Cancer Gene Clustering Using Computational Model

Shivam Vishwakarma[1], Satyendra Kumar Bhardwaj[1], Anand Bihari[2*],
Sudhakar Tripathi[1], Sanjay Agrawal[3], and Puneet Joshi[3]

ARTICLE INFO

Article history:
Received: 02 October 2022
Revised: 23 November 2022
Accepted: 11 December 2022

Keywords:
Cancer data
Gene expression
Dataset
Clustering
Bioinformatics
Gene-expression data

ABSTRACT

Clustering plays an important role in the analysis of genetic datasets. In the gene dataset, there is a lot amount of hidden information. Understanding functional genomics needs to be strengthened to explore hidden information in gene expression data. Many challenges arise when dealing with the gene dataset. These challenges are the huge amount of data, dimensionality, and dataset changes over time. This kind of problem can be solved with the help of clustering algorithms. Therefore, clustering techniques are the first step in solving these challenges and are needed for data mining processes to uncover gene dataset's structure and hidden patterns. Many clustering algorithms analyze gene expression datasets. In this research, the cancer gene database has been clustered using five clustering algorithms, including K-Means Clustering, Hierarchical Clustering, SOM (Self-Organizing Map), and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Various external and internal clustering evaluation indices determine the clustering effectiveness of all clustering techniques and Calinski-Harabaz Index value is good for all the clustering techniques'. The outcome of the k-means clustering algorithm is best as compare to other clustering techniques.

1. INTRODUCTION

A variety of Genes are small chromosomal segments that perform additional functions such as storing encrypted data for protein synthesis. These sequences have been found in chromosomes of varying lengths, and some of them share specific functions.. The analysis of DNA sequences is a critical application area in computational biology, and discovering similarities between genes and DNA sub-sequences provides critical information about their functions and structures [2]. To find similarities between biological sequences, clustering, a popular data mining technique has been applied [3]. For instance, by grouping genes, the functions of those genes can be inferred based on the known functions of genes in other groups [4]. Several common pattern recognition methods, including k-means, k-nearest neighbors, and neural networks, can be used to cluster sequential data.

However, when observations include sequences with varying lengths, such as genes, these methods become extremely complex. Gene expression levels were measured simultaneously using DNA microarray technology [1], and the analysis of genetic data revealed the functional activity of genes. But in this technology, some useful information is lost, and then the clustering algorithms help in the analysis and find the missing information. The clustering groups the similar objects. The DNA microarray experiment generated the data in the form of matrix $G=\{Epq \mid 1<= p<= K; 1 <= q <=L\}$. The total number of rows is represented by K, which is the total number of genes. L represents the total number of columns, that is experimental condition. To identify the cluster, different clustering algorithms have been used. Several clustering algorithms are available to cluster the data. For the gene clustering authors are generally applied the K-Means Clustering [2], SOM [5-8], DBSCAN [3-4], and Hierarchical Clustering [9-11] and implemented on cancer gene expression datasets [12, 24, 25, 27, 28,29]. The implemented clustering has been validated by several internal and external validation techniques.

In this paper, section 2 highlights the several clustering techniques including k-menas, hierarchical clustering, SOM and DBSCAN and the important external and internal validation techniques. Section 3 discussed about the datasets and section 4 highlights and implementations and results. Finally section 5 highlights the importance of this research.

[1]Department of Information Technology, Rajkiya Engineering College Ambedkarnagar, U. P., India
[2]Department of Computational Intelligence, School of Computer Sciene and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
[3]Department of Electrical Engineering, Rajkiya Engineering College Ambedkarnagar, U. P., India
**Corresponding author:* Anand Bihari, Email: anand.bihari@vit.ac.in, csanandk@gmail.com

2. METHODOLOGY

This section will establish the groundwork concepts and background knowledge necessary for the subsequent discussions. This section highlights the important algorithms and the validation approach related to the clustering of genes.

2.1. Clustering Algorithms

K-Means: It is coms under the unsupervised machine learning algorithm. It is based on centroid-based techniques and uses the partition method for clustering. Here K represents the number of clusters, which is initialized randomly.

