"in the name of Allah the most beneficent the most merciful"

Majority Score Clustering Algorithms to Identify the Chemical Compounds Having Alike Antibacterial Activity

By

Hira Mahmood
Reg No . 330869

A thesis submitted in partial fulfillment of the requirements for the degree of **Master of Science** in **Statistics**

Supervised by: Tahir Mehmood

School of Natural Sciences
National University of Sciences and Technology
H-12, Islamabad, Pakistan
2022

FORM TH-4

National University of Sciences & Technology

MS THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: **Hira Mahmood**, Regn No. **00000330869** Titled: **"Majority Score Clustering Algorithms to Identify the Chemical Compounds Having Alike Antibacterial Activity"** accepted in partial fulfillment of the requirements for the award of **MS** degree.

Examination Committee Members

1. Name: DR. MUDASSIR IQBAL Signature: _____

2. Name: DR. SHAKEEL AHMED Signature: _____

Supervisor's Name: DR. TAHIR MEHMOOD Signature: _____

_____ 19/08/2022
Head of Department Date

COUNTERSINGED

Date: 22.08.2022 _____
Dean/Principal

I dedicate this thesis to my beloved mother and sister.

Acknowledgments

Allah Almighty, the most beneficent and gracious, who created the whole universe, deserves all honour and glory. I am deeply grateful and indebted to Him for bestowing countless blessings upon me, including the courage and strength to complete my thesis effectively. Without a doubt, my sincerest appreciation goes to my supervisor, **Dr. Tahir Mehmood**, who is one of the best teachers I have ever had. I owe him a lot of gratitude for his support, advice, and especially his constant patience during this journey. May Allah bless him with an abundance of blessings. This research would not have been accomplished without his knowledge, experience, and support. Furthermore, it is completely because of his efforts and appreciative responses to my questions that I have gained a complete grasp of this field.

I would like to pay my gratitude to my GEC members **Dr. Mudasser Iqbal** and **Dr. Shakeel Ahmed** for their support and guidance in completing this thesis. Lastly, I want to thank the support of my sister and friends throughout my studies.

Abstract

This work presents a detailed study of majority-based clustering algorithms decision on three different data sets of anti-microbial evaluation, the minimum inhibitory concentration of antibacterial, antibacterial, and anti-fungal activity of chemical compounds against 04 bacteria (E. Coli, P. Aeruginosa, S. Aureus, S.Pyogenes) and 02 Fungus (C. Albicans, As. Fumigatus). Clustering is an unsupervised machine learning method used to divide the chemical compounds on the bases of their similarity. In this thesis we applied the K-means clustering, Gaussian mixture model (GMM), and Mixtures of multivariate t distribution on antibacterial activity data sets. For an optimal number of clusters and to determine which clustering algorithm performs best we used a variety of clustering validation indices (CVI) which are within sum square (to be minimized), connectivity (to me minimized), silhouette width (to be maximized), the Dunn index (to be maximized). On the bases of the majority score clustering algorithm, we conclude that K-means and the mixture of multivariate t distribution satisfy the maximum and the Gaussian mixture model satisfies the minimum cluster validation indices. The K-means algorithm and mixture of multivariate t distribution give 3 optimal number of clusters in an anti-microbial evaluation of antibacterial activity data set and 5 number of optimal clusters in minimum inhibitory concentration (MIC) of anti bacteria's data set. K-means, Mixtures of multivariate t distribution and Gaussian mixture model give 3 optimal number of clusters in the antibacterial and anti-fungal activity data set. The K-means clustering algorithm gives the best performance on the bases of a majority-based decision. This study may help the pharmaceutical industry, alchemists as well as doctors in the future.

Keywords: Clustering, K-means, GMM, Mixtures of multivariate t distribution, Hierarchical clustering, Silhouette Width, Within sum square, DI

Contents

List of Tables

List of Figures

List of Abbreviations

GMM	Gaussian Mixture Model
ML	Machine Learning
SL	Supervised Learning
UL	Unsupervised Learning
RL	Reinforcement Learning
CVI	Cluster Validation Indices
WSS	Within Sum Square
DI	Dunn Index
SW	Silhouette Width
EM	Expectation Maximization
MLE	Maximum Likelihood Estimation
MIC	Minimum Inhibitory Concentration
ICL	Integrated Complete-data Likelihood
BIC	Bayesian Information Criteria

Chapter 1

Introduction

Despite the use of antibiotics to combat infectious diseases such as bacteria, viruses, fungi, and parasites. Bacteria continue to be the cause of many diseases and deaths as the human population grows. This is evident by the fact that infectious diseases caused by microbial pathogens account for roughly one-quarter of all fatalities worldwide each year, and such bacterial and fungal infections are treated with antibiotics. A molecule's antibacterial and antivirus activity is completely related to substances that kill or slow down the rate of growth of bacteria, fungus, and viruses.

Because of the peculiar shape and chemical composition of the bacterial cell wall, most medicines are ineffective, making different diseases extremely challenging. Chemotherapy for microbial infections has become a serious concern as a result of the rise in multi-drug resistant organisms, which makes the management of infectious diseases shakier.

1.1 Background of Machine Learning

Arthur Samuel first proposed the concept of machine learning in 1959, describing it as a branch of research that allows computers to learn without being explicitly programmed[1] The act of creating computer systems that automatically get better with use and incorporate a learning process is referred to as machine learning (ML). Draw Inference, best to model fit, or learning from examples are still ways to describe machine learning, which automatically learns the theory from data:

- Due to the lack of a universal theory, it is most suitable for domains with a lot of data.
- Automated extraction of significant information using sound probabilistic models from a corpus of data.

1.2 Definition of Machine Learning

Machine learning is more feasible and relevant in modern times thanks to current processing power and breakthroughs in the field. ML is a branch of AI(Artificial Intelligence). ML define that computers programs or machines can learn or pick up new information and adjust to it without human involvement.
The application area of ML, which adds a new dimension to ML, is the other factor for categorizing learning systems. The areas that various existing learning systems have been applied to are listed below as follows: ML can be applied in various fields and the list are given below as follows:

1.3 Application of Machine Learning

1. Game playing and computer programming
2. Speech and image recognition
3. Pharmaceutical and Medical Evaluation
4. Physics and Agriculture
5. Robotics and email management
6. Music
7. Mathematics and numerous other topics

1.4 Types of Machine Learning

Scientists divide the ML into three different types which are given below:

1. Supervised Machine Learning [2]
2. Unsupervised Machine Learning [3]
3. Reinforcement Machine Learning[4]

1.4.1 Supervised Machine Learning (SL)

Supervised machine learning is referred to observe components as input and output data points. Labels (specified as) must be included in the input data. The goal is to figure out how they are related to one another. Regression, prediction, and classification are some of the subjects covered in supervised learning.

1.4.2 Unsupervised Machine Learning (UL)

Unsupervised machine learning is similar to exploratory data analysis and data mining. In UL there are only explanatory variables and no output observations in UL. Because the data is unlabelled. The aim of UL is to find the pattern of data on the relation of data points. It's limited to figuring out what can be deduced from the data. The main topic of unsupervised machine learning is clustering.

1.4.3 Reinforcement Machine Learning (RL)

RL is a type of ML in which an agent makes a decision and takes measures based on the situation and is optimally rewarded for it. The agent is merely rewarded for executing the correct actions in this learning technique, hence there is no need to define how the action should be handled. The goal is to create an agent that performs actions accurately after learning via trial and error in a dynamic environment.

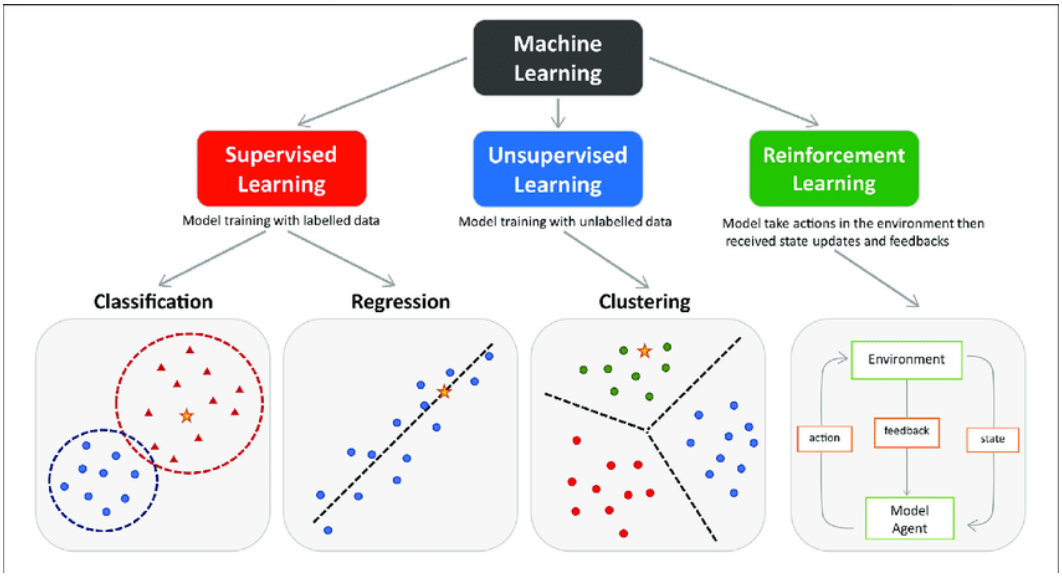


Figure 1.1: The flow chart of machine learning types

1.5 Clustering

In this thesis, we used clustering which is an unsupervised ML method[5].In clustering, data points are partition into groups based on distance dissimilarity among data points[6]. Data points that are alike or close to each other are placed in the same group, while data objects that are unlike or far apart are placed in a different cluster. Clustering classifies data objects in the same way as classification does, however unlike classification, the class labels are unknown because clustering is based on unsupervised learning. Domain specialists investigate the behavior or attributes of the data items to define the clusters[7]. 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 from each other as much as possible.
- The distance/similarity measure must have some practical ability and be clear. clustering is also used in many application domains i.e. statistics, image segmentation,

pharmaceutical industry, object recognition, information retrieval, bioinformatics, etc [8].

The Figure 1.2 represents the procedure of the clustering algorithm. We can see the fruits are divided into different groups/clusters on the bases of there similarity or dissimilarity characteristics.

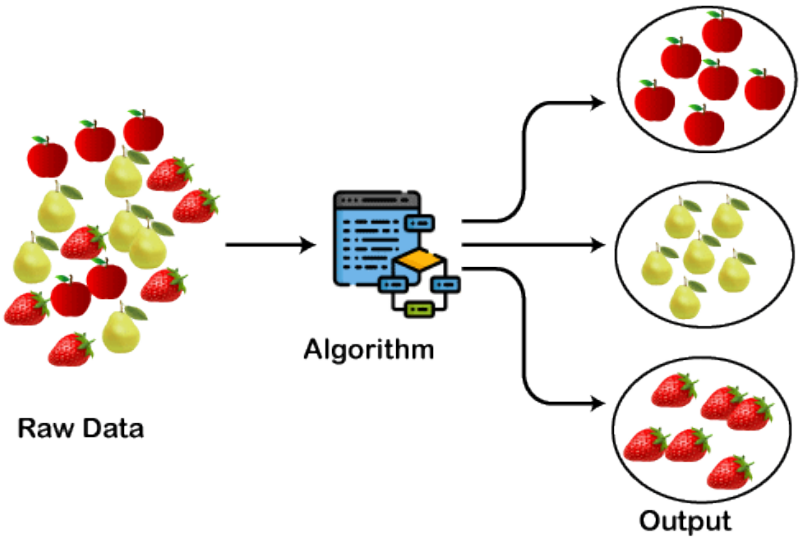


Figure 1.2: This figure illustrates the clustering procedure

1.5.1 Types of Clustering

Soft clustering and hard clustering are the two types of clustering[9]. Soft clustering is when all of the data points belong to just one cluster and hard clustering is when a data point belongs to multiple clusters.

