration of the six data modalities, protein expression, miRNA expression, mRNA expression, DNA methylation,

DNA copy number, and somatic point mutations. The outcomes have discovered prognostic molecular subgroups that unsupervised clustering had not before identified.

Zhu et al. [21] in order to distinguish between tumor microenvironment (TME) and drug sensitivity, they created a "Signature associated with FOLFIRI resistant and Microenvironment" (SFM). K-means clustering was also used to identify SFM subtypes. Since the K-means clustering algorithm is one of the most significant and straightforward clustering approaches. Also, this algorithm is statistically deterministic without defining seed centers, they employed it to accomplish the clustering. Assuming k clusters, it is a simpler method to categories the dataset. In case that the clusters' number is relatively small, there is a benefit for the K-means algorithm, a benefit of faster calculation for big variables. They used the "factoextra" R package's K-means clustering implementation to analyze gene expression profiles. This analysis depends on about 250 distinct genes were used to create the SFM signature. The value that maximizes the gap statistics is used to estimate the ideal clusters, indicating that the clustering structure is not at all similar to the random uniform distribution of points.

Xu et al. [22] Consensus clustering was used to identify BRCA subtypes. The ICI pattern was used as a base for this identification. It was used to identify three distinct ICI pattern subtypes. The "ConsensusClusterPlus" package was used to divide the BRCA samples into three distinct subgroups, this was done using a hierarchical agglomerative consensus, and this algorithm depends on the ICI matrix, this was done with a view to more effectively show the immune cell's biological importance infiltration in BRCA. They employed the unsupervised clustering "Pam" method based on the linkage of Euclidean and Ward. Also, to guarantee the stability of the classification, it was repeated 1,000 times. Samples were grouped into ICI subgroups, which depends on the prior consensus clustering algorithm with a view to identify the genes associated with the ICI patterns.

When Mycobacterium tuberculosis was profiled in 2002 using HC, 826 genes were found to have poor expression in most of replicates of the hybridizations, which are stationary-phase, and logarithmic [23]. As noticed everyday there are new clustering algorithms that add value to the field of gene expression data analysis, and identifying signature genes.

Wang et al. [24] The prognostic lncRNA expression was used through non-negative matrix factorization (NMF), which was used in order to explore molecular subgroups. Based on the cophenetic correlation coefficient, the optimal K cluster was selected. Filtering out low expression level lncRNAs, and this to ensure a reliable, and robust subtype. Also, associated lncRNAs with survival of GC (Gastric cancer) were retained. NMF clustering analysis processed data, and the optimal k value was determined using the cophenetic correlation coefficients. As a result, four subtypes were identified named C1, C2, C3, and C4.

Li et al. [25] Oncogenic, DNA damage repair, stromal, or immune pathways were considered as enrichment levels of 15 pathways. Based on them GCs were clustered in three datasets (GSE84437, ACRG-STAD, and TCGA-STAD) using the consensus clustering algorithm. The clustering results of these three datasets were displayed similar. Immunity-enriched (ImE), stroma-enriched (StE), and immunity-deprived (ImD) are considered as the three subtypes of GCs after being clustered. This was confirmed with principal component analysis that in all three datasets, and based on the pathway scores of them, separating GCs could be into three subgroups.

6. Conclusion

In order to comprehend the biological processes that occur in a certain organism in connection to its environment, it is essential to understand the information that is hidden in gene expression data. These facts prevent ambiguity, noise, and imprecision. Researchers have created numerous clustering algorithms to glean important data regarding gene activity in relation to various systemic conditions. This review looks at popular clustering validity techniques and finds that exceedingly of them are biased toward one type of clustering methodology. The higher validity rating is what needed, and required to give them, and also it results in a false impression of the clustering output. For instance, the existing functional annotations could be used in order to confirm whether our goal of clustering genes with similar functions has been achieved. This review discussed many studies about various clustering algorithms focusing on the identification of signature genes, and the prognosis.

