IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 May 2023, accepted 11 June 2023, date of publication 21 June 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3288344

RESEARCH ARTICLE

Optimizing Clustering Algorithms for Anti-Microbial Evaluation Data: A Majority Score-Based Evaluation of K-Means, Gaussian Mixture Model, and Multivariate T-Distribution Mixtures

HIRA MAHMOOD[1], TAHIR MEHMOOD[1], AND LAILA A. AL-ESSA[2]
[1]Department of Mathematics, School of Natural Sciences, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Tahir Mehmood (tahime@gmail.com)

This work was supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, through the Princess Nourah bint Abdulrahman University Researchers Supporting Project under Grant PNURSP2023R443.

ABSTRACT This study presents a detailed analysis of the performance of the majority score clustering algorithm on three different datasets of anti-microbial evaluation, namely the minimum inhibitory concentration (MIC) of bacteria, and the antifungal activity of chemical compounds against 4 bacteria (E. coli, P. aeruginosa, S. aureus, S. pyogenes) and 2 fungi (C. albicans, As. fumigatus). Clustering is an unsupervised machine learning method used to group chemical compounds based on their similarity. In this paper, we apply the k-means clustering, Gaussian mixture model (GMM), and mixtures of multivariate t distribution to antibacterial activity datasets. To determine the optimal number of clusters and which clustering algorithm performs best, we use a variety of clustering validation indices (CVIs) which include within sum square (to be minimized), connectivity (to be minimized), Silhouette Width (to be maximized), and the Dunn Index (to be maximized). Based on the majority score clustering algorithm, we conclude that the k-means and mixture of multivariate t-distribution methods perform best in terms of the maximum CVIs, while GMM performs best in terms of the minimum CVIs. K-means clustering and mixture of multivariate t-distribution provide 3 optimal clusters for the anti-microbial evaluation of antibacterial activity dataset and 5 optimal clusters for the MIC bacteria dataset. K-means clustering, mixture of multivariate t-distribution, and GMM provide 3 optimal clusters for both the antibacterial and antifungal activity datasets. K-means clustering algorithm performs the best in terms of the majority-based clustering algorithm. This study may be useful for the pharmaceutical industry, chemists, and medical professionals in the future.

INDEX TERMS Clustering, K-means, GMM, multivariate t distribution, Silhouette width, within sum square, Dunn index.

I. INTRODUCTION

Clustering is an unsupervised machine learning method [1]. In clustering, data objects are partitioned into groups based on distance dissimilarity among data objects [2]. Data objects which are like or near to each other are placed within the same cluster while unlike or far-off data objects are placed in another cluster. Like classification, clustering is also classifying the data objects but unlike classification, the class labels are unknown because clustering is based on unsupervised learning. The clusters are defined based on the study of the behavior or characteristics of the data objects by the domain experts [3]. The clustering algorithms must have the following properties: Data objects within the cluster must be like or near to each other as much as possible. Data objects belonging to different clusters must be dissimilar or far off from each other as much as possible. The distance/similarity measure must have some practical ability and be clear. Clustering is also extensively used in many application domains i.e. statistics, image

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

segmentation, pharmaceutical industry, object recognition, information retrieval, bioinformatics etc [4]. There are two types of clustering algorithms soft and hard clustering algorithms. The data points completely belonging to just one cluster are called hard clustering and a data point belonging to more than one cluster is called soft clustering. There are many clustering algorithms known, But few of the clustering algorithms are used mostly, which is k means clustering algorithm [5], fuzzy c means clustering algorithm [6], GMM [7], hierarchical clustering (agglomerative and divisive algorithm) [8], mixture of multivariate t-distribution [9] and density-based spatial clustering [10]. K means, hierarchical, and density-based spatial clustering is the type of hard clustering on the other hand GMM, mixture of multivariate t-distribution, and fuzzy c mean clustering is the type of soft clustering. Here we discussed one of the applications related to clustering in the medical and chemistry field which is chemical compounds of antibacterial activity. Antibiotics are the most important weapons in the fight against microbial infections and have enormously benefited the health-related quality of human life since their introduction.We adapted three different clustering algorithms which are k means, GMM, and mixture of multivariate t-distribution clustering algorithms for the grouping of chemical compounds having alike antibacterial activity. A variety of indices aimed at validating the results of clustering analysis and determining which clustering algorithm performs best.