There are various clustering algorithms, most commonly use of algorithm which are K-means clustering algorithm[10], Fuzzy c means clustering algorithm[11], Gaussian mixture model (GMM)[12], Hierarchical clustering (agglomerative and divisive algorithm)[13], Mixture of multivariate t distribution[14] and Density-based spatial clustering[15]. K-means, Hierarchical clustering, and Density-based spatial clustering is the type of hard clustering on

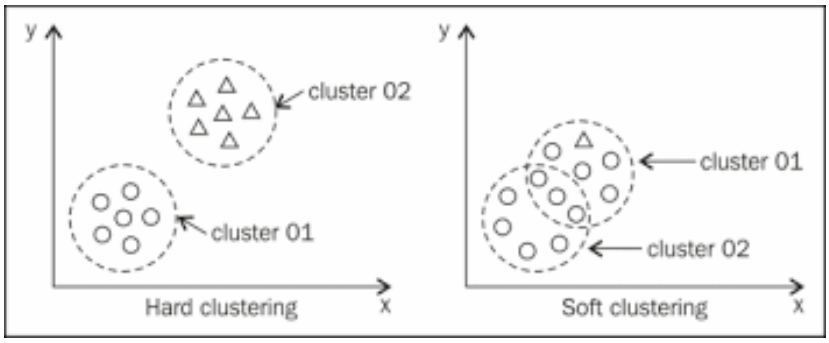


Figure 1.3: This figure represents the clustering types

the other hand GMM, the mixture of multivariate t distribution, and Fuzzy c mean is the type of soft clustering.

1.5.2 Application of Clustering

One of the uses of clustering applications that we covered in the thesis is chemical compounds with antibacterial activity in the medical and chemistry fields. Antibiotics/Antibacterial are the most essential weapons in the fight against microbial illnesses, and their introduction has had a huge impact on the health-related quality of human life. We used several clustering techniques for the distinct antibacterial activity portioning/grouping data sets.

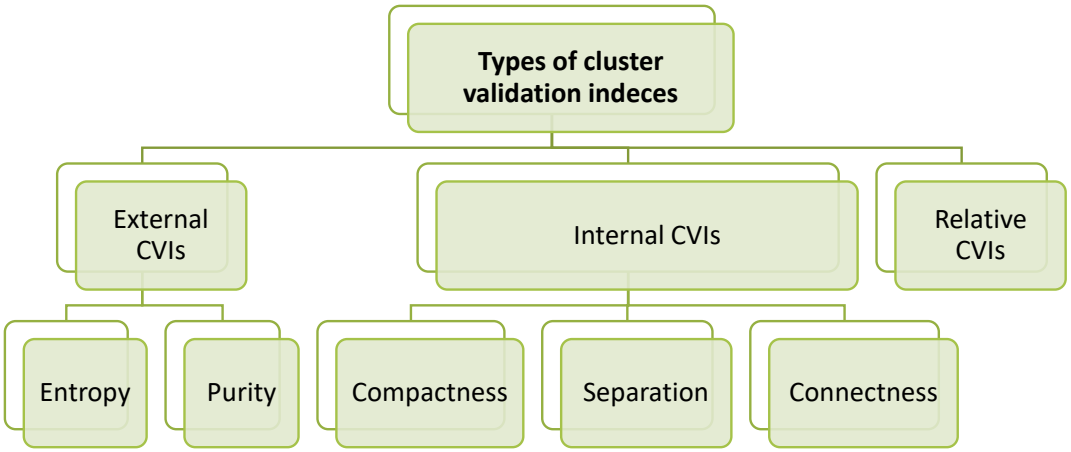
1.5.3 Clustering Algorithms

Centroid-based clusters which are also known as K-means clustering algorithm. Distribution-based clustering GMM and Mixture of multivariate t distribution, Connectivity-based clustering (Hierarchical clustering). Several indexes for verifying clustering analysis results and identifying which clustering algorithm performs best. The most important question is whether much optimal number of clusters is enough. To solve this problem, we will use different cluster validation techniques/indices.

1.6 Cluster Validation Indices/Technique

It is necessary to establish a technique/indices to validate the goodness of partitions of data points after clustering as an unsupervised learning method. In this thesis, we introduce the term "majority score clustering ". The "majority score clustering" rule depends on an individual decision of each CVI, where the final decision is made by the majority of the total CVI votes. This method provides quick results and adheres to a strict requirement of using separate CVI in the clustering algorithm validation procedure.

Figure 1.4: The diagram representation of cluster validation indices



1.6.1 Types of Cluster Validation Indices

The three main categories of clustering validation are external, internal, and relative[16]. The main difference between clustering validation is that evaluating the results of clustering algorithm based on prior information of data is called external validation, whereas internal validation does not. The example of internal validation is entropy which analyses the purity of data points of clusters based on the given class labels [17].

Relative cluster validation indices which measure the clustering structure by shifting parameter values for the same method varying the number of clusters k. It's commonly used to

figure out how many clusters are best.

1.6.2 Internal Clustering Validation Techniques/Indices

In the thesis, we apply the internal clustering validation techniques/indices because we have the class labels of chemical compounds of antibacterial activity data sets. Internal validation metrics frequently represent the cluster partitions, compactness, contentedness, and separation[18].

1.6.2.1 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 a cluster such as a cluster group-wise within average or median distances.

1.6.2.2 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.

1.6.2.3 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).

Chapter 3 explained the methodology and chapter 4 about data explanation. In chapter 5 results and discussion. Chapter 6 presents the conclusion and future scope.

1.7 Problem Statement

How do identify alike chemical compounds characteristics of antibacterial activity using different clustering algorithms on the basis of cluster validation indices (CVI)?

1.8 Research Objective

The research objective of the thesis is given as follows:

1. In clustering which algorithm satisfy the most cluster validation indices (CVI).
2. How many clusters are enough?
3. Which cluster algorithm is best?
4. Clusters of alike chemical compounds characteristics of antibacterial activity

Chapter 2

Review of Literature

In 1943 Walter Pitts and neuroscientist Warren McCulloch used the mathematical modeling of neural networks where machine learning originally emerged[19]. Arthur Samuel first proposed the concept of machine learning in 1959, describing it as a branch of research that allows computers to learn without being explicitly programmed. The act of creating computer systems that automatically get better with use and incorporate a learning process is referred to as machine learning (ML). Draw Inference, best to model fit or learning from examples are still ways to describe machine learning, which automatically learns the theory from data[1].

Research produced over a wide range of groups makes up cluster analysis, a primitive inquiry with little to no prior knowledge. On the one hand, diversity gives us a wide range of resources. Clustering techniques for data sets found in statistics, computer science, the medical field, and machine learning, and explain how they are used in the traveling salesman problem, several benchmark data set, and the emerging discipline of bioinformatics. Also covered are a number of closely related subjects including cluster validation and closeness measure[20].

UL (clustering) method is similar to exploratory data analysis and data mining. In UL there are only explanatory variables and no output observations in UL, because the data is unlabelled. UL is a method of learning where examples are based on their similarities and are automatically grouped into relevant groups. It introduces the in this paper while surveying current clustering algorithms, introduces essential unsupervised learning concepts. More-

over, Recent developments in unsupervised learning, including distributed clustering and ensembles of clustering algorithms are explained[21].

One of the most used and easy-to-understand data clustering algorithms is the K-means clustering technique, which divides data sets with "n" data points into "k" groups or clusters. The K-means grouping technique was first proposed by MacQueen in 1967[22], and Hartigan and Wong later improved it [23]. The convergence of the K-means algorithms and calculations was illustrated by Bottou and Bengio [24]. It has been said to be incredibly helpful for a collection of obvious rules and popular uses. Although the true K-means algorithm does not work with incomplete or outliers' data set. The computational time is less than another algorithm.

K. A. Abdul Nazeer et al. [25] propose an improved clustering algorithm to efficiently allocate data points to clusters by finding initial centroids. Soumi Ghosh et al. Improve the K-means algorithm's efficiency and accuracy[26]. On the basis of temporal complexity, a comparison between Fuzzy C-means and K-means. The K-means method appears to be superior to the Fuzzy C-means algorithm[27]. Shafeeq et al. suggested approach finds the number of clusters on the run is based on the cluster quality output. It is suitable for both known and unknown numbers of clusters in advance. Junatao Wang et al. [28] modified algorithm reduces the impact of noise data on the K-means algorithm, resulting in more accurate clustering results. Reduce the computational complexity of the K-means to improve clustering speed and accuracy[29].

A GMM is a parametric probability density function that may be written as the weighted sum of Normal component densities. GMM is frequently employed in biometric systems as a parametric model of the probability distribution of continuous measurements or data, for as vocal-tract-related spectral features in a speaker recognition system. The iterative Expectation-Maximization (EM) or Maximum posterior technique is used to measure the unknown parameters of GMM.[30].

EM algorithm introduced by Dempster, Laird, and Rubin in 1977[31]. After providing a brief historical overview of the EM algorithm, they look at methods for accelerating convergence while keeping the algorithm's stability and simplicity (e.g., automatic monotone convergence in likelihood). To make it easier to find effective data augmentation strategies

and, consequently, quick EM implementations, we first propose the concept of a "working parameter." Second, a summary of different current EM algorithm extensions. The paper's key finding is that it is feasible to create simple, dependable, and quick algorithms with the aid of statistical considerations[32].

Using a mathematical method based on finite mixes of distributions, modeling random occurrences using statistics [33]. Increasingly, continuous multivariate data sets are clustered using normal mixture models. In terms of their fitting posterior probability of membership of the mixture components corresponding to the clusters, they give a probabilistic (soft) clustering of the data. Following that, one may get an unequivocal (hard) clustering by the component to which each observation is assigned has the highest measured posterior likelihood of membership. Outliers in the statistics, however, can influence the parameter estimations in the inferred clustering and the normal component densities.

A stronger strategy is to multivariate t-distribution fit mixes, which have larger tails than the typical elements[34]. The algorithm of expectation-maximization (EM) may be used to fit t-distribution mixes using maximum likelihood. On an actual data set, the model's use to give a reliable technique for clustering is demonstrated. It is shown how the use of Less extreme estimates of the posterior probability of cluster memberships is provided by t-components.

The cluster validation problem is described as figuring out how many clusters there are in a data collection [35]. Cluster validation's major goal is to analyze clustering findings in order to discover the optimal partitioning of a data set. As a result, cluster validity methodologies are employed to quantify and assess a clustering algorithm's output. These methods have representative indicators known as validity indices. The conventional method for determining the maximum number of clusters is to execute the algorithm[36].

Traditional clustering algorithm classifications [37, 38, 39] discriminate largely between hierarchical, density-based approach, and partitioning techniques. A kind of different classification is utilized in this case, dependent on the clustering method (implicit or explicit) optimized by every algorithm. The relationship between clustering methods and cluster validations/indices provides a good understanding of them by using clustering criteria. These procedures have one of the core issues of clustering is capturing the intuitive concept of a cluster through any specific, formal description [40]. There are various legitimate features

that may be assigned to a good partitioning, but they are somewhat contradictory and often difficult to explain in terms of objective functions.

The compactness approach is often accomplished by limiting intra-cluster variance to a minimum [41]. In this article, the authors investigate whether connectedness is more validated for the clustering appro-depending on the assumption that data items or points that are close to each other should be clustered together. Density-based approaches [42]and famous methods such as k mean clustering are examples of algorithms that apply this notion[43]. They are well-suited for arbitrarily formed clusters, but they can lack resilience when the clusters are close together in space.

For the clustering methods, spatial separation or space is a criterion that, on its own, provides little direction and can simply lead to easy solutions. As a result, it is frequently paired with other objectives, most notable measurements of compactness or cluster size balance.[44, 45]. The literature has a variety of improved techniques that integrate metrics of the many sorts mentioned above. Combinations of compactness and separation are particularly popular in this regard, as the two types of measurements display opposite trends: while intra-cluster homogeneity improves as the clusters number increases, and then the distance between the clusters automatically decreases. As a result, some algorithms can measure both intra & inter clusters homogeneity and separation and provide the final results for the data [46, 47].

Chapter 3

Reference Methods

3.1 Clustering Algorithms

We outline the three clustering methods utilized in the thesis in this part.