The appropriate value of K is decided by the Elbow graph by selecting the knee value (where abrupt change occurs) or based on the previous experience. The elbow graph is drawn between the WCSS (within-cluster sum of square) and the number of clusters. The optimal number of clusters is determined by selecting the value under the middle region and the reasonable number of nodes.

This algorithm clusters the data into k disjoint subsets that optimize the objective function, which is given as:

$$E = \sum_{g=1}^{n} \sum_{ob \in Cg} |ob - \mu|^2 \qquad (1)$$

where, ‘ob’ represents the object in cluster Cg and μ is the Centroid or mean of the objects in Cg. If the initialized centroid is very close, then the number of the cluster may be increased, i.e. incorrect. So initialized centroid should be very far.

Hierarchical Clustering: Hierarchical clustering is based on the unsupervised machine learning algorithm. This is a simple and proven method to analyze gene expression data by generating gene clusters with similar expression patterns. In the Hierarchical clustering algorithm, the clustering has been done hierarchically and represented like a tree is called a dendrogram [13]. For finding the number of clusters in hierarchical clustering, find the longest vertical line that has no horizontal line passing through it. It is very useful for the small dataset. It takes less time than the K-Means clustering algorithm. it can be used to group data in agglomerative (bottom-up) or divisive (top-down) ways.

DBSCAN: DBSCAN is a clustering algorithm that uses the density approach with noise. It finds clusters of varying shapes and sizes in enormous data volumes containing noise and outliers. DBSCAN takes two parameters:

1.eps: Distance measuring of a point to the neighborhood of any point.

2. MinPts: MinPts is the minimum number of neighbors (data points) inside the eps radius. The value of the MinPts should be determined based on the data size. The minimal MinPts are MinPts >= dimensions +1 in the number of dimensions in the data collection. The minimum MinPts value must be set to 3 or greater.

SOM (Self-Organizing Map): A self-Organizing Map is a clustering technique. T. Kohonen developed it in 1977. It is a trained model without any supervision. Hence SOM is unsupervised learning. SOM learns on their own through unsupervised competitive learning. SOM is a neural network that maps multi-dimensional data into lower-dimensional data. SOM has two layers. One is the input, and the other is the output. The input and output in this clustering technique are in a two-dimensional grid. The closest vector is used to map the input and output neurons. It takes a high-dimensional input and produces a low-dimensional output, and each input-data acts as training data during the computation process. The cluster was discovered after data training by mapping all data to possible output neurons.

2.2 Clustering validation index

Silhouette Index: This metric is used to assess the effectiveness of clustering techniques. The values of the silhouette index range from -1 to 1 [14]. If the silhouette index is negative, it indicates that the sample was assigned to the incorrect cluster, while 1 indicates that the clusters are well separated and distinct.
A silhouette plot shows how close each point in a cluster is to another point in the same cluster. The silhouette index has the following mathematical equation:

$$Sl(k) = \frac{(A_k - B_k)}{\max(A_k, B_k)} \qquad (2)$$

where, Sl(k) represents the silhouette coefficient of the kth data point. A_K represents the minimum average distance between the k^{th} data point and all other data points in the nearby cluster. B_k represents the average distance between the nth data point and all other data points in the same cluster.

Davies-Bouldin Index (DBI): It is defined as the ratio of intra-cluster (inside the cluster) to inter-cluster (between the clusters) distance [15]. It is used to find the centroid of the cluster. The Davies-Bouldin index ranges from 0 to 1. The best results are obtained when the value is close to zero.

Calinski-Harabaz Index (CHI): It is the ratio of the inter-cluster (with-in the cluster) dispersion sum to the intra-cluster (between the clusters) dispersion sum for all clusters [16]. Inter cluster dispersion means the sum of the with-in cluster dispersion for all the clusters and Intra cluster means the weighted sum of the between the cluster dispersion. It is known as the variation ratio criterion. In the Calinski-Harabaz index, good clusters have a large value of intra-cluster variance and small inter-cluster variance.