The most important question is how many optimal clusters are enough. To solve this problem, we will use different CVIs. As an unsupervised learning task, it is necessary to find a way to validate the goodness of partitions after clustering. In this present paper, we introduce the term ''majority-based decision''. The ''majority-based decision'' rule depends on an individual decision of each CVIs, where the final decision is made by the majority of the total CVIs votes. This method delivers fast solutions and follows a clear rule of using independent CVIs in the validation process of clustering algorithms. The two main categories of clustering validation are external, and internal clustering [11]. The main difference between clustering validation is that evaluating the results of a clustering algorithm based on prior information of data is called external validation, whereas internal validation does not. An example of an external validation measure is entropy, which evaluates the ''purity'' of clusters based on the given class labels [12]. In this paper, we did the internal clustering validation techniques/indices because we have the class labels of antibacterial activity data sets. Internal validation measures reflect often the compactness, the contentedness, and the separation of the cluster partitions [13]. *Cluster cohesion or compactness:* Measures how near are the data points within the same cluster or groups. The cluster is compact when the variation within a cluster should be minimum. Different Distance metrics can be used to measure the compactness of clusters such as the cluster or group-wise within average or median distances. *Separation:* Separation is used to measure the segregation of clusters or groups from each other. Distances between cluster centers and pairwise minimum distances between items in various clusters are among the cluster validation indices used as separation metrics. *Connectivity:* In the data space, connectivity refers to the extent to which things are clustered with their closest neighbors. The connection, which ranges from 0 to infinity, should be kept to a minimum. There are other internal clustering validation approaches, but we used four of the most relevant ones here: Within sum square (to gauge cluster compactness), Connectivity (how data points connect), DI, and SW (how well separate clusters). Section II gives an overview of the methodology and section III contains the data explanation. In section IV we explained results. Section V presents the conclusion and future scope.

II. REFERENCE METHODS

For clustering of chemical compounds having alike antibacterial activity, we applied proposed methods K means, GMM, and a mixture of multivariate t-distribution clustering algorithms.

A. K MEANS CLUSTERING

K means is a partitional clustering algorithm. Data points are divided into non-overlapping groups. It is most easy to understand and useful method. Its gives better results as compare to other algorithms. K means groups data points using distance from the cluster centroid. The objective of K means clustering is to minimize total intra-cluster variance or the squared error function [14]:

$$X = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{1}$$

where X,K, and n objective function, number of clusters, and number of cases

Calculations steps of K means clustering:

1) Initially, k must be predefined.
2) Select k points at random as centroids.
3) Assign data points to their closest cluster center according to the Euclidean distance function.
4) Calculate the centroid or mean of all objects within the cluster.
5) Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive iteration.

Limitations of K means clustering algorithm:

1) It requires to specify the number of clusters (k) in advance.
2) It can not handle noisy data and outliers.
3) It is not suitable to identify clusters with non-convex shapes.
4) Curse of dimensionality.

B. GAUSSIAN MIXTURE MODEL CLUSTERING

In clustering, the mixture model helps us to identify the cluster model that describes a data set by combining a mix

of two or more probability distributions. Each component of the cluster is considered as a model with mean and variance. Mixture models are to estimate the parameters of the probability distribution for each cluster like mean and variance [15].

Probability density function of Gaussian distribution:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (2)$$

Calculation steps for Gaussian mixture model given below:

1) Initialise $\mu, \sigma, \pi(c_j)$ for all clusters

$$\mu_k = \frac{k * \max_j + 1}{N+1}\sigma^2 = max(j) + 1\pi(c_j) = \frac{1}{N} \quad (3)$$

2) Suppose the probability x_i of belonging to any class c_j

$$p(c_j|x_i) = \frac{p(x_i|c_j) * p(c_j)}{p(x_i)} \quad (4)$$

$$p(x_i|c_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu_j}{\sigma_j}\right)^2\right) \quad (5)$$

$$p(x_i) = \sum_j p(x_i|c_j) * p(c_j) \quad (6)$$

Here, $p(x_i|c_j)$,$p(c_j)$, and $p(x_i)$ is likelihood, prior information, and probability of chemical compounds

3) Re-estimate the parameter μ, σ, and $p(c_j)$ as

$$\mu_k = \frac{\sum_i p(c_j|x_i) * x_i}{\sum_i p(c_j|x_i)} \sigma_k = \frac{\sum_i p(c_j|x_i) * (x_i - \mu_k)^2}{\sum_i p(c_j|x_i)} \quad (7)$$

$$p(c_j) = \frac{\sum_i p(c_j|x_i)}{n} \quad (8)$$

4) Iterate until convergence.

Limitations of GMM:

1) GMM is a complicated algorithm and cannot be implemented to larger data.
2) It is difficult to find clusters if the data is not Gaussian, hence a lot of data preparation and information is required.

C. MIXTURE OF MULTIVARIATE T-DISTRIBUTION CLUSTERING ALGORITHM

The mixture of multivarite t-distributions assume that each sub-population of the observed data follow the multivariate-t distribution [16].

$$P(x_i|\theta_k) = \frac{\Gamma((v+p)/2)|\sigma_k|^{-1/2}}{\Gamma(1/2)\Gamma(1/2)v^{p/2}} * \frac{1}{[1+(\delta(x_i,\mu_k;\Sigma_k)/v)]^{(v+p)/2}} \quad (9)$$

δ represented mahalanobis distance between x_i and μ_k squared is given below:

$$\delta(x_i,\mu_k;\Sigma_k) = (x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k) \quad (10)$$

Calculation steps for mixture of multivariate t-distribution given below:

Expectation Maximization algorithm(EM) to find the unknown parameter of mixture of multivariate t-distribution