3.1.1 K-means Clustering Algorithm

K-means clustering method was initially discussed by James MacQueen in 1967 [48], although the notion originated in 1957 by Hugo [49].
K-means is a partitional clustering technique. Data points are classified into non-overlapping categories. It is the most straightforward and practical way. It outperforms other algorithms. K-means clusters data points/objects based on distance from the cluster centroid. The aim of K-means clustering is to minimize the total intra-cluster variance, also known as the squared error function[50].

$$Y = \sum_{j=1}^{k} \sum_{i=1}^{n} ||Y_i^{(j)} - d_j||^2 \tag{3.1}$$

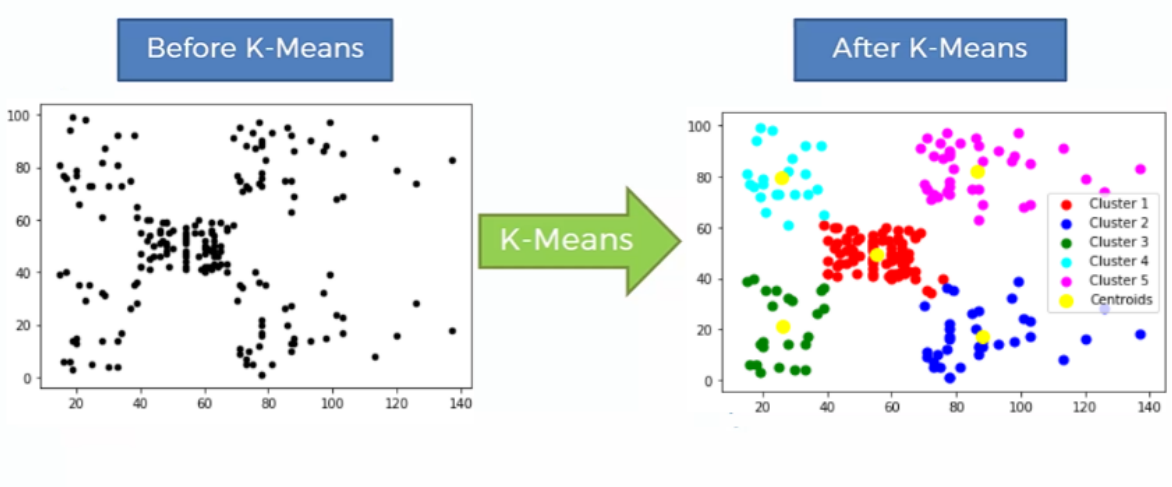
Where,

Y =objective function

k = number of clusters

n = number of data points

Figure 3.1: The diagram represent procedure of K-means clustering algorithm



3.1.1.1 Algorithm

There are five calculation steps of the K-means clustering algorithm are explained below:

1. Initially, k must be specified.
2. Choose k random sites as centroids.
3. Apply the Euclidean distance function to assign data points to their nearest cluster center.
4. Determine the center or mean of all the items in the cluster.
5. Until, the K-means approaches the convergence criteria, repeat steps 2, 3, and 4.

3.1.1.2 Benefits

1. Relatively easy to put into practice.
2. Scale able to big data sets Convergence is ensured.
3. Can warm up the locations of the centroids.
4. Adapts well to new examples.

5. Generalizes to circular clusters and other clusters of various sizes and forms.

3.1.1.3 Limitations

1. Predefine the number of clusters (k) ahead of time.
2. For noisy data or outliers it is not applicable.
3. Difficult to identify clusters with non-convex geometries
4. Curse of dimensionality.

3.1.2 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. GMM is the type of soft Clustering. Each component of the cluster is considered a model with mean and variance. Mixture models are to estimate the parameters of the probability distribution for each cluster like mean and variance.

In the actual world, complex data collection is often made up of a number of stochastic processes. As a result, a single Gaussian distribution cannot fit such data. A GMM, on the other hand, is used to describe a combination of j Gaussian distributions. Assume we have a data collection of N independent data points $y = y_1, y_2 \ldots y_j \ldots y_N$ with various peaks. A GMM may be used to model this data collection. The Pdf of GMM is:

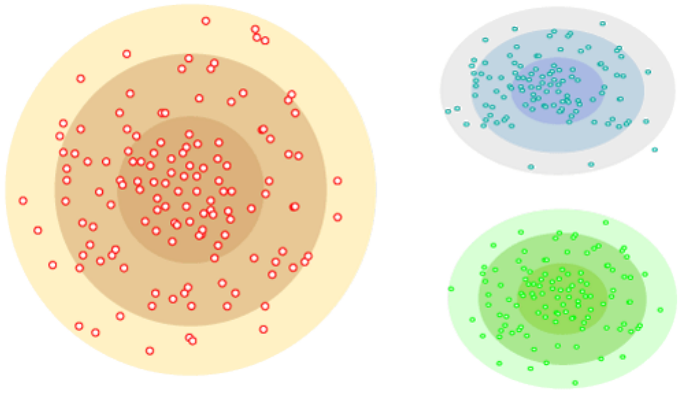
$$p(y \mid \Theta) = \sum_j \pi_j N\left(y; \mu_j, \sigma_j\right) = \sum_j \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{\left(y-\mu_j\right)^2}{2\sigma_j^2}\right] \tag{3.2}$$

In this model: Θ is the parameters, π_j is the prior probability of the j^{th} Gaussian model (GM), and

$$\sum_j \pi_j = 1$$

The Figure 3.2 explained the clustering of GMM through EM algorithm.

Figure 3.2: This figure representation of GMM of EM algorithm



3.1.2.1 Expectation-Maximization (EM)

An iterative supervised training approach is the expectation-maximization (EM) algorithm. We have a data collection of N independent data points, and y is the task. We can infer or know for sure that the data set consists of a combination of j Gaussian distributions. To estimate the GMM parameters is the task: j set of $\left(\pi_j, \mu_j, \sigma_j\right)$

Likelihood Function(LF):

Maximum-likelihood estimation (MLE) is a highly popular and simple approach for doing an estimate for issues based on data sets of independent samples. The product of the probabilities of each test is used to express the likelihood of N independent tests. The probability function is what we refer to as:

$$p(y \mid \Theta) = \prod_j p\left(y_i \mid \Theta\right) \tag{3.3}$$

The likelihood function is to be maximized using MLE in order to estimate the parameters $Theta$.

$$\hat{\Theta} = \arg\max_{\Theta} \prod_j p\left(y_i \mid \Theta\right) \tag{3.4}$$

The likelihood function for single-variable Gaussian mixture models is quite complex when the MLE is used.

$$p(y \mid \Theta) = \prod_j p\left(y_i \mid \Theta\right) = \prod_j \left[\sum_j \pi_j N\left(y_i \mid \mu_j, \sigma_j\right)\right] \tag{3.5}$$

It is difficult to estimate a set of Gaussian parameters directly and openly. The EM algorithm reduces GMM's likelihood function and provides an iterative method for optimizing estimates. We will try to discuss the EM technique for GMM parameter estimation succinctly here.

First, choosing the log-likelihood function will simplify the likelihood function of a GMM model. It is simpler to separate independent data samples and compute parameter derivatives with a formula that has a summation form.

$$L(y \mid \Theta) = \sum_j \ln\left[p\left(y_i \mid \mu_j, \sigma_j\right)\right] = \sum_j \ln\left[\sum_j \pi_j N\left(y_i \mid \mu_j, \sigma_j\right)\right] \tag{3.6}$$

Latent Parameters:

There is no efficient technique to maximize the log-likelihood function for GMM above equation. The EM method includes a latent parameter α, which has the value $\alpha \in 1, 2 \ldots j \ldots j$. Particular complete GMM parameters, this is used to represent the likelihood of a given training sample y_j belonging to cluster α:

Complete GMM parameters $p\left(\alpha \mid y_i, \mu_j, \sigma_j\right)$ In the probability distribution of y_i include the latent parameter α.

$$p\left(y_i \mid \Theta\right) = \sum_j p\left(y_i \mid \alpha = j, \mu_j, \sigma_j\right) p(\alpha = j) \tag{3.7}$$

Compared with $p(y \mid \Theta) = \sum_j \alpha_j N\left(y \mid \mu_j, \sigma_j\right)$, we can demonstrate that α_j is the prior probability of $p(\alpha = j)$.

$$p(\alpha = j) = \pi_j$$

and the conditional probability of y given $\alpha = j$ is the j^{th} GM.

$$p\left(y_i \mid \alpha = j, \mu_j, \sigma_j\right) = N\left(y_i; \mu_j, \sigma_j\right) \tag{3.8}$$

The log likelihood function can now include the latent parameter. In order to fit the form of Jensen's inequality, a superfluous term $p\left(\alpha \mid y_i, \mu_j, \sigma_j\right)$ is added.

$$\begin{aligned} L(y \mid \Theta) &= \sum_j \ln\left[p\left(y_i, \alpha \mid \mu_j, \sigma_j\right)\right] \\ &= \sum_i \ln \sum_j p\left(y_i \mid \alpha = j, \mu_j, \sigma_j\right) p(\alpha = j) \end{aligned}$$

$$= \sum_i \ln \sum_j p\left(\alpha = j \mid y_i, \mu_j, \sigma_j\right) \frac{p\left(y_i \mid \alpha = j, \mu_j, \sigma_j\right) p(\alpha = j)}{p\left(\alpha = j \mid y_i, \mu_j, \sigma_j\right)} \tag{3.9}$$

Generalize the Lf:

It is challenging to maximize a log function due to the summing. Here, consider Jensen's inequality:

$$f[E(y)] \geq E[f(y)]$$

Let v represent $\frac{p\left(y_i|\alpha=j,\mu_j,\sigma_j\right)p(\alpha=j)}{p\left(\alpha|y_i,\mu_j,\sigma_j\right)}$ to match Jensen's inequality. We get

$$f(v) = \ln v$$

$$E(v) = \sum_j p\left(\alpha \mid y_j, \mu_j, \sigma_j\right) v$$

Therefore,

$$L(y \mid \Theta) \geq \sum_i \sum_j p\left(\alpha = j \mid y_i, \mu_j, \sigma_j\right) \ln \frac{p\left(y_i \mid \alpha = j, \mu_j, \sigma_j\right) p(\alpha = j)}{p\left(\alpha = j \mid y_i, \mu_j, \sigma_j\right)} \tag{3.10}$$

The posterior probability can be derived by Bayes' law.

$$\begin{aligned} p\left(\alpha = k \mid y_i, \mu_j, \sigma_j\right) &= \frac{p\left(y_i \mid \alpha = j, \mu_j, \sigma_j\right)}{\sum_j p\left(y_i \mid \alpha = j, \mu_j, \sigma_j\right)} \\ &= \frac{\pi_j N\left(y_i \mid \mu_j, \sigma_j\right)}{\sum_j \pi_j N\left(y_i \mid \mu_j, \sigma_j\right)} \end{aligned}$$

define $\omega_{i,j} = p\left(\alpha = j \mid y_i, \mu_j, \sigma_j\right) = \frac{\pi_j N\left(y_i|\mu_j,\sigma_j\right)}{\sum_j \pi_j N\left(y_i|\mu_j,\sigma_j\right)}$

Then

$$\begin{aligned} L(y \mid \Theta) &= \sum_i \ln \sum_j \omega_{i,j} \frac{\pi_j N\left(y_j \mid \mu_j, \sigma_j\right)}{\omega_{i,j}} \\ &\geq \sum_i \sum_j \omega_{i,j} \ln \frac{\pi_j N\left(y_i \mid \mu_j, \sigma_j\right)}{\omega_{i,j}} \end{aligned} \tag{3.11}$$

The log LF lower bound is defined by this equation. Consequently, a stated iterative target function for the EM algorithm:

$$Q\left(\Theta,\Theta^{t}\right)=\sum_{i}\sum_{j}\omega_{i,j}^{t}\ln\frac{\pi_{j}N\left(y_{i}\mid\mu_{j},\sigma_{j}\right)}{\omega_{i,j}^{t}} \tag{3.12}$$

Apply the most recent latent parameters in $Q\left(\Theta,\Theta^{t}\right)$ and then we may update Θ^{t}, via maximization. After t iterations, we have Θ^{t} and the latent $\omega_{i,j}^{t}$.

iterative Optimization:

Ω and Θ are adjusted repeatedly after the initialization of the Θ parameter.