References

1. Daxin Jiang, Chun Tang, and Aidong Zhang, “Cluster analysis for Gene Expression Data: A survey,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370–1386, 2004.
2. G. Kerr, H. J. Ruskin, M. Crane, and P. Doolan, “Techniques for clustering gene expression data,” Computers in Biology and Medicine, vol. 38, no. 3, pp. 283–293, 2008.
3. J. E. Gentle, L. Kaufman, and P. J. Rousseuw, “Finding groups in data: An introduction to cluster analysis.,” Biometrics, vol. 47, no. 2, p. 788, 1991.
4. G. Karypis, Eui-Hong Han, and V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” Computer, vol. 32, no. 8, pp. 68–75, 1999.
5. T. Zhang, R. Ramakrishnan, and M. Livny, “Birch,” ACM SIGMOD Record, vol. 25, no. 2, pp. 103–114, 1996.
6. T. Handhayani and L. Hiryanto, “Intelligent kernel K-means for clustering gene expression,” Procedia Computer Science, vol. 59, pp. 171–177, 2015.
7. T. Kohonen, “The self-organizing map,” Proceedings of the IEEE, vol. 78, no. 9, pp. 1464–1480, 1990.
8. Z. S. Qin, “Clustering microarray gene expression data using weighted Chinese restaurant process,” Bioinformatics, vol. 22, no. 16, pp. 1988–1997, 2006.
9. J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy C-means clustering algorithm,” Computers & Geosciences, vol. 10, no. 2-3, pp. 191–203, 1984.
10. L. Fu and E. Medico, “Flame, a novel Fuzzy Clustering Method for the analysis of DNA microarray data,” BMC Bioinformatics, vol. 8, no. 1, 2007.
11. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications,” ACM SIGMOD Record, vol. 27, no. 2, pp. 94–105, 1998.
12. D. R. Edla and P. K. Jana, “A prototype-based modified DBSCAN for gene clustering,” Procedia Technology, vol. 6, pp. 485–492, 2012.

13. J. Handl and J. Knowles, “An evolutionary approach to multiobjective clustering,” IEEE Transactions on Evolutionary Computation, vol. 11, no. 1, pp. 56–76, 2007.
14. S. Saha and S. Bandyopadhyay, “A generalized automatic clustering algorithm in a multiobjective framework,” Applied Soft Computing, vol. 13, no. 1, pp. 89–108, 2013.
15. S. Saha, A. Ekbal, K. Gupta, and S. Bandyopadhyay, “Gene expression data clustering using a multiobjective symmetry based clustering technique,” Computers in Biology and Medicine, vol. 43, no. 11, pp. 1965–1977, 2013.
16. J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi, “Clustering algorithms: Their application to gene expression data,” Bioinformatics and Biology Insights, vol. 10, 2016.
17. Z. Kakushadze and W. Yu, “*k-means and cluster models for cancer signatures,” Biomolecular Detection and Quantification, vol. 13, pp. 7–31, 2017.
18. A. Dharmarajan and T. Velmurugan, “Lung cancer data analysis by K-means and farthest first clustering algorithms,” Indian Journal of Science and Technology, vol. 8, no. 15, 2015.
19. Y. A. Bochkov, K. M. Hanson, S. Keles, R. A. Brockman-Schneider, N. N. Jarjour, and J. E. Gern, “Rhinovirus-induced modulation of gene expression in bronchial epithelial cells from subjects with asthma,” Mucosal Immunology, vol. 3, no. 1, pp. 69–80, 2009.
20. A. Arora, A. B. Olshen, V. E. Seshan, and R. Shen, “Pan-cancer identification of clinically relevant genomic subtypes using outcome-weighted integrative clustering,” Genome Medicine, vol. 12, no. 1, 2020.
21. X. Zhu, X. Tian, L. Ji, X. Zhang, Y. Cao, C. Shen, Y. Hu, J. W. Wong, J.-Y. Fang, J. Hong, and H. Chen, “A tumor microenvironment-specific gene expression signature predicts chemotherapy resistance in colorectal cancer patients,” npj Precision Oncology, vol. 5, no. 1, 2021.
22. Q. Xu, S. Chen, Y. Hu, and W. Huang, “Landscape of immune microenvironment under immune cell infiltration pattern in breast cancer,” Frontiers in Immunology, vol. 12, 2021.
23. A. M. Talaat, “Genomic DNA standards for gene expression profiling in mycobacterium tuberculosis,” Nucleic Acids Research, vol. 30, no. 20, 2002.
24. H. Wang, Q. Meng, and B. Ma, “Characterization of the prognostic M6A-related lncrna signature in Gastric cancer,” Frontiers in Oncology, vol. 11, 2021.
25. L. Li and X. Wang, “Identification of gastric cancer subtypes based on pathway clustering,” npj Precision Oncology, vol. 5, no. 1, 2021.