Adjusted Mutual Information (AMI): This is the coordination of mutual information and is used for the similarities between the two data labels [17]. Adjusted Mutual information is used for the unbalanced ground truth clustering and there should be a small cluster. Higher the adjusted mutual information value shows the purity of the cluster. It takes 0 when the mutual information between two partitions is equal and it takes 1 when both partitions are the same.

Normalized Mutual Information (NMI): This is the normalization of the mutual information score. It is used for scaling the result between 0 and 1 [18]. The value of 0 means no mutual information or the sets are dissimilar and 1 means perfect correlation or the sets are identical. It is used for comparing two partitions whether a different number of clusters. It tells about the uncertainty of the class labels. It is the same as the information gained in the decision tree.

Fowlkes mallows Index (FMI): It is used for finding the similarities between a clustering and benchmark classification or between two hierarchical clustering or among different clustering algorithms [19]. It is promotional to the number of true positives. The higher the value of the Fowlkes-Mallows index, the greater will be the similarities. For the unrelated data sets its value will be 0. The mathematical formula for the Fowlkes-Mallows index is given as follows:

$$FM = \sqrt{PPV * TPR} = \sqrt{\frac{TP}{TP+FP}} + \frac{TP}{TP+FN} \qquad (3)$$

where,

TPR = true positive rate

PPV= positive predictive norm.

TP = True Positive,

FP = False Positive, and

FN = False Negative.

Homogeneity Score: A cluster will be homogeneous if the data of all the clusters belong to the member of a single class [20]. It ranges from 0.0 to 1.0 where 1.0 shows the homogeneous labeling. In this technique, the permutation of the label of the cluster never changes. Hence it is free from label value in any way.

Completeness: It is the complement of the homogeneity score [21]. It is used for providing information about the sample of the same class. The output of the clustering will be completeness if all the data points that are a member of the assigned class are the element of that cluster. In this method, the permutation of the cluster's label never changes. So it is free from the data label's value in any way.

Rand Index: It is used for measuring the similarity between two clustering algorithms [22]. Mathematically, Rand indexes represent precision (accuracy) and are applied when class labels are not used.

The mathematical equation for calculating the Rand index is given below:

$$R = \frac{a+b}{N_{C_2}} \qquad (4)$$

Where, a denotes the number of times a pair of data points belong to the same cluster, b denotes the number of times a pair of data points belong to a different cluster and NC2indicates the number of unordered pairs in a set of n data points. It ranges from 0 to 1. The value of rand index zero means the two clustering algorithms do not perfectly match and 1 show that the two clustering algorithms perfectly matched.

Adjusted Rand Index: This is the rand index adjustment [23]. It is used for measuring the agreement between two partitions. It ranges from 0 to 1. If the value of the Adjusted Rand index is 0 means clustering is not good and 1 means clustering is identical and the data are well clustered.

3. DATASET

Gene expression datasets are huge, so it isn't easy to process those using traditional techniques. Before applying analysis, many datasets need to be pre-processed to remove outliers and missing values. To analyze the behavior of the gene expression, there is a need for an experimental value of the gene expression. Therefore, in this article, we will use the RNA-Seq cancer gene expression dataset. Many classical and heuristic clustering algorithms have been used to analyse the biological data related to genes and their expression. A large amount of gene expression data have been generated, but there is a great requirement for developing the methods to analyze and explore the genes and related information. Clustering is a very useful technique for analyzing gene expression data. Here, we are studying the RNA-Seq cancer dataset to classify gene expression in patients with different types of tumors. Samples (instances) are stored line by line (row-wise). The variable (property or attributes) for each sample is the degree of RNA-Seq gene expression measured using the Illumina HiSeq platform. The number of instances in the dataset is 801, and the number of attributes is 20531[26].