- A set of parameters Θ^{t} have been reached after iteration t

- By apply Θ^{t} to the GMM determine the latent parameters $\omega_{i,j}^{t}$. It is known as the expectation step.

$$\omega_{i,j}^{t}=\frac{\pi_{j}N\left(y_{i}\mid\mu_{j},\sigma_{j}\right)}{\sum_{j}\pi_{j}N\left(y_{i}\mid\mu_{j},\sigma_{j}\right)} \tag{3.13}$$

- Apply the latest latent parameters $\omega_{i,\,\mathrm{a}}^{t}$ in the target function. By using Jensen's inequality to simplify the log-likelihood function, the target function is created.

$$Q\left(\Theta,\Theta^{t}\right)=\sum_{i}\sum_{j}\omega_{i,j}^{t}\ln\frac{\pi_{j}N\left(y_{i}\mid\mu_{j},\sigma_{j}\right)}{\omega_{i,j}^{t}} \tag{3.14}$$

- With $\omega_{i,j}$, for the purpose of updating the GMM parameters Θ^{t+1}, the target log-likelihood function should be maximized. This step is known as the maximization step maximizing phase is the next stage.

$$\hat{\Theta}^{t+1}=\arg\max_{\Theta}\sum_{i}\sum_{j}\ln\omega_{i,j}^{t}\frac{\pi_{j}N\left(y_{i}\mid\mu_{j},\sigma_{j}\right)}{\omega_{i,j}^{t}} \tag{3.15}$$

3.1.2.2 EM Algorithm for GMM

The EM target function for a GMM has the following full form:

$$Q\left(\Theta,\Theta^{t}\right)=\sum_{i}\sum_{j}\omega_{i,j}^{t}\ln\frac{\pi_{j}}{\omega_{i,j}^{t}\sqrt{2\pi\sigma_{j}^{2}}}\exp\left[-\frac{\left(y_{i}-\mu_{j}\right)^{2}}{2\sigma_{j}^{2}}\right] \tag{3.16}$$

E-Step:

The E-step entails estimating the latent parameters on j Gaussian models for each training

sample. Therefore, the N by K matrix represents the latent parameter ω.

The most recent Gaussian parameters $\left(\pi_j, \mu_j, \sigma_j\right)$ are used to compute $\omega_{i,j}$ for each iteration

$$\omega_{i,j}^t = \frac{\pi_j^t N\left(y_i \mid \mu_j^t, \sigma_j^t\right)}{\sum_j \pi_j^t N\left(y_i \mid \mu_j^t, \sigma_j^t\right)} \tag{3.17}$$

M-Step:

$$\hat{\Theta} = \arg\max_{\Theta} Q\left(\Theta, \Theta^t\right)$$

To clearly separate elements, the target likelihood function might be increased.

$$Q\left(\Theta, \Theta^t\right) = \sum_i \sum_j \omega_{i,j}^t \left(\ln \pi_j - \ln \omega_{i,j}^t - \ln \sqrt{2\pi\sigma_j^2} - \frac{\left(y_i - \mu_j\right)^2}{2\sigma_j^2}\right) \tag{3.18}$$

Update π_j:

As defined in GMM, π_j is constrained by $\sum_j \alpha_j = 1$, A restricted optimization issue, then, is estimating π_j

$$\hat{\pi}_j^{t+1} = \arg\max_{\pi_j} \sum_i \sum_j \omega_{i,j}^t \ln \pi_j$$
$$\text{subject to } \sum_j \pi_j = 1$$

To locate the local maxima of such a restricted optimization problem, Lagrange multipliers are utilized. A Lagrange function can be created as follows:

$$\mathscr{L}\left(\pi_j, \lambda\right) = \sum_i \sum_j \omega_{i,j}^t \ln \pi_j + \lambda \left[\sum_j \pi_j - 1\right]$$

The local maxima π_j^{t+1} should cause the Lagrangean function's derivative to equal zero. Hence,

$$\frac{\partial \mathscr{L}\left(\pi_j, \lambda\right)}{\partial \pi_j} = \sum_i \omega_{i,j}^t \frac{1}{\pi_j} + \lambda = 0$$
$$\pi_j = -\frac{\sum_i \omega_{i,j}^t}{\lambda}$$

The value of λ may be found by adding the equations for all j.

$$\sum_j \pi_j = -\frac{\sum_i \sum_j \omega_{i,j}^t}{\lambda}$$

$$1 = -\sum_i \frac{1}{\lambda} = -\frac{N}{\lambda}$$
$$\lambda = -N$$

As a result, π_j is updated on iteration $t+1$ depending on latent parameters on iteration t.

$$\pi_j^{j+1} = \frac{\sum_i \omega_{i,j}^t}{N} \tag{3.19}$$

Update μ_j: μ_j is unconstrained and can be derived by taking the derivative of the target likelihood function.

$$\hat{\mu}_j^{t+1} = \arg\max_{\mu_j} Q\left(\Theta, \Theta^t\right)$$

Let $\frac{\partial Q(\Theta,\Theta^t)}{\partial \mu_j} = 0$, hence

$$\frac{\partial \sum_i \sum_j \omega_{i,j}^t \left(\ln \pi_j - \ln \omega_{i,j}^t - \ln \sqrt{2\pi\sigma_j^2} - \frac{\left(y_i - \mu_j\right)^2}{2\sigma_j^2} \right)}{\partial \mu_j} = 0$$
$$\sum_i \omega_{i,j}^t \frac{y_i - \mu_j}{\sigma_j^2} = 0$$
$$\Rightarrow \sum_i \omega_{i,j}^t \mu_j = \sum_i \omega_{i,j}^t y_i$$
$$\Rightarrow \mu_j \sum_i \omega_{i,j}^t = \sum_i \omega_{i,j}^t y_i$$

Hence μ_j on iteration $t+1$ can be updated as a form of weighted mean of y.

$$\mu_j^{t+1} = \frac{\sum_i \omega_{i,j}^t y_i}{\sum_i \omega_{i,j}^t} \tag{3.20}$$

Update σ_j: Similarly, the derivative of the target likelihood function with respect to σ_j is used to calculate the updated σ_j.

$$\hat{\sigma}_j^{t+1} = \arg\max_{\sigma_j} Q\left(\Theta, \Theta^t\right)$$

Let

$$\frac{\partial Q\left(\Theta, \Theta^t\right)}{\partial \sigma_j} = \frac{\partial \sum_i \sum_j \omega_{i,j}^t \left(\ln \pi_j - \ln \omega_{i,j}^t - \ln \sqrt{2\pi\sigma_j^2} - \frac{\left(y_i - \mu_j\right)^2}{2\sigma_k^2} \right)}{\partial \sigma_j} = 0.$$

We get $\sum_i \omega_{i,j} \left[-\frac{1}{\sigma_j} + \frac{\left(y_i - \mu_j\right)^2}{\sigma_j^3} \right] = 0$

$\sum_i \omega_{i,j} \sigma_j^2 = \sum_i \omega_{i,j} \left(y_i - \mu_j\right)^2 \Rightarrow \sigma_j^2 \sum_i \omega_{i,j} = \sum_i \omega_{i,j} \left(y_i - \mu_j\right)^2$

For σ_j, we can update σ_j^2, this is sufficient for calculating the Gaussian model. Because new sigma a_j^2 is dependent on μ_j, μ_j^{t+1} is generally computed first and then added to the update equation for σ_j^2.

$$\left(\sigma_j^2\right)^{t+1} = \frac{\sum_i \omega_{i,j} \left(x_i - \mu_j^{t+1}\right)^2}{\sum_i \omega_{i,j}} \tag{3.21}$$

3.1.2.3 Benefit of GMM

1. Using probability metrics, which are simple to comprehend, the associativity of a data point to a cluster is measured.
2. Accurate for real-time data sets as demonstrated.
3. Some GMM implementations provide mixed membership of the data points, making them a viable alternative to Fuzzy C Means for fuzzy clustering.

3.1.2.4 Limitations of GMM

1. 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.

3.1.3 Mixtures of Multivariate t Distributions Clustering Algorithm

The multivariate-t distribution mixes presume that each sub population of the observed data follows the multivariate-t distribution [51].

$$P(y_i|\theta_j) = P(y_i|\mu_j, \Sigma_j; \nu) = \frac{\Gamma((\nu+p)/2)|\sigma_j|^{-1/2}}{\Gamma(1/2)\Gamma(1/2)\nu^{p/2}} * \frac{1}{[1+(\delta(y_i,\mu_j;\Sigma_j)/\nu)]^{(\nu+p)/2}} \tag{3.22}$$

Calculation steps for Mixture of multivariate t distribution given below:

The distance between y_i and μ_j squared as defined by Mahalanobis is given below:

$$\delta(y_i, \mu_j; \Sigma_j) = (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j) \tag{3.23}$$

where μ_j is a random scalar produced using a gamma distribution The gamma distribution is represented as

$$\gamma(\mu_j, \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma\alpha} * \mu_j^{\alpha-1} exp(-\mu_j|\beta) \quad \mu_j, \alpha, \beta > 0 \tag{3.24}$$

We suppose that, according to a combination of multivariate t-distributions the observed sample N points are generated at randomly, i.i.d with the mixing percentage π_j. The log-likelihood of sample N points is

$$L = log(\Pi_{i=1}^N (\sum_{j=1}^g \pi_j P(y_i|\theta_j))) \tag{3.25}$$

$$= \sum_{i=1}^N log(\sum_{j=1}^g \pi_j P(y_i|\theta_j)) \tag{3.26}$$

3.1.3.1 EM for Mixtures of Multivariate t Distribution Clustering Algorithm

Exception Step

Determine the Q function: $Q(\theta) = E_{f(y_j)} \left[\log p\left(y_i \mid y_j; \theta\right)\right]$ with:

$$f\left(y_{\mathrm{j}}\right) = p\left(y_j \mid y_i; \theta^{(t)}\right)$$

$$f\left(y_{\mathrm{j}}\right) = \frac{p\left(y_{\mathrm{j}}; \theta^{(t)}\right) p\left(y_i \mid y_{\mathrm{j}}; \theta^{(t)}\right)}{\int p\left(y_{\mathrm{j}}; \theta^{(t)}\right) p\left(y_i \mid y_{\mathrm{j}}; \theta^{(t)}\right) \mathrm{d}y_j}, \tag{3.27}$$

Maximization Step

$$\begin{aligned} p\left(y_i \mid \pi_{ij} = 1, u_i, \mu_j, \Sigma, \nu\right) &\sim N\left(\mu_j, \Sigma/u_i\right) \\ &= \frac{1}{(2\pi/u_i)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{u_i}{2} \cdot \delta\left(y_i, \mu_j; \Sigma\right)\right) \\ p\left(u_i \mid w\right) &= \gamma\left(\frac{w}{2}, \frac{2}{w}\right) = \frac{1}{(2/w)^{w/2}\Gamma(w/2)} \\ &\quad \times u_i^{w/2-1} \exp\left(-\frac{w}{2} u_i\right), \quad u_i > 0 \end{aligned}$$

$$
\begin{aligned}
=&E_{\pi,u}\left[\sum_{i=1}^{N}\sum_{j=1}^{g}\pi_{ij}\left[\log\left(p\left(y_i \mid u_i,\pi_{ij},\mu_j,\Sigma_j,\nu\right)p\left(\pi_{ij}\right)\right.\right.\right.\\
&\left.\left.\left.\times p\left(u_i \mid w\right)\right)\right]\right]\\
=&E_{\pi,u}\left[\sum_{i=1}^{N}\sum_{j=1}^{g}\pi_{ij}\left[-\frac{u_i}{2}\delta\left(y_i,\mu_j;\Sigma_j\right)-\frac{1}{2}\log\left|\Sigma_j\right|+\log\pi_j\right.\right.\\
&\left.\left.+\frac{w}{2}\left(\log u_i-u_i+\log\frac{w}{2}\right)-\log\Gamma(w/2)+C\right]\right]
\end{aligned}
$$

where the expectation is in terms of the missing data probability $f(z,u)$, and C is a constant that is independent of the parameters (and hence has no effect on the maximization step). The component j contributes to $Q(\theta)$, this is ensured by the z_{ij} multiplication.