E-step:

$$\hat{Z}_{ik} = E_{f(y,w)}(Z_{ik}) = \frac{\pi_k P(y_i|\mu_k^{j-1}), \sum_k^{j-1}, v}{\sum_{i=1}^g \pi_j P(y_i|\mu_k^{j-1}, \sum_k^{j-1}, v)} \quad (11)$$

with $P(x_i|\mu_k, \Sigma_k, v)$ define in eq (7)

$$\hat{\mu}_{ik} = \frac{P+V}{\delta(x_i, \mu_k^{j-1}), \Sigma_k^{j-1} + V}$$

Here, P and V are matrix dimension and degree of freedom

M-steps:

$$\pi_k = \frac{\Sigma_{i-1}^N \hat{Z}_{ik}}{N} \quad (12)$$

$$\mu_k^j = \frac{\Sigma_{i-1}^N \hat{Z}_{ik}\hat{\mu_{ik}}x_i}{\Sigma_{i-1}^N \hat{Z}_{ik}\hat{\mu_{ik}}} \quad (13)$$

$$\Sigma_k^j = \frac{\sum_{i=1}^N (\hat{Z}_{ik}\hat{\mu_{ik}})(x_i-\mu_k^j)(x_i-\mu_k^j)^T}{\Sigma_{i-1}^N \hat{Z}_{ik}} \quad (14)$$

for mixture model:

$$\Sigma_j^k = \frac{\sum_{i=1}^N (\hat{Z}_{ij}\hat{\mu_{ij}})(x_i-\mu_j^k)(x_i-\mu_j^k)^T}{\Sigma_{i-1}^N \hat{Z}_{ij}\hat{\mu_{ij}}} \quad (15)$$

$$\Sigma^j = \frac{\sum_{i=1}^N \sum_{k=1}^g (\hat{Z}_{ik}\hat{\mu_{ik}})(x_i-\mu_k^j)(x_i-\mu_k^j)^T}{\Sigma_{i-1}^N \sum_{k=1}^g \hat{Z}_{ik}\hat{\mu_{ik}}} \quad (16)$$

Check for convergence.

Limitations of mixture of multivariate t-distribution algorithm:

1) It converges slowly.
2) It just reaches the local optimum.

D. CLUSTERING VALIDATION INDICES

External and Internal validation are two types of CVIs. In this paper we applied the external validation techniques [17], [18].

1) WITHIN SUM SQUARE (WSS)

How distinct or well isolated from one another a cluster is measured using the WSS method. Elbow curve is another name for WSS. To determine the maximum number of clusters that may be constructed for a certain data set, WSS is used as a metric. The squared distance between each cluster member and its centroid is added to create the WSS.

$$WSS = \sum_{i=1}^{n} (x_i - c_i^2) \quad (17)$$

Here, x_i is data points and c_i is closest point to centroid.

2) SILHOUETTE WIDTH (SW)

The SW measures how similar a data points is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette width or index is between +1 and −1. A high value or close to +1 is d indicates that the data point is placed into the correct cluster. If many data points have a negative Silhouette width it may indicate that we have created too many or too few clusters [19].

https://www.datanovia.com/en/lessons/cluster-validation - statistics-must-know-methods. SW is defined as follows:

$$s_i = \frac{x_i - y_i}{\max(x_i, y_i)} \tag{18}$$

Were, x_i is the average distance and y_i is the minimum average distance of

$$x_i = \frac{1}{|c_i| - 1} \sum_{j \in c_i, j \neq i} d(i,j) \;\; and \;\; y_i = \min_{i \neq j} \frac{1}{|c_i|} \sum_{j \in c_i} d(i,j) \tag{19}$$

$d(i, j)$ is the distance between data points i and j. Generally, Euclidean Distance is used to measure the distance metric.

3) CONNECTIVITY

In the data space, connectivity measures how closely entities are clustered with their closest neighbours. The connection should be kept to a minimum because its value ranges from 0 to infinity [20]. Most internal clustering validation methods often incorporate compactness and separation metrics as shown below:

$$index = \frac{(a * separation)}{(b * Compactness)} \tag{20}$$

where a and b are weights.

4) DUNN INDEX (DI)

J. C. Dunn created DI in 1974, which is another method for validating clusters. The DI measures the ratio of the highest intra-cluster distance or diameter to the shortest distance between observations that are not in the same cluster. It is best to maximize DI, which ranges from 0 to infinity. The most practical index for cluster validation is DI [19].

To calculate DI:

$$D = \frac{min.separation}{max.diameter} \tag{21}$$

Maximum diameter as the intra-cluster compactness and minimum separation as inter-cluster separation.

E. PROPOSED METHOD: MAJORITY SCORE CLUSTERING ALGORITHM

The different CVIs provides a different optimal number of cluster, which result of clustering algorithm is not reliable. Instead of selecting optimal number of cluster based on individual CVIs, we have introduced a majority scoring clustering algorithm where clustering algorithm is selected if it is satisfy by more than two or more combinations of CVIs.

TABLE 1. The contents of antimicrobial evaluation, Microbial inhibitory concentration, and antibacterial activity against considered microbes are presented.