4. IMPLEMENTATION AND RESULT

In this section, we have implemented different clustering algorithms such as K-Means, SOM, and DBSCAN and Hierarchical Clustering in the python environment. Further, the defined cluster has been evaluated by the evaluation metrics such as Fowlkes-Mallows Index, NMI, Adjusted Mutual Information, Homogeneity Score, Completeness, Rand Index, Adjusted Rand Index, Calinski-Harabasz Index, Davies-Bouldin Index, and Silhouette Index. Based on the literature review, the k-value for the given dataset is fixed as 5. With 5 k-value the result of k-means clustering algorithm is given in figure 1.

Form the figure 1, it is clearly seen that the nodes are clustered together very tightly, if we consider k=5. Five different clusters are clearly visible in figure 1. The Hierarchical clustering with K = 5 is given in figure 2. In this figure, we can also point the same factor that given by the k-means. Next the same parameter k=5 is given for the DBSCAN and it can be seen from the figure 3. The DBSCAN have the similar outcome for the given datasets. Next, the SOM clustering method is applied with same parameter k=5. The outcome of the SOM is given in figure 4.

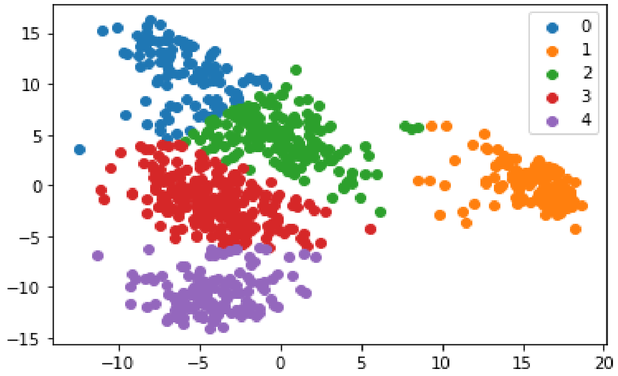

Fig. 1. K- Means Clustering with K = 5 on the cancer RNA-Seq Dataset.

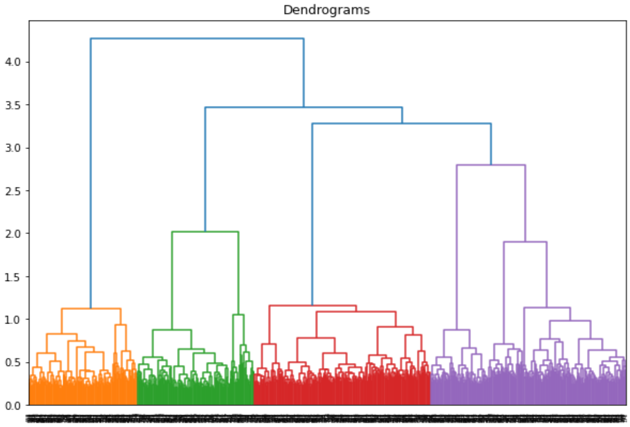

Fig. 2. Hierarchical clustering with K = 5 on the cancer RNA-Seq Dataset.

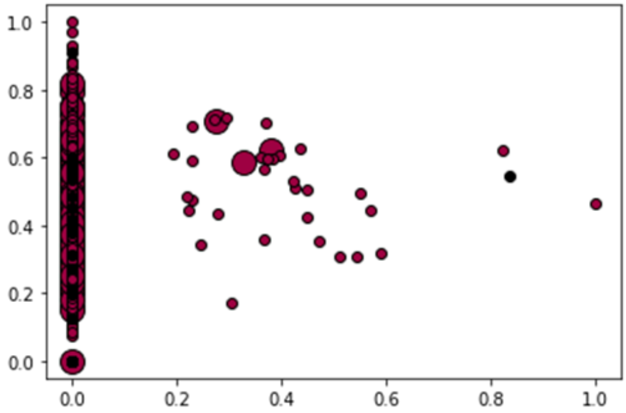

Fig. 3. DBSCAN clustering with K = 5 on the cancer RNA-Seq Dataset.

As similar to the K-means and hierarchal clustering, here we have set the k-value to 5, but it is giving similar to k-means or hierarchal clustering. The same can be seen in figure 3.

The result of the SOM clustering with k=5 is given in figure 4.