In zij, this equation is obviously linear. To establish linearity in u_i as well, which will substantially simplify the method, we use a new parameter set: $\theta' \equiv \left\{\pi_{1\ldots g},\mu_{1\ldots g},\Sigma_{1\ldots g}\right\}$, assuming that v is a known constant. We now have

$$
\begin{aligned}
Q\left(\theta'\right) =&E_{\pi,u}\left[\sum_{i=1}^{N}\sum_{j=1}^{g}\pi_{ij}\left[-\left(\frac{u_i}{2}\left(y_i-\mu_j\right)^{\mathrm{T}}\Sigma_j^{-1}\left(y_i,\mu_j\right)\right)\right.\right.\\
&\left.\left.-\frac{1}{2}\log\left|\Sigma_j\right|+\log\pi_j+C'\right]\right]\\
&\pi_j u_i^{w/2-1}\exp\left(-\frac{w}{2}u_i\right)\cdot\left[\frac{\exp\left(-\left(u_i/2\right)\cdot\delta\left(y_i,\mu_j;\Sigma\right)\right)}{\left|\Sigma_j\right|^{1/2}\left(2\pi/u_i\right)^{p/2}}\right]\\
&\propto \pi_j\left|\Sigma_j\right|^{-1/2}u_i^{(\nu+p)/2-1}\exp\left(-u_i\frac{\delta\left(y_i,\mu_j;\Sigma\right)+\nu}{2}\right).
\end{aligned}
$$

The expectation is then calculated using the following integral:

$$
\int_0^\infty u_i^e\exp\left(-u_i/f\right)\mathrm{d}u_i=\Gamma(e+1)\cdot f^{e+1}.
$$

We finally obtain

$$
\begin{aligned}
\hat{\pi}_{ij} &= E_{f\left(y_{\mathrm{j}}\right)}\left(\pi_{ij}\right)\\
&= \frac{\pi_j\left|\Sigma_j\right|^{-1/2}\left(\delta\left(y_i,\mu_j;\Sigma_j\right)+\nu\right)^{-(p+\nu)/2}}{\sum_{j=1}^{g}\pi_j\left|\Sigma_j\right|^{-1/2}\left(\delta\left(y_i,\mu_j;\Sigma_j\right)+\nu\right)^{-(p+\nu)/2}},
\end{aligned}
$$

$$E_{f(y_j)}\left(\pi_{ij}u_i\right) = \frac{\pi_j\left|\Sigma_j\right|^{-1/2}\left(\delta\left(y_i,\mu_j;\Sigma_j\right)+\nu\right)^{-(p+\nu)/2}\left((p+\nu)/\left(\delta\left(y_i,\mu_j;\Sigma_j\right)+\nu\right)\right)}{\sum_{j=1}^{g}\pi_j\left|\Sigma_j\right|^{-1/2}\left(\delta\left(y_i,\mu_j;\Sigma_j\right)+\nu\right)^{-(p+\nu)/2}}$$
$$= \hat{z}_{ij}\hat{u}_{ij}$$

Here

$$\hat{u}_{ij} = \frac{p+\nu}{\left(\delta\left(y_i,\mu_j;\Sigma_j\right)+\nu\right)} \tag{3.28}$$

M Step:

keeping the limitation $\sum_{j=1}^{g}\pi_j = 1$, the parameters we find, should maximizes $Q(\theta)$. Lets use a Lagrange multiplier λ as:

$$\frac{\mathrm{d}}{\mathrm{d}\pi_j}\left[Q-\lambda\left(\sum_j\pi_j-1\right)\right] = \sum_{i=1}^{N}\frac{\pi_{ij}}{\pi_j}-\lambda = 0$$

This produces the expected effect:

$$\pi_j = \sum_{i=1}^{N}\frac{\hat{pi}_{ij}}{N} \tag{3.29}$$

Means:

$$\frac{\mathrm{d}}{\mathrm{d}\mu_j}[Q] = \sum_{i=1}^{N}\pi_{ij}u_{ij}\Sigma_j^{-1}\left(y_i-\mu_j\right) = 0$$

Then we have

$$\mu_j^{(j)} = \frac{\sum_{i=1}^{N}\hat{\pi}_{ij}\hat{u}_{ij}y_i}{\sum_{i=1}^{N}\hat{\pi}_{ij}\hat{u}_{ij}} \tag{3.30}$$

Covariances:

$$\frac{\mathrm{d}}{\mathrm{d}\left(\Sigma_j^{-1}\right)}[Q] = -\frac{1}{2}\sum_{i=1}^{N}\pi_{ij}\left(\Sigma_j - u_i\left(y_i-\mu_j^{(j)}\right)\left(y_i-\mu_j^{(j)}\right)^{\mathrm{T}}\right)$$
$$= 0$$

which gives

$$\Sigma_j^{(j)} = \frac{\sum_{i=1}^{N}\left(\hat{\pi}_{ij}\hat{u}_{ij}\right)\left(y_i-\mu_j^{(k)}\right)\left(y_i-\mu_j^{(k)}\right)^{\mathrm{T}}}{\sum_{i=1}^{N}\hat{\pi}_{ij}} \tag{3.31}$$

3.1.3.2 Benefits of Mixtures of Multivariate t Distribution Clustering Algorithm

1. It has a long tail distribution.
2. It has a parameter for adjusting its robustness.
3. It provides less extreme estimates of the component membership posterior probability.

3.1.3.3 Limitations of Mixtures of Multivariate t Distribution Clustering Algorithm

1. It converges slowly.
2. It just reaches the local optimum.
3. It accounts for both forward and backward probability. This is in contrast to numerical optimization, which solely takes forward probabilities into account.

3.1.4 Clustering Validation Indices (CVI)

External and Internal validation are two types of CVI. In this thesis, we applied external validation techniques [52, 53].

3.1.4.1 Separation (WSS)

Cluster Separation: How distinct or well isolated from one another in a cluster is measured using the within-sum-of-squares method. Elbow curve is another name for WSS [54].To find out the maximum number of clusters that may be constructed for a certain data set, within-sum-of-squares is used as a metric. The squared distance between each cluster member and its centroid is added to create the WSS.

$$WSS = \sum_{i=1}^{n} (y_i - c_i^2) \tag{3.32}$$

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

3.1.4.2 Silhouette Width

The silhouette width measures how alike a data point its own cluster and compared to other clusters (separation). The range of the Silhouette width or index is between +1 and -1. if the value of SW is close to +1,this signifies that the data points are correctly classify into cluster.If the value of SW is nearby to -1 it indicate that the data points are correctly classify into cluster. [55].

silhouette width is defined as follows:

$$s_i = \frac{x_i - y_i}{\max(x_i, y_i)} \tag{3.33}$$

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

$$x_i = \frac{1}{|c_i| - 1} \sum_{j \in c_i, j \neq i} d(i, j) \quad and \quad y_i = \min_{i \neq j} \frac{1}{|c_i|} \sum_{j \in c_i} d(i, j) \tag{3.34}$$

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

3.1.4.3 Connectivity

In the data space, connectivity measures how closely entities are clustered with their closest neighbors. The connection should be kept to a minimum because its value ranges from 0 to infinity.[56].

Most internal clustering validation methods often incorporate compactness and separation metrics as shown below:

$$index = \frac{(a * separation)}{(b * Compactness)} \tag{3.35}$$

Where a and b weight.

3.1.4.4 Dunn Index (DI)

J. C. Dunn created DI in 1974, which is another method for validating clusters. The DI compute 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[55].
To calculate DI:

$$D = \frac{min.separation}{max.diameter} \tag{3.36}$$

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

3.1.5 Proposed Method: Majority Score Clustering Algorithm

It has been observed several cluster validation indices measure provides a different optimal number of clusters, which result of the clustering algorithm is not reliable. Instead of selecting an optimal cluster based on individual clustering validation indices (CVI), we have introduced a majority scoring clustering algorithm where the clustering algorithm is selected if it is satisfied by more than two or more combinations of clustering validation indices (CVI). We group/cluster the chemical compounds of antibacterial activity on the bases of the majority score clustering algorithm.

Chapter 4

Data Explanation and Statistical Software

4.1 Data Source

The source of data used in this paper is given below:

1. Dhinoja, Karia ,Shah, V. D. A. (2014). Acid Promoted One Pot Synthesis of Some New Coumarinyl 3,4'-Bipyrazole and Their In Vitro Antimicrobial Evaluation. Chemistry & Biology Interface, 4(4), 232–245.
   https://www.researchgate.net/publication/268502562.

2. Sankappa Rai, U., Isloor, A. M., Shetty, P., Isloor, N., Padaki, M., & Fun, H. K. (2011). A novel series of homoallylic amines as potential antimicrobials. Medicinal Chemistry Research, 21(7), 1090–1097.
   https://doi.org/10.1007/s00044-011-9607-3.

3. Shruthi, N., Poojary, B., Kumar, V., Hussain, M. M., Rai, V. M., Pai, V. R., Bhat, M., & Revannasiddappa, B. C. (2016). Novel benzimidazole–oxadiazole hybrid molecules as promising antimicrobial agents. RSC Advances, 6(10), 8303–8316.
   https://doi.org/10.1039/c5ra23282a.

In 1st data set, we have 6 variables which are four bacteria (P. Aeruginosa, E. Coli, S. Aureus, S. Pyogenes) and two Fungi (C. Albicans, As. Fumigatus). The Broth Dilution procedure

used to find the yield of Antimicrobial evaluation of all synthesized compounds.

The content of Table 4.1 is R, R', M.F, M.W (g/mole), M.P $^{\circ}C$, yield and R_f. The letters R and R' stand for various molecules or substances. Molecular formula (M.F), is an equation that specifies the quantity and kind of atoms that make up a substance's molecules. Molecular weight is denoted by M.W(g/mole) or gram per mole. The total atomic weights of the atoms in a molecule are measured by its molecular weight. M.W is measured in grams per mole. M.P $^{\circ}C$ stands for melting point. The typical definition of the melting point is the temperature at which a substance transforms from a solid to a liquid. The scale of M.P used in this data is Celsius or $^{\circ}C$. The R_f value is the ratio of the compound's distance to the solvent's distance. Ratio-to-front or retardation factor are both represented by the sign R_f . It has a decimal fractional form and it is always between 0 to 1.

In 2nd data set we have four variables which are four bacteria (E. Coli, B. Subtilis, S.

Table 4.1: The table of the chemical composition against bacteria's evaluation

Entry as	R	R'	M.F.	M.W. (g/ mole)	M.P. °C	Yield %	R_f
5a	H	H	$C_{27}H_{20}N_4O_3$	448.47	200-202	88	0.50
5b	H	4-CH_3	$C_{28}H_{22}N_4O_3$	462.50	210-212	90	0.54
5c	H	4-Cl	$C_{27}H_{19}ClN_4O_3$	482.92	232-234	80	0.56
5d	H	4-F	$C_{27}H_{19}FN_4O_3$	466.46	194-196	83	0.52
5e	H	4-Br	$C_{27}H_{19}BrN_4O_3$	527.37	206-208	81	0.55
5f	H	4-NO_2	$C_{27}H_{19}N_5O_5$	493.47	212-214	84	0.53
5g	H	3-NO_2	$C_{27}H_{19}N_5O_5$	493.47	224-226	85	0.54
5h	H	3-OH	$C_{27}H_{20}N_4O_4$	464.47	216-218	80	0.42
5i	H	2-OCH_3	$C_{28}H_{22}N_4O_4$	478.50	201-203	89	0.53
5j	7, 8-di Me	H	$C_{29}H_{24}N_4O_3$	476.53	218-220	82	0.50
5k	7, 8-di Me	4-CH_3	$C_{30}H_{26}N_4O_3$	490.55	226-228	88	0.51
5l	7, 8-di Me	4-Cl	$C_{29}H_{23}ClN_4O_3$	510.97	230-232	87	0.53
5m	7, 8-di Me	4-F	$C_{29}H_{23}FN_4O_3$	494.52	236-238	84	0.52
5n	7, 8-di Me	4-Br	$C_{29}H_{23}BrN_4O_3$	555.42	246-248	81	0.51
5o	7, 8-di Me	4-NO_2	$C_{29}H_{23}N_5O_5$	521.52	238-240	83	0.50
5p	7, 8-di Me	3-NO_2	$C_{29}H_{23}N_5O_5$	521.52	242-244	80	0.54
5q	7, 8-di Me	3-OH	$C_{29}H_{24}N_4O_4$	492.53	228-230	81	0.45
5r	7, 8-di Me	2-OCH_3	$C_{30}H_{26}N_4O_4$	506.55	198-200	90	0.53

Aureus, P. Aeruginosa). The serial dilution method is used to find the yield of MIC of

antibacterials.