Antimicrobial evaluation	Min	Max	Mean	S.D
E. Coli	62.5	500	207.64	122.41
P. Aeruginosa	100	250	200	53.55
S. Aureus	62.5	500	204.17	102.54
S. Pyogenus	100	500	222.22	89.07
C. Albicans	200	1200	561.11	402.4
As. Fumigatus	250	1200	858.33	337.92
Microbial inhibitory concentration	Min	Max	Mean	S.D
S. Aureus	1.61	25	6.27	7.58
B. Subtilis	1.61	25	6.58	7.59
E. Coli	1.61	25	9.31	9.40
P. Aeruginosa	1.61	25	11.67	10.45
Antibacterial activity	Min	Max	Mean	S.D
S. Aureus	1.60	50	12.97	12.50
E. Faecalis	1.60	50	13.52	14.75
E. Coli	3.12	50	16.72	16.04
K. Pneumoniae	3.12	50	20.00	16.64
C. Albicans	0.80	25	8.64	8.27
A. Fumigates	1.60	25	9.15	8.89

We cluster the chemical compounds of antibacterial activity on the bases of majority score clustering algorithm.

III. DATA EXPLANATION AND STATISTICAL SOFTWARE

The data is taken from the studies [21], [22], [23] in which, the 1st and 3rd data set contains six variables (E. Coli, P. Aeruginosa, S. Aureus, S. Pyogenus, C. Albicans, and As. Fumigatus) and 2nd data set only contains four variables (E. Coli, P. Aeruginosa, S. Aureus, and S. Pyogenus). In these studies, the chemical compounds of antibacterial activity are labeled alphabetically. We applied clustering algorithms and CVIs to classify these labels of chemical compounds on the bases of antibacterial activity.

A. COMPUTATION

R is used for both computations statistical analysis and modelling. https://www.R-project.org/. R packages used kmeans, mclust, and teigen for clustering algorithm and for CVIs used WithinSS, SilWidth, Conn and Dunn.

IV. RESULTS AND DISCUSSION

In antimicrobial evaluation study sample of 18 ionic liquids is considered. The antibacterial activity of these samples against six microbes E. Coli, E. Aerogenes, K. Pneumoniae, P. Vulgaris, P. Aeruginosa, and S. Pyogenes is monitored. The mean, maximum, minimum, and standard deviation (S.D) of antimicrobial evaluation against considered microbes. The Table 1 shows the most antibacterial activity against S. Pyogenes is 858.33 and the least antibacterial activity against P. Aeruginosa is 200.

The number of clusters (k) must be set before we start the algorithm, it is often advantageous to use several different values of k and examine the differences in the results. We can execute the same process for 2, 3, 4, and 5 clusters, The Figures reffig:1,2,3,4 represents k means, and the mixture of multivariate t-distribution clustering algorithms curves gives

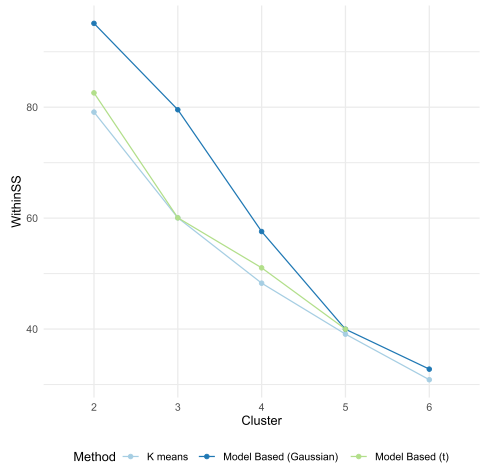


FIGURE 1. Within sum square CVI of Antimicrobial evaluation.

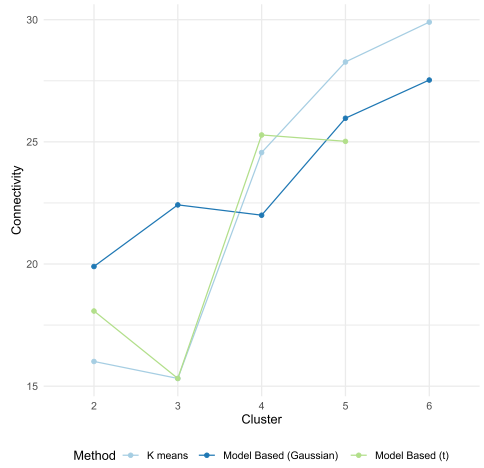


FIGURE 2. Connectivity CVI of Antimicrobial evaluation.

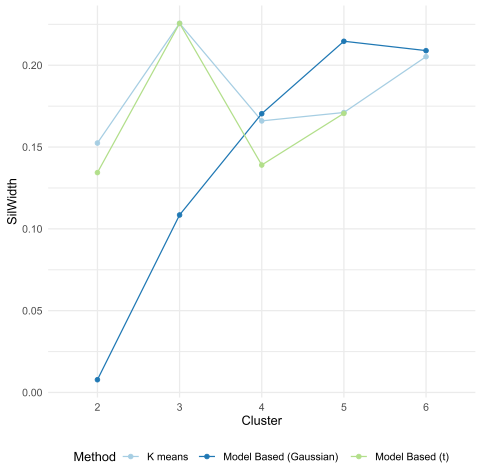


FIGURE 3. Silhoutte width CVI of Antimicrobial evaluation.