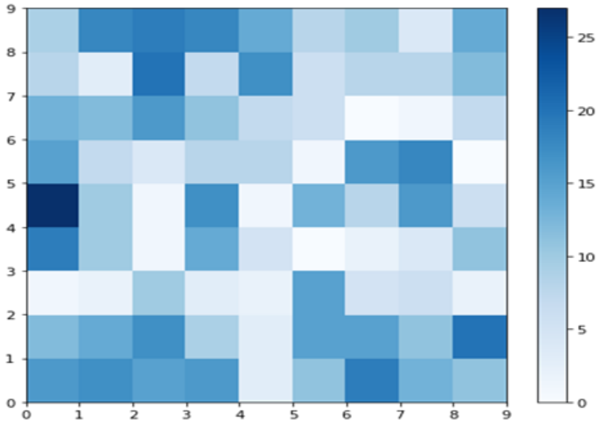

Fig. 4. SOM with K = 5 on the cancer RNA-Seq Dataset.

Table 1: Result of validation techniques for all the clustering techniques

Validation Index	Clustering Algorithm			
	K-Means	Hierarchical Clustering	DBSCAN	SOM
Silhouette Index	0.170	0.172	0.166	0.134
Davies-Bouldin index	2.335	2.220	4.252	2.880
Calinski-Harabaz Index	88.012	86.916	6.141	76.179
Adjusted Mutual Information	0.874	0.975	0.010	0.816
Normalized Mutual Information	0.875	0.975	0.0128	0.832
Fowlkes mallows Index	0.851	0.987	0.476	0.744
Homogeneity Score	0.883	0.974	0.007	0.837
Completeness	0.867	0.977	0.081	0.798
Rand Index	0.93	0.993	0.269	0.88
Adjusted Rand Index	0.805	0.983	-0.002	0.739

From the above figures, we can see that all of the above clustering algorithms are working well on the cancer RNA-Seq dataset and the cluster of the genes in a very skew range of width. Next we have validated the above discussed clustering method using different clustering validation techniques. The computed cluster has been validated with internal and external validation techniques that have been mentioned in the previous section. The result of the validation techniques is given in table 1.

Form the table 1, it can be clearly seen that the most of the validation techniques gives better results for the hierarchical clustering technique except Silhouette Index, DBI and CHI . The Hierarchical clustering fits well for the Silhouette Index, FMI, NMI, AMI, Homogeneity Score, Completeness, Rand index, and Adjusted Rand Index. K-Means clustering algorithm gives better results for the CHI and the DBI gives better results for the DBSCAN. Apart from that SOM gives a better result than DBSCAN (except DBI) and worse than K-Means and Hierarchical clustering for all the indices values. Based on this analysis, we can conclude that the hierarchal clustering technique can be used for cancer gene classification and prediction. To validate the above statement, we have drawn the comparative graph of all the clustering techniques with the validation techniques used. The plot graph is shown in figure 5.

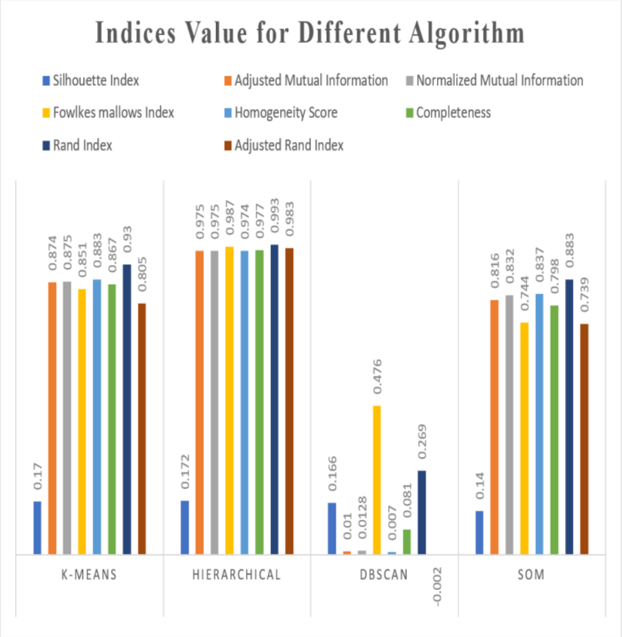

Figure 5: Comparative analysis of clustering algorithm with the validation index value.