The content of Table 4.2 is R, R^1, Reaction time (min) and Yield %. The letters R and R^1 stand for various molecules or substance. Reaction time (min) represent that, how much time require for reaction of chemical in minutes. A yield % is a ratio of moles of product to moles of reactant used to calculate the efficiency of a chemical process. Often stated as a percentage.

In 3rd data set, we also have six variables like data set one and also the same four bacteria

Table 4.2: The table of the different chemical compositions against MIC of Bacteria

S. no	R	R^1	Reaction time (min)	Yield (%)
3a	Benzaldehyde	Napthylamine	30	95
3b	Benzofuran-2-aldhyde	3,4-Diflurobenzylamine	35	90
3c	Cyclopropanecarboxaldehyde	4-t-Butylaniline	30	98
3d	2,4-Difluorobenzaldehyde	2,4,5-Trifluoroaniline	45	89
3e	Benzofuran-2-aldhyde	Napthylamine	30	90
3f	2,4-Difluoro-benzaldehyde	4-t-Butylaniine	45	95
3g	2-Fluoro-5-methoxy benzaldehde	3-Fluoroaniline	45	82
3h	Cyclopropane carboxaldehyde	4-Fluoro-3-trifluoromethylaniline	30	85
3i	Cyclohexane carboxaldehyde	4-Morpholinobenzenamine	60	78
3j	Cyclohexane carboxaldehyde	2,5-Dimemethylaniline	30	88
3k	2-Allyloxy benzaldehyde	4-(4-Chlorophenoxy)benzenamine	45	75
3l	5-(2-Chlorophenyl)furan-2-carbaldehyde	4-(4-Chlorophenoxy)benzenamine	60	73
3m	1-Acetyl-1H-indole-3-carbaldehyde	Benzo[d]thiazol-7-amine	60	68
3n	2,4-Difluorobenzaldehyde	2,4-Difluoroaniline	45	75
3o	2,6-Difluorobezaldehyde	4-Chloro-3-fluoroaniline	30	78
3p	Thiophen-3-carboxaldehyde	4-Cyanoaniline	45	68
3q	3-Ethoxybenzaldehyde	4-Fluoroaniline	30	72

and two fungi just use the different methods to investigate the antimicrobial potency by the broth dilution method.

The content of Table 4.2 is Ar, R', Mol. formula, Mol. weight and M.P. $^{\circ}C$. The letters Ar and R stand for various compounds or substances. Ar stands for aromatic compounds. Any one of a vast group of unsaturated chemical compounds known as aromatic compounds is distinguished by having one or more planar rings of atoms connected by covalent bonds of two distinct types. Molecular formula (M.F), is an equation that specifies the quantity and kind of atoms that make up a substance's molecules. Molecular weight is denoted by M.W. The total atomic weights of the atoms in a molecule are measured by its molecular weight.

Table 4.3: The table of the chemical composition against bacteria and fungi

Compound	Ar	R	Mol. formula	Mol. weight	M. p. (°C)
8a	4-Chlorophenyl	2-Chloro-6-fluoro	$C_{22}H_{13}Cl_2FN_4O$	439.27	212–214
8b	Pyridine-3-yl	2-Chloro-6-fluoro	$C_{21}H_{13}ClFN_5O$	405.81	222–224
8c	2-Chlorophenyl	2-Chloro-6-fluoro	$C_{22}H_{13}Cl_2FN_4O$	439.27	202–204
8d	3-Chlorobenzyl	2-Chloro-6-fluoro	$C_{23}H_{15}Cl_2FN_4O$	453.3	192–194
8e	4-(Methylsulfonyl)phenyl	2-Chloro-6-fluoro	$C_{23}H_{16}ClFN_4O_3S$	482.91	>225
8f	4-Chlorophenyl	2,4-Dichloro	$C_{23}H_{15}Cl_2FN_4O$	455.72	200–202
8g	Pyridine-3-yl	2,4-Dichloro	$C_{21}H_{13}Cl_2N_5O$	422.27	188–190
8h	2-Chlorophenyl	2,4-Dichloro	$C_{22}H_{13}Cl_3N_4O$	455.72	201–203
8i	3-Chlorobenzyl	2,4-Dichloro	$C_{23}H_{15}Cl_3N_4O$	469.75	198–200
8j	4-(Methylsulfonyl)phenyl	2,4-Dichloro	$C_{23}H_{16}Cl_2N_4O_3S$	499.37	184–186
8k	4-Chlorophenyl	3-Chloro-2-fluoro	$C_{22}H_{13}Cl_2FN_4O$	439.27	>225
8l	Pyridine-3-yl	3-Chloro-2-fluoro	$C_{21}H_{13}ClFN_5O$	405.81	186–189
8m	2-Chlorophenyl	3-Chloro-2-fluoro	$C_{22}H_{13}Cl_2FN_4O$	439.27	190–192
8n	3-Chlorobenzyl	3-Chloro-2-fluoro	$C_{23}H_{15}Cl_2FN_4O$	453.3	>225
8o	4-(Methylsulfonyl)phenyl	3-Chloro-2-fluoro	$C_{23}H_{16}ClFN_4O_3S$	482.91	>225
8p	4-Chlorophenyl	2,3-Dichloro	$C_{23}H_{15}Cl_2FN_4O$	455.72	194–196
8q	Pyridine-3-yl	2,3-Dichloro	$C_{21}H_{13}Cl_2N_5O$	422.27	>225
8r	2-Chlorophenyl	2,3-Dichloro	$C_{22}H_{13}Cl_3N_4O$	455.72	180–182
8s	3-Chlorobenzyl	2,3-Dichloro	$C_{23}H_{15}Cl_3N_4O$	469.75	172–174
8t	4-(Methylsulfonyl)phenyl	2,3-Dichloro	$C_{23}H_{16}Cl_2N_4O_3S$	499.37	>225

4.2 Computation

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

Chapter 5

Results and Discussion

Results

Eighteen chemical compounds sample is examined for the antimicrobial evaluation, six microbes (4 bacteria and 2 Fungi) E. coli, E. Aerogenes, K. Pneumonia, P. Vulgaris, P. Aeruginosa, and S. Pyogenes are used to analyze the antibacterial activity against the sample. The avg, max, min, and S.D of antimicrobial evaluation against considered microbes have been observed.

Table 5.1 shows the minimum value of E. Coli and k. Pneumoniae is 62.50, the maximum value of S. Pyogenes as well as P. Aeruginosa is 1200 and most average of antibacterial activity against S. pyogenes is 858.33, and the least average of antibacterial activity against P. aeruginosa is 200. The most variation shown in anti-microbial evaluation against P. aeruginosa is 402.40, data values far away from the mean, and the least variation in E. aerogenes is 53.55, and data values are close to the mean.

Seventeen chemical compounds sample is examined for the antimicrobial evaluation, six microbes (4 bacteria) E. Coli, B. Subtilis, P. Aeruginosa, and S. Aureus are used to analyze the antibacterial activity against the sample. The avg, max, min, and S.D of minimum inhibitory concentration (MIC) of antimicrobial evaluation against considered microbes.

The Table 5.2 illustrates that the minimum value of MIC against all the bacteria is 1.61 and the maximum value is 25. The maximum average or center point of P. Aeruginosa is 11.67 and the least average value is 6.27 for S. Aureus. The lowest variation around the mean/average MIC of antibacterial evaluation against S. aureus bacteria is 7.58 and the most variation

Table 5.1: In this table the characteristics of the antimicrobial evaluation are showcased:

Antimicrobial Evaluation	Min	Max	Mean	S.D
E. Coli	62.50	500	207.64	122.41
P. Aeruginosa	100	250	200	53.55
S. Auresus	62.50	500	204.17	102.54
S. Pyogenes	100	500	222.22	89.07
E. Elbicans	200	1200	561.11	402.40
As. Fumigatus	250	1200	858.33	337.92

of values away from the average point of P. aeruginosa bacteria is 10.45.

In the antimicrobial evaluation research sample of 20 chemical compounds is considered.

Table 5.2: The contents of minimum inhibitory concentration (MIC) of antibacterial considered microbes are presented:

Inhibitory concentration of antibacterial	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

The microbes (4 bacteria and 2 Fungi) against which the sample is observed are E. coli, S. Aureus, K. Pneumoniae, E. Aecalis, P. Aeruginosa, A. Fumigates, and C. Albicans is observed. The avg, max, min, and S.D of antimicrobial evaluation (antibacterial activity) against considered microbes.

In Table 5.3 maximum value against K. Pneumoniae, E. Coli, S. Aureus, and E. Aecalis is 50 and the least S.D for antibacterial activity against C. Albicans is 8.27 and the maximum S.D of K. Pneumomiae is 16.64. The maximum mean value of bacteria K. Pneumoniae is 20.00 and the minimum mean value of C. Albicans is 8.64.

Bertin in 1967 first propose the Distance Matrix visualization [57] as a method of methodically showing data structures and interactions using a reorderable matrix. The visualization of the raw data matrix (subjects by variables) has gained a lot of attention over the last few years, but little work has been done on the visualization of the equivalent proximity matrices (subjects by subjects, variables by variables).

clustering algorithms & visualization were performed using factoextra and cluster libraries

Table 5.3: The contents of antibacterial and antifungal activity against the microbes that are considered are presented:

Antibacterial and Antifungal 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

of R. To Visualize the dissimilarity matrix in R we used the fviz_dist function. This colour-based representation of distance matrix or ordered data matrices aims to depict tabular numbers and connections in a natural and accessible way[58].

The below Figure 5.1, the blue colour indicates high similarity while orange colour indicates low similarity. Where pure blue represents zero and pure orange represents six in the value of dissimilarity between observations, the colour level is proportional to that value. There are a lot of nearby observations that are similar. Orange colour denotes a great distance between observations, whereas blue colour represents a small distance.

The visualization matrix Figure 5.2 shown that, the blue colour indicates high similarity

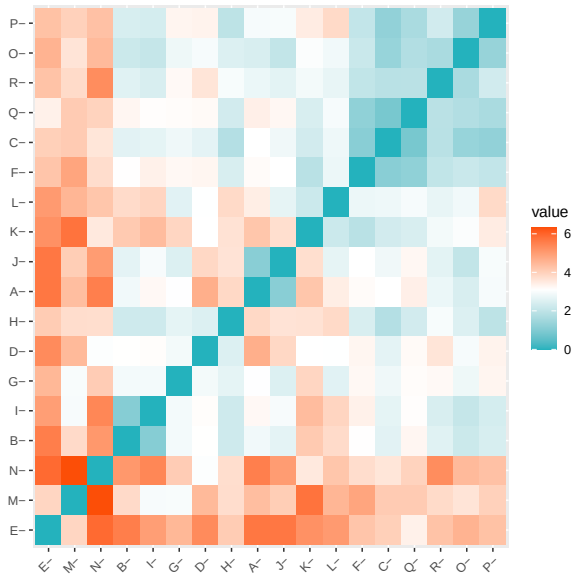


Figure 5.1: This figure represent distance matrix visualization of antimicrobial evaluation

while orange colour indicates less similarity. The colour level is related to the magnitude of dissimilarity between observations, where pure blue represents zero and pure orange represents six. Similar observations are near together. Blue denotes a small distance between observations, whereas orange suggests a large distance.