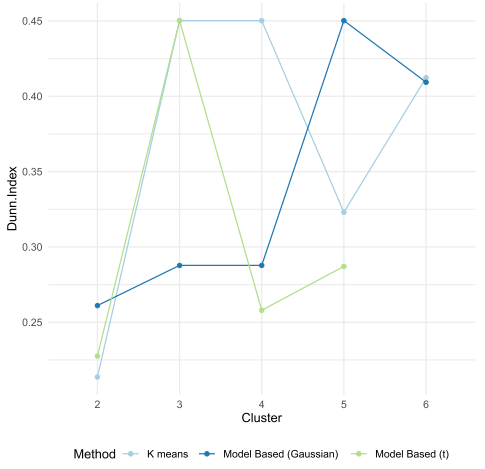


FIGURE 4. Dunn Index CVI of Antimicrobial evaluation.

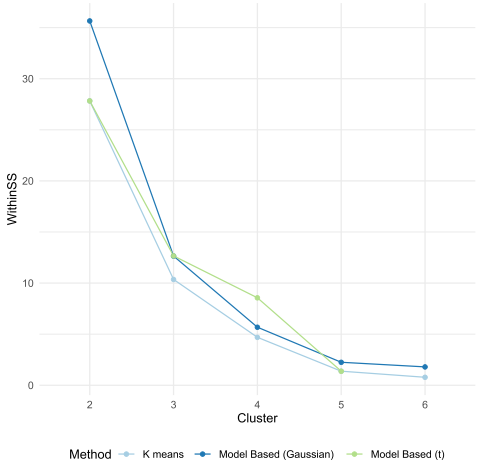


FIGURE 5. Within sum Saqure CVI's of minimum inhibitory concentration.

maximum SW, minimum connectivity, and maximum DI of antimicrobial evaluation of the antibacterial activity at 3 optimal number of clusters. As illustrated in Figures 5,6,7,8 for MIC of antibacterial activity, k means and multivariate t distribution clustering algorithms give almost the same results as the curves of WSS is minimum of k means, SW at 5 optimal number of clusters, and the minimum connectivity of k means and a mixture of multivariate t-distribution clustering algorithms lowest point at 2 number of clusters. The curve of DI is maximum at optimal 5 number of clusters of k means and multivariate t distribution. In Figures 9,10,11,12 Antibacterial activity of chemical compounds distributed the k means gives minimum WSS at 6 optimal clusters and maximum SW of k means and GMM at 3 optimal clusters. The minimum connectivity of k means cluster at 2 optimal clusters. The maximum point of the DI curve at 3 optimal clusters. In figures 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 we demonstrate the results on the bases of the "majority score" term, the k means and mixture of multivariate t-distribution clustering algorithms give 3 optimal numbers of clusters in an antimicrobial evaluation of antibacterial activity and 5 optimal clusters MIC of bacteria's. K means, mixture of multivariate t- distribution

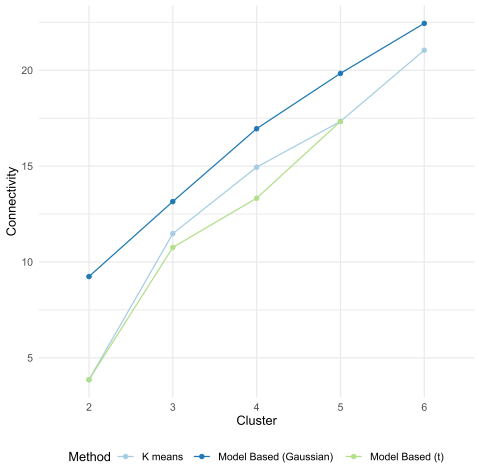


FIGURE 6. Connectivity CVI's of minimum inhibitory concentration.

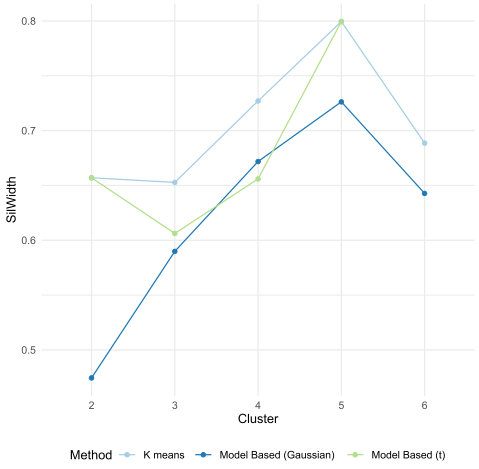


FIGURE 7. Silhoutte width CVI's of minimum inhibitory concentration.

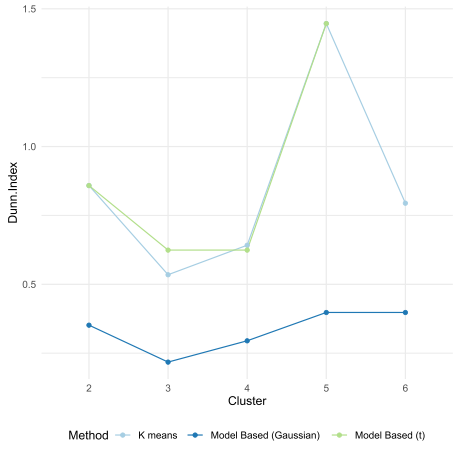


FIGURE 8. Dunn index CVI's of minimum inhibitory concentration.

and GMM give 3 optimal numbers of clusters in antibacterial activity data set. The k means clustering algorithm gives the best performance on the bases of the "Majority Score Clustering Algorithm".

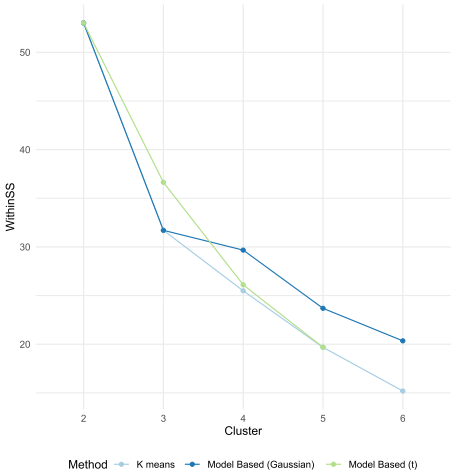


FIGURE 9. Within sum saqure CVI of Antibacterial activity.

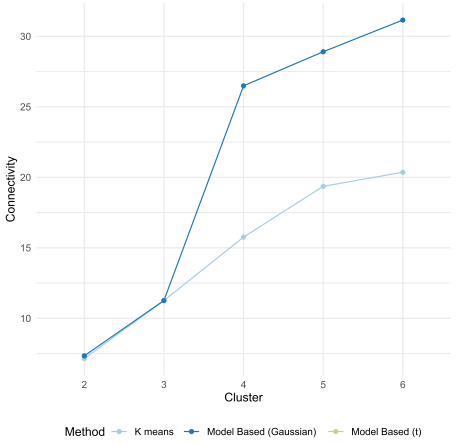


FIGURE 10. Connectivity CVI of Antibacterial activity.

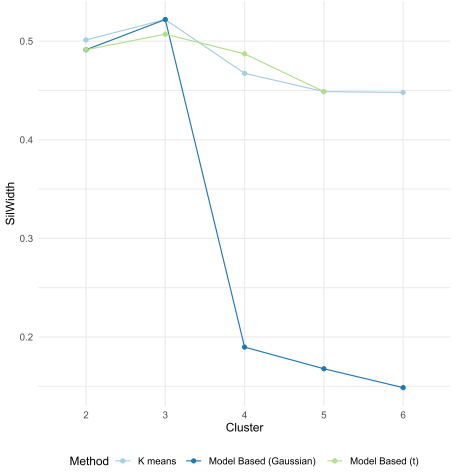


FIGURE 11. Silhoutte CVI of Antibacterial activity.

The Table 3 represents the best model according to Bayesian Information Criteria (BIC) is an unequal-Covariance model with 3 optimal number of components

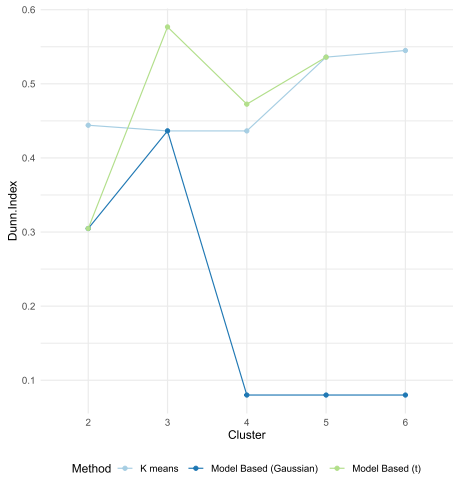


FIGURE 12. Dunn Index CVI of Antibacterial activity.

TABLE 2. The table explained cluster means of antimicrobial evaluation of antibacterial activity, MIC of bacteria's, and antibacterial activity.

No.of clusters	E. Coli	P. Aeruginosa	S. Aureus	S. Pyogenus	C. Albicans	As. Fumigatus
1	0.62	0.70	0.02	-0.53	0.12	-0.17
2	-0.45	-0.77	0.28	0.77	-0.81	0.09
3	-0.57	-0.23	-0.46	-0.10	1.46	0.19
No.of clusters	S. Aureus	B. Subtilis	E. Coli	P. Aeruginosa		
1	2.47	2.29	1.64	1.19		
2	-0.61	-0.65	1.61	1.22		
3	0.10	0.26	0.01	0.67		
4	2.47	2.42	1.66	1.27		
5	-0.45	-0.49	-0.72	-0.84		
No.of clusters	S. Aureus	E. Faecalis	E. Coli	K. Pneumoniae	C. Albicans	A. Fumigates
1	-0.65	-0.59	-0.52	-6.05	-0.72	-0.69
2	0.46	1.28	-0.18	2.00	0.91	1.22
3	1.21	0.03	1.68	1.80	0.84	0.37

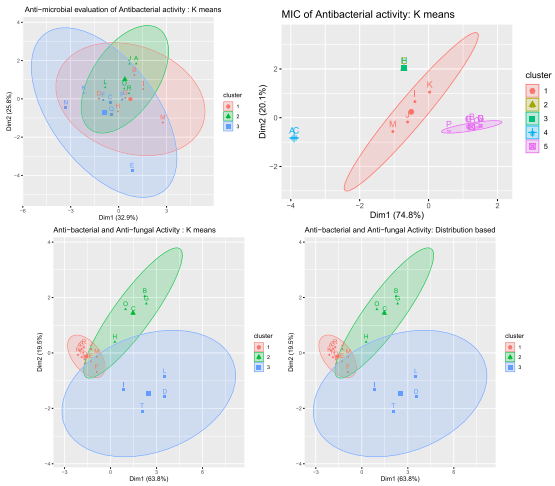


FIGURE 13. The clustering of chemical compounds with similar antibacterial activity using K means and GMM clustering algorithm.

TABLE 3. The table represents GMM clusters of antibacterial activity.

log-likelihood	n	df	BIC	ICL
-83.3754	20	23	-235.6526	-235.6644

or clusters. The Figure 15 explained the dendrogram clustering of chemical compounds, 1st dendrogram figure shows the maximum 3 optimal number of clusters and in cluster 1 (G, M, L, A, J) 5 components having alike chemical compounds characteristics, 4 chemical compounds (B, I, N, O, and H) having same properties in cluster two

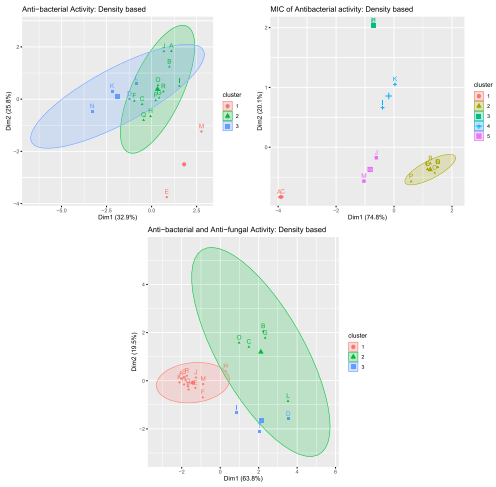


FIGURE 14. The clustering of chemical compounds with similar antibacterial activity using multivariate t-distribution clustering algorithm.

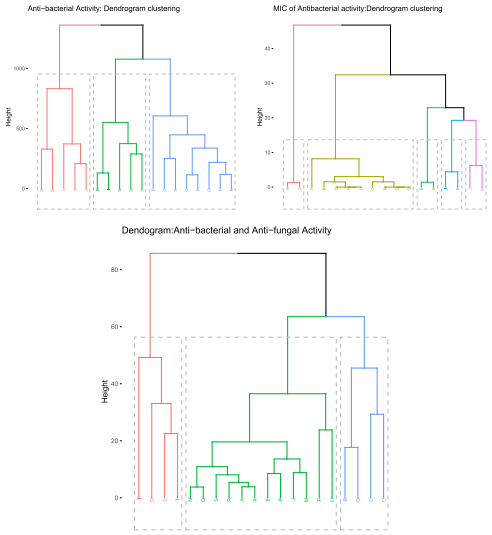


FIGURE 15. The dendrogram clustering of chemical compounds having alike antibacterial activity.

and in cluster 3 remaining 8 chemical compounds (E, O, R, C, P, Q, F, and K) having alike chemical compounds characteristics 04 bacteria's (E. Coli, P. Aeruginosa, S. Aureus, S. Pyogenes) and 02 Fungus (C. Albicans, As. Fumigatus). In 2nd dendrogram reveals that a maximum 5 number of clusters and in clusters 1,2,3,4,5 the chemical compounds having the same properties are (A, C), (P, B, Q, F, L, O, N, D, Q), (E, and H), (I, and K) and (J, and M) against 4 bacteria's (E. coli, B. Subtilis, P. aeruginosa, and S. Aureus). The 3rd dendrogram figure shows maximum 3 optimal number of clusters and in cluster 1 (I, D, L, and T) 4 components having alike chemical compounds characteristics, 12 chemical compounds (N, Q, S, R, A, K, E, P, J, M, F, and H) having same properties in cluster two and in cluster 3 remaining 4 chemical compounds (B, G, C, and O) having alike chemical compounds characteristics 04 bacteria's (E. Coli,

P. Aeruginosa, S. Aureus, and S. Pyogenes) and 02 Fungus (C. Albicans, and As. Fumigatus).

V. CONCLUSION

In this research, we discussed the problem of clustering chemical compounds having alike antibacterial activity using unsupervised machine learning methods and different CVIs. The study identifies the antibacterial activity of ionic liquids against several microbes through the newly proposed majority-based clustering algorithm. k means and mixture of multivariate t-distribution satisfy the maximum and, GMM satisfy the minimum CVIs. The k means algorithm and mixture of multivariate t-distribution give 3 optimal number of clusters in an antimicrobial evaluation of antibacterial activity data set and 5 number of optimal clusters in MIC of bacteria's data set. K means, mixture of multivariate t-distribution and GMM gives 3 optimal numbers of the cluster in the antibacterial activity data set. At the last, we demonstrate that the performance of K means clustering algorithm is better. The proposed method produces more correctly classify chemical compounds, which motivates its application in diverse areas.

CONFLICTS OF INTEREST

The authors affirm that they have no known financial or interpersonal conflicts that would have appeared to have an impact on the research presented in this study.

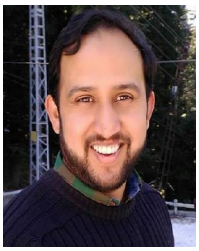
ACKNOWLEDGMENT

This work was supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, through the Princess Nourah bint Abdulrahman University Researchers Supporting Project under Grant PNURSP2023R443.

REFERENCES

[1] R. Gentleman and V. J. Carey, "Unsupervised machine learning," in *Bioconductor Case Studies*. New York, NY, USA: Springer, 2008, pp. 137–157.
[2] C. C. Aggarwal, "An introduction to cluster analysis," in *Data Clustering*. London, U.K.: Chapman & Hall, 2018, pp. 1–28.
[3] V. S. Moertini, "Introduction to five data clustering algorithms," *Integral*, vol. 7, no. 2, pp. 1–10, Oct. 2002.
[4] I. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 3rd ed. Amsterdam, The Netherlands: Elsevier, 2012.
[5] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
[6] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, Jan. 1984.
[7] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, "A robust EM clustering algorithm for Gaussian mixture models," *Pattern Recognit.*, vol. 45, no. 11, pp. 3950–3961, Nov. 2012.
[8] F. Nielsen, "Hierarchical clustering," in *Introduction to HPC with MPI for Data Science*. Cham, Switzerland: Springer, 2016, pp. 195–211.
[9] J. L. Andrews, J. R. Wickins, N. M. Boers, and P. D. McNicholas, "Teigen: An R package for model-based clustering and classification via the multivariate t distribution," *J. Stat. Softw.*, vol. 83, no. 7, pp. 1–32, 2018.
[10] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 231–240, May 2011.
[11] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.
[12] E. Rendón, I. M. Abundez, C. Gutierrez, S. D. Zagal, A. Arizmendi, E. M. Quiroz, and H. E. Arzate, "A comparison of internal and external cluster validation indexes," in *Proc. Amer. Conf.*, vol. 29, San Francisco, CA, USA, 2011, pp. 1–10.
[13] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, Aug. 2005.
[14] Y. Li and H. Wu, "A clustering method based on K-means algorithm," *Phys. Proc.*, vol. 25, pp. 1104–1109, Dec. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1875389212006220
[15] E. Patel and D. S. Kushwaha, "Clustering cloud workloads: K-means vs Gaussian mixture model," *Proc. Comput. Sci.*, vol. 171, pp. 158–167, Jan. 2020.
[16] S. Shoham, "Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions," *Pattern Recognit.*, vol. 35, no. 5, pp. 1127–1142, May 2002.
[17] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 911–916.
[18] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, Jun. 2013.
[19] T. Gupta and S. P. Panda, "Clustering validation of CLARA and K-means using silhouette & Dunn measures on Iris dataset," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCon)*, Feb. 2019, pp. 10–13.
[20] J. C. Rojas-Thomas and M. Santos, "New internal clustering validation measure for contiguous arbitrary-shape clusters," *Int. J. Intell. Syst.*, vol. 36, no. 10, pp. 5506–5529, Oct. 2021.
[21] V. Dhinoja, D. Karia, and A. Shah, "Acid promoted one pot synthesis of some new coumarinyl 3,4'-bipyrazole and their in vitro antimicrobial evaluation chemistry & biology interface," *Chem. Biol. Interface*, vol. 4, pp. 232–245, Aug. 2014.
[22] U. S. Rai, A. M. Isloor, P. Shetty, N. Isloor, M. Padaki, and H.-K. Fun, "A novel series of homoallylic amines as potential antimicrobials," *Medicinal Chem. Res.*, vol. 21, no. 7, pp. 1090–1097, Jul. 2012.
[23] N. Shruthi, B. Poojary, V. Kumar, M. M. Hussain, V. M. Rai, V. R. Pai, M. Bhat, and B. C. Revannasiddappa, "Novel benzimidazole-oxadiazole hybrid molecules as promising antimicrobial agents," *RSC Adv.*, vol. 6, no. 10, pp. 8303–8316, 2016, doi: 10.1039/C5RA23282A.

HIRA MAHMOOD received the M.Phil. degree in statistics from the School of Natural Sciences, National University of Sciences and Technology, Islamabad, Pakistan, in 2022. Her research interests include machine learning, clustering, and bio-statistics.



TAHIR MEHMOOD received the Ph.D. degree in statistics from the Norwegian University of Life Science (NMBU), Norway, in 2012. He is currently a Professor of statistics with the School of Natural Science (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan. His research interests include multivariate statistics, statistical learning, classification, clustering, variable selection, and the application of these methods/algorithm covers chemometrics, enviormetrics, public health, and in related areas.

LAILA A. AL-ESSA received the Ph.D. degree from Princess Nourah bint Abdulrahman University. She is currently an Assistant Professor of mathematics with the Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, Saudi Arabia. Her research interests include reliability estimation, probability distributions, and ordinal statistics.