Further, we made a comparative analysis for the validation index against clustering techniques. The plot of comparison is shown in figure 6.

From figures 5 and 6, we can also conclude that the hierarchical clustering methods can be used for cancer RNA-Seq gene classification and prediction. This statement also validates the previous statement.

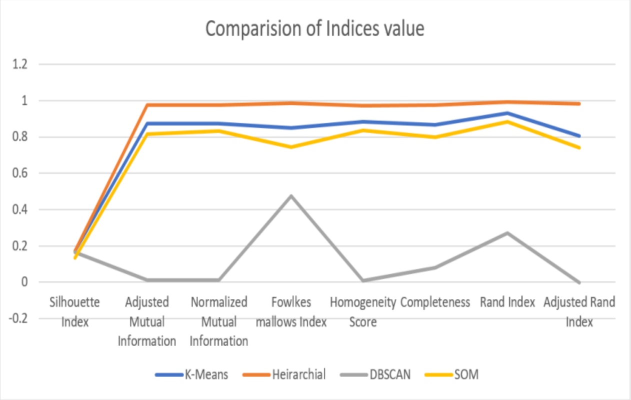

Fig. 6. Comparative analysis of the validation indices against clustering technique.

5. CONCLUSION

Gene expression data hides important information needed to understand the biological processes occurring in a particular organism concerning its environment. Several clustering algorithms have been developed to extract useful information about the behavior of genes concerning various systemic conditions. Clustering has been continuously applied in the Medical field to detect and analyses various diseases such as cancer, malaria, asthma, and tuberculosis. Biological data has increased exponentially due to new technologies and advanced research. The traditional method for data analysis generally fails to find the hidden patterns in the datasets. So, data mining is a useful technology for finding meaningful patterns from the genomics.

REFERENCES

[1] Ishida, Seiichi, Erich Huang, Harry Zuzan, Rainer Spang, Gustavo Leone, Mike West, and Joseph R. Nevins. “Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis.” Molecular and cellular biology 21, no. 14 (2001), pp. 4684-4699.

[2] Meriem Kherbache, David Espes, Kamal Amroun, "An Enhanced approach of the K-means clustering for Anomaly-based intrusion detection systems", 2021 International Conference on Computing, Computational Modelling and Applications (ICCMA), pp.78-83, 2021.

[3] A. M. Fahim, A. M. Salem, F. A. Torkey, and M.A. Ramadan, "Density Clustering Based on Radius of Data (DCBRD)," World Academy of Science, Engineering and Technology 2006.

[4] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In Kdd, Vol. 96, No. 34, pp. 226–231. 1996.

[5] Kumar Utkarsh, Fei Ding, "Self-Organizing Map-Based Resilience Quantification and Resilient Control of Distribution Systems Under Extreme Events", IEEE Transactions on Smart Grid, vol.13, no.3, pp.1923-1937, 2022.

[6] Mallick P, Ghosh O, Seth P, Ghosh A (2019) Kohonen's self-organizing map optimizing prediction of gene dependency for cancer mediating biomarkers. In: Advances in Intelligent Systems and Computing. pp. 863–870

[7] Chen N, Chen L, Ma Y, Chen A (2019) Regional disaster risk assessment of China based on self-organizing map: clustering, visualization and ranking. Int J Disaster Risk Reduct 33:196–206.

[8] Tamayo, Pablo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub. "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proceedings of the National Academy of Sciences 96, no. 6 (1999): 2907–2912.

[9] Alican Dogan, Derya Birant, K-centroid link: a novel hierarchical clustering linkage method, Applied Intelligence, 10.1007/s10489-021-02624-8, 52, 5, (5537-5560), (2021).

[10] Jiahui Wei, Huifang Ma, Yuhang Liu, Zhixin Li, Ning Li, Hierarchical high-order co-clustering algorithm by maximizing modularity, International Journal of Machine Learning and Cybernetics, 10.1007/s13042-021-01375-9, 12, 10, (2887-2898), (2021).