In the visualization matrix Figure 5.3 explained that, the blue colour indicates less similar-

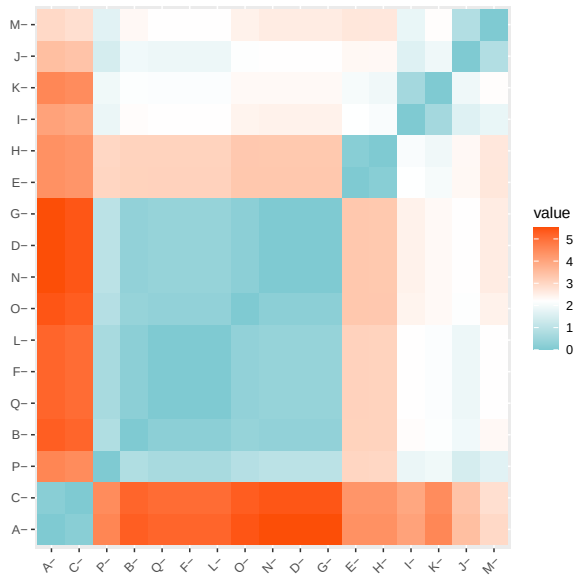


Figure 5.2: This figure represent that distance matrix visualization of minimum inhibitory concentration (MIC)

ity while orange colour indicates low similarity. The colour level is related to the value of dissimilarity between observations or data point where pure blue represents zero and pure orange represents six. Blue colour represents to small distance and orange colour indicates large distance between observation.

Before we begin the process, we must specify the number of clusters/groups. It is frequently useful to experiment with different values of k and compare the results. The same procedure may be used for 2, 3, 4, and 5 clusters, with the results displayed in the figures: The Figure 5.4 explained that K-means clustering anti-microbial evaluation, this visual analysis shows us where actual delineations between clusters exist, it does not indicate the ideal number of clusters. The Figure 5.5 explained that K-means clustering of MIC of bacteria,

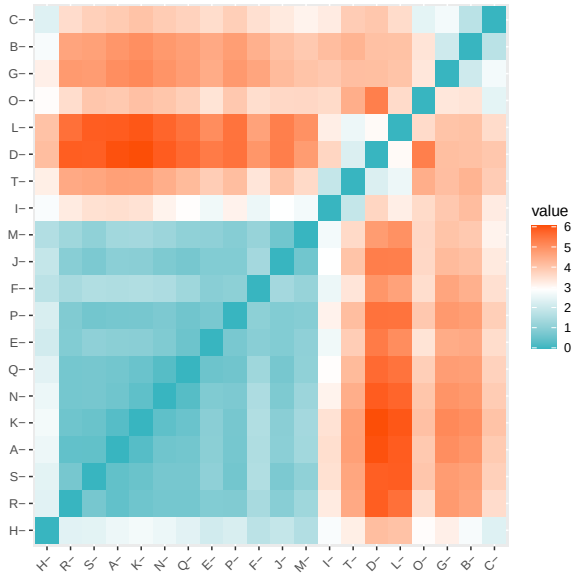


Figure 5.3: This figure represents distance matrix visualization of antibacterial and antifungal activity of chemical compounds

when we take the value of k is 2,3,4, and 5.In the cluster contain those data points whose have close to centroid point.If we explained the value of k=3, the green colour represents those data points that lie in the 2nd cluster and contain a total of five data points and the remaining seven data points lie in the blue colour of the 3rd cluster with k=4& 5 we have a number cluster are 4 and 5.

The Figure 5.6 explained that K-means clustering of anti-bacteria and anti-fungus, when we take the value of k is 3. The number of clusters of K-means the clustering algorithm will be three, in the 1st cluster containing those 3 data points that have close to a centroid point. The red colour represents those data points that lie in the 1st cluster and contain the total of four data points and the remaining data points lie in the green and blue colour of the 2nd and 3rd. The value of k=4& 5 we have a number of clusters are 4 and 5. The Table 5.4 represents that K-means clustering has a minimum value of within sum square is 30.87 with 6 maximum number of clusters. K-means clustering algorithm and mixture of multivariate t distribution with 3 optimal number of clusters gives maximum Silhouette Width is 0.22 and maximum Dunn index is 0.45 of anti-microbial evaluation of the antibacterial activity. We stop the process of clustering with a total number of clusters is 6 because the table at the

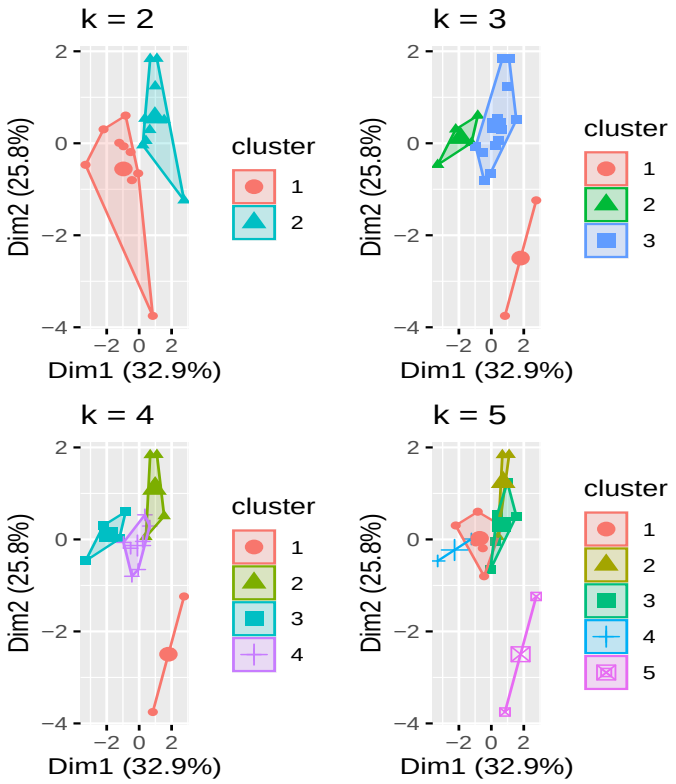


Figure 5.4: This figure describe a different number of the cluster for antibacterial activity

k=6 mixtures of multivariate t distribution clustering algorithm does not work and gives NA column.

The Table 5.5 shows that for the minimum inhibitory concentration of antibacterial activity

Table 5.4: The clustering validation of antimicrobial evaluation of antibacterial activity for different values of k

Anti-microbial evaluation	k=2			k=3			k=4			k=5			k=6		
	K-means	*Mclust*	*teigen*	*K-means*	*Mclust*	*teigen*	*K-means*	*Mclust*	*teigen*	K-means	Mclust	teigen	*K-means*	*Mclust*	*teigen*
Within SS	79.11	95.11	82.58	60.06	79.54	60.06	48.27	57.58	51.04	39.09	39.96	39.98	30.87	32.78	NA
Silhoutte_width	0.15	0.007	0.13	0.22	0.10	0.22	0.16	0.17	0.13	0.17	0.21	0.17	0.20	0.20	NA
Connectivity	16.01	19.84	18.07	15.31	22.42	15.31	24.56	21.99	25.28	28.27	25.96	25.02	29.90	27.53	NA
Dunn_index	0.21	0.26	0.22	0.45	0.28	0.45	0.45	0.28	0.25	0.32	0.45	0.28	0.41	0.40	NA

K-means clustering algorithm and multivariate t distribution clustering algorithm gives almost same results, the minimum within sum square of K-means is 0.78 and Silhouette Width of K-means and mixture of multivariate t distribution is 0.79 at the five maximum number of clusters, the minimum connectivity of K-means and a mixture of the multivariate t distri-

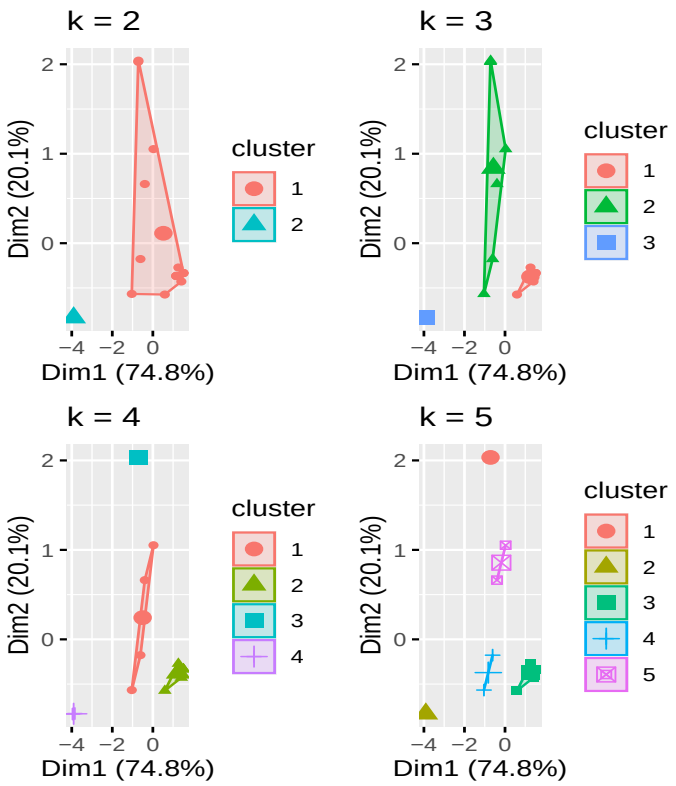


Figure 5.5: This figure describe a different number of cluster for minimum inhibitory concentration (MIC)

bution is 3.85 at 2 number of clusters. The Dunn Index (DI) is 1.44 maximum at optimal 5 number of clusters of K-means and multivariate t distribution and A mixture of multivariate t distribution does not give the answer at the k=6.

The Table 5.6 antibacterial activity of chemical compounds distributed the K-means gives

Table 5.5: The clustering validation of minimum inhibitory concentration (MIC) of antibacterials activity for different values of k

MIC	k =2			k=3			k=4			k=5			k=6		
of anti bacteria	K-means	Mclust	teigen	K-means	Mclust	teigen	K-means	Mclust	teigen	K-means	Mclust	teigen	K-means	Mclust	teigen
Within SS	27.82	35.65	27.82	10.35	12.64	12.66	4.69	5.68	12.66	1.73	2.25	1.37	0.78	1.79	NA
Silhoutte_width	0.65	0.47	0.65	0.65	0.58	0.60	0.72	0.67	0.65	0.79	0.72	0.79	0.68	0.64	NA
Connectivity	3.85	9.24	3.85	11.47	13.15	10.75	14.94	16.95	13.32	17.33	22.44	17.33	21.04	3.85	NA
Dunn_index	0.85	0.35	0.85	0.53	0.21	0.62	0.64	0.29	0.62	1.44	0.39	1.44	0.79	0.39	NA

minimum within sum square is 15.19 at 6 optimal cluster and maximum silhouette width of K-means and Gaussian mixture model is 0.52 at 3 optimal clusters. The connectivity 7.14 is a minimum of K-means cluster at 2 optimal clusters. The Dunn index is 0.57 maximum at 3

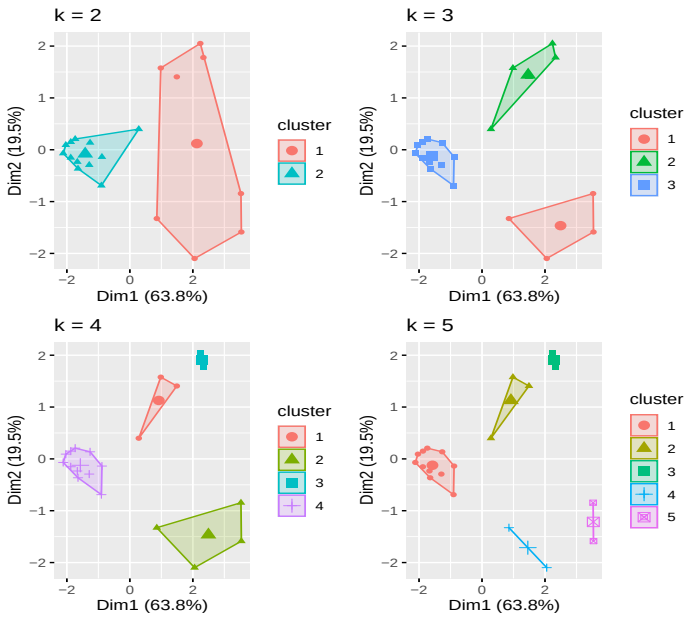


Figure 5.6: This figure describe a different number of the cluster for antibacterial and anti-fungal activity

optimal clusters.

On the bases of the "majority score" term and from the results of tables 5.4, 5.5 and 5.6 the

Table 5.6: The clustering validation of antibacterials and antifungal activity for different values of k

anti bacteria and anti fungal	k =2			k=3			k=4			k=5			k=6		
	K-means	Mclust	teigen	K-means	Mclust	teigen	K-means	Mclust	teigen	K-means	Mclust	teigen	K-means	Mclust	teigen
Within SS	52.99	53.01	53.01	31.70	31.70	36.63	25.48	29.67	26.10	19.68	23.69	19.68	15.19	20.69	NA
Silhoutte_width	0.50	0.49	0.49	0.52	0.52	0.50	0.46	0.18	0.48	0.44	0.16	0.44	0.44	0.14	NA
Connectivity	7.14	7.33	7.33	11.26	11.26	11.92	15.76	26.48	14.16	19.36	28.90	19.36	20.6	31.15	NA
Dunn_index	0.44	0.30	0.30	0.43	0.43	0.57	0.43	0.08	0.47	0.54	0.08	0.53	0.54	0.08	NA

K-means algorithm and mixture of multivariate t distribution give 3 optimal number of clusters in an anti-microbial evaluation of antibacterial activity data set and 5 number of optimal clusters in minimum inhibitory concentration (MIC) of anti bacteria's data set. K-means and mixtures of multivariate t distribution and GMM gives 3 optimal number of clusters in the antibacterial and anti-fungal activity data set. The K-means clustering algorithm gives the best performance on the bases of the majority score clustering algorithm.

The Table 5.9 describes the mean value of bacteria and fungus, the mean of E. Coli in 1st cluster is 0.62, and the minimum mean of fungus S. Aureus is -0.53. In the 2nd cluster max-

imum mean of S. aureus is 0.77 and the minimum value of E. Coli is -0.77, in the 3rd cluster fungus S. Pyogenes has a most of value 1.46. The variation explained by K-means cluster with 3 optimal clusters of anti-microbial evaluation of antibacterial activity 33.2 %.

The Table 5.10 describes the mean or average value of bacteria, the mean of S. aureus in

Table 5.7: The table shows cluster means of antimicrobial evaluation of antibacterial activity

no.of Clusters	E.coli	P. Aeruginosa	S. Aureus	S. Pyogenes	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

(between_SS / total_SS of antimicrobial evaluation of antibacterial activity = 33.2 %)

the 1st cluster is 2.47 and the minimum mean of P. aeruginosa is 1.19. In the 2nd cluster maximum mean of E. coli is 1.66 and the minimum value of B. subtilis is -0.77, in the 3rd cluster P. aeruginosa has a most of value 0.67 and same as the in clusters 4 and 5 the maximum mean value of S. aureus is 2.47 & -0.45. The variation explained by K-means cluster with 5 optimal clusters of MIC of antibacterial activity 92.75%. The Table 5.8 describes the

Table 5.8: The table shows cluster means of a minimum inhibitory concentration of antibacterial activity

No. of cluster	S. Aureus	B. Subtilis	E. Coli	P. Aeruginos
1	2.47	2.29	1.64	1.19
2	-0.61	-0.65	1.61	1.22
3	0.10	0.26	0.00	0.67
4	2.47	2.42	1.66	1.27
5	-0.45	-0.49	-0.72	-0.84

(between_SS / total_SS of minimum inhibitory concentration (MIC) of antibacterial = 92.7 %)

clusters centroid of bacteria and fungus, the mean of E. Coli in 1st cluster is 0.62, and the minimum mean of k. Pneumoniae is -6.05. In the 2nd cluster maximum mean of k. Pneumoniae is 2.00 and the minimum value of E. Coli is -0.18, in the 3rd cluster, E. Aecalis has a least mean value is 0.03. The variation explained by K-means cluster with 3 optimal clusters of anti-microbial evaluation of antibacterial and anti-fungal activity 72.2 %.

The Figure 5.7 represents that, elliptical shapes of clusters contain a different chemical composition of bacteria and fungus. The 1st cluster contains (I, B, G, D, H, and M) labels of

Table 5.9: The table shows cluster means of antibacterial and antifungal activity

no.of Clusters	S. Aureus	E. Aecalis	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

(between_SS / total_SS of antibacterial and antifungal activity = 72.2%)

a chemical compound, 2nd cluster contains those compounds that have the same chemical characteristics (J, A, R, L, and O) and in the 3rd cluster (N, K, F, C, Q, O, and E) have same chemical compounds characteristics.

The Figure 5.8 represents that K-means clustering of MIC antibacterial activity, spherical

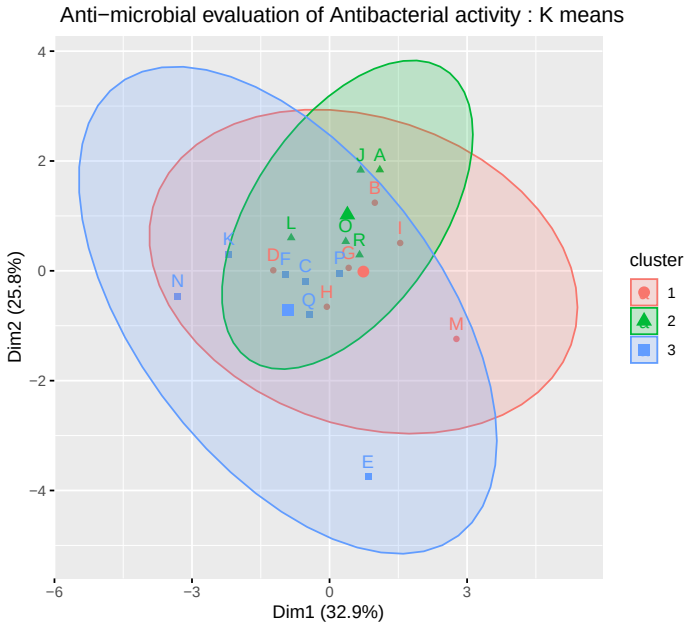


Figure 5.7: This figure illustrates antimicrobial evaluation of antibacterial activity

shapes of clusters contain a different chemical composition of bacteria and fungus. The 1st cluster contains (A and C) labels of a chemical compound, 2nd cluster contains those compounds that have the same chemical characteristics (P, B, O, G, N, Q, E, and D), and the 3rd cluster (L, H, and F) and the 4th cluster contain (K and I), 5th cluster contains (J and M) have same chemical characteristics.

The Figure 5.9 describes that K-means clustering of antibacterial and anti-fungal activity, el-

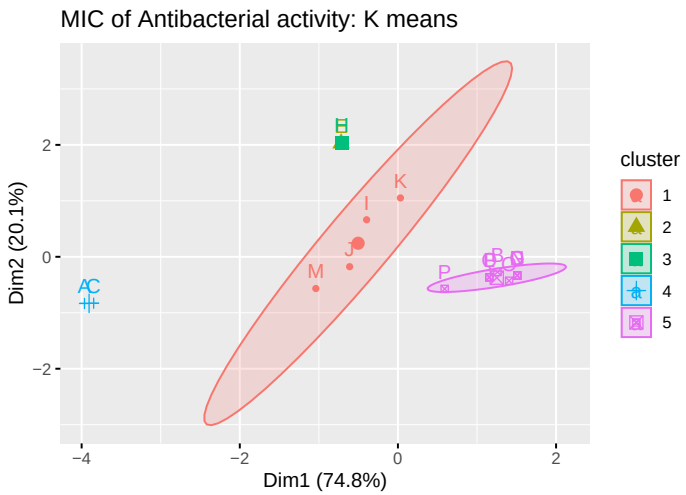


Figure 5.8: This figure represents that minimum inhibitory concentration antibacterial activity

liptical shapes of clusters contain the different chemical compositions of bacteria and fungus. The 1st cluster contains (J, M, E, F, P, O, N, K, A, S, and R) labels of a chemical compound, 2nd cluster contains those compounds that have the same chemical characteristics (I, T, D, and, L) and in the 3rd cluster (C, O, B, G, and H) have same chemical characteristics.

The content of Table 5.10 describes the log-likelihood, n, df, BIC, and ICL. For the optimal model log-likelihood procedure used in GMM; however, BIC works well in distribution-based clustering and the value of BIC and ICL is -235.65. The minimum value of BIC means that the distribution of chemical compounds are correctly classify into clusters on the bases of characteristics.

The Figure 5.10 describes that GMM clustering of antibacterial and anti-fungal activity,

Table 5.10: The table shows that GMM clusters of antibacterial and antifungal activity

log-likelihood	n	df	BIC	ICL
-83.3754	20	23	-235.65	-235.66

In this case, the best model according to BIC is an unequal-Covariance model with 3 optimal numbers of components or clusters.

elliptical shapes of clusters contain different chemical compositions of bacteria and fungus. The 1st cluster contains (J, M, E, F, P, O, N, K, A, S, and R) labels of a chemical compound,

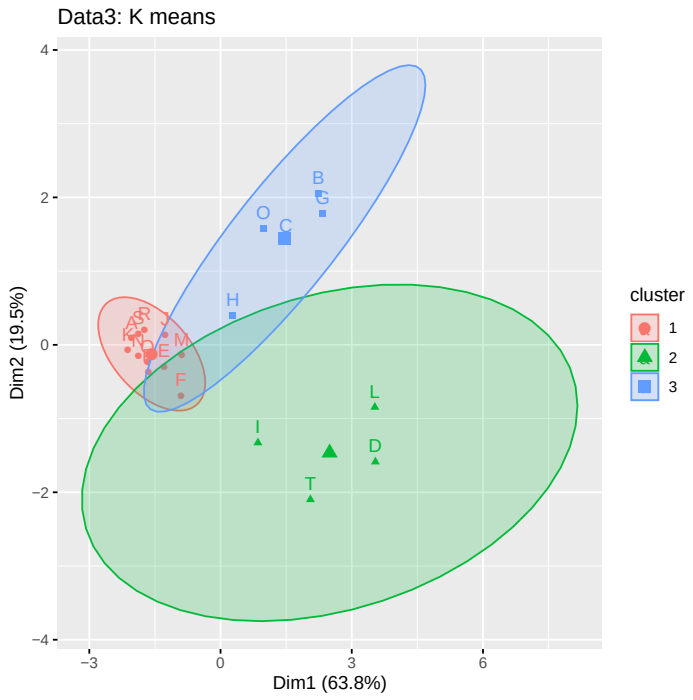


Figure 5.9: This figure represents that antibacterial and antifungal activity

2nd cluster contains those compounds that have the same chemical characteristics (I, T, D, and L) and in the 3rd cluster (C, O, B, G, and H) have same chemical characteristics.

The best model (BIC of -328.82, ICL of -328.8218) is CICC with G=3 of anti-microbial evaluation of the antibacterial activity. The best model (BIC of 11.6815, ICL is 11.6815) is CCCU with G=5 of, minimum inhibitory concentration(MIC) of antibacterial and the best model (BIC of -248.12, ICL of -250.0513) is CICC with G=3.

The Figure 5.11 describes that mixture of multivariate t distribution model clustering of antibacterial activity, elliptical shapes of clusters contain a different chemical composition of bacteria and fungus. The 1st cluster contains (E and M) labels of a chemical compound, 2nd cluster contains those compounds that have the same chemical characteristics (J, A, B, I, O, R, G, P, C, F, H, and Q) and in the 3rd cluster (L, D, K, and N) have same chemical characteristics.

The best model (BIC of 11.6815, ICL is 11.6815) is CCCU with G=5 of Minimum inhibitory concentration(MIC) of antibacterial.