[11] Karypis, George, Eui-Hong Han, and Vipin Kumar. "Chameleon: Hierarchical clustering using dynamic modeling." Computer 32, No. 8 (1999): 68–75.

[12] Jiang, Daxin, Chun Tang, and Aidong Zhang. "Cluster analysis for gene expression data: a survey." IEEE Transactions on knowledge and data engineering 16, No. 11 (2004): 1370–1386.

[13] Kerr, Grainne, Heather J. Ruskin, Martin Crane, and Padraig Doolan. "Techniques for clustering gene expression data." Computers in biology and medicine 38, No. 3 (2008): 283–293.

[14] Starczewski, A., Krzyżak, A. (2016). A Modification of the Silhouette Index for the Improvement of Cluster Validity Assessment. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L., Zurada, J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2016. Lecture Notes in Computer Science (), vol 9693. Springer, Cham.

[15] D. L. Davies and D. W. Bouldin, "A cluster separation measure", IEEE transactions on pattern analysis and machine intelligence, no. 2, pp. 224-227, 1979.

[16] CALINSKI T, HARABASZ J. A dendrite method for cluster analysis [J]. Communications in Statistics, 1974, 3(1):1-27.

[17] S. Romano, J. Bailey, V. Nguyen, and K. Verspoor. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In International Conference on Machine Learning, pages 1143–1151, 2014.

[18] P. A. Estévez, M. Tesmer, C. A. Perez and J. M. Zurada, "Normalized mutual information feature selection", IEEE Trans. Neural Netw., vol. 20, no. 2, pp. 189-201, Feb. 2009.

[19] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings", Journal of The American Statistical Association, vol. 78, no. 383, pp. 553-569, 1983.

[20] Xue, J., Jiang, N., Liang, S. et al. Quantifying the spatial homogeneity of urban road networks via graph neural networks. Nat Mach Intell 4, 246–257 (2022).

[21] Liu, M., Thomas, P.D. GO functional similarity clustering depends on similarity measure, clustering method, and annotation completeness. BMC Bioinformatics 20, 155 (2019).

[22] Steinley, D. and Brusco, M. J. (2018). A note on the expected value of the Rand index. British Journal of Mathematical and Statistical Psychology, 71, 287–299.

[23] Steinley, D., Brusco, M. J. and Hubert, L. (2016). The variance of the adjusted Rand index. Psychological Methods, 21, 261–272.

[24] Bihari, Anand and Tripathi, Sudhakar and Deepak, Akshay, Gene Expression Analysis Using Clustering Techniques and Evaluation Indices (March 11, 2019). Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019, Available at SSRN: https://ssrn.com/abstract=3350332 or http://dx.doi.org/10.2139/ssrn.3350332

[25] Nies, H. W., Zakaria, Z., Mohamad, M. S., Chan, W. H., Zaki, N., Sinnott, R. O., ... & Corchado, J. M. (2019). A review of computational methods for clustering genes with similar biological functions. Processes, 7(9), 550.

[26] https://archive.ics.uci.edu/ml//datasets/gene+expression+cancer+RNA-Seq

[27] Uoc, N. Q., Duong, N. T., Son, L. A., & Thanh, B. D. A (2022) Novel Automatic Detecting System for Cucumber Disease Based on the Convolution Neural Network Algorithm. GMSARN International Journal 16 (2022) 295-301

[28] KS, A. K., Sarita, K., Kumar, S., Saket, R. K., & Swami, A.(2022) Machine Learning-based Approach for Prevention of COVID-19 using Steam Vaporizer GMSARN International Journal 16 (2022) 399-404.

[29] Chandra, Girish, Akshay Deepak, and Sudhakar Tripathi. "A Graph-Based Method for Clustering of Gene Expression Data with Detection of Functionally Inactive Genes and Noise." Advances in Machine Learning and Data Science: Recent Achievements and Research Directives. Springer Singapore, 2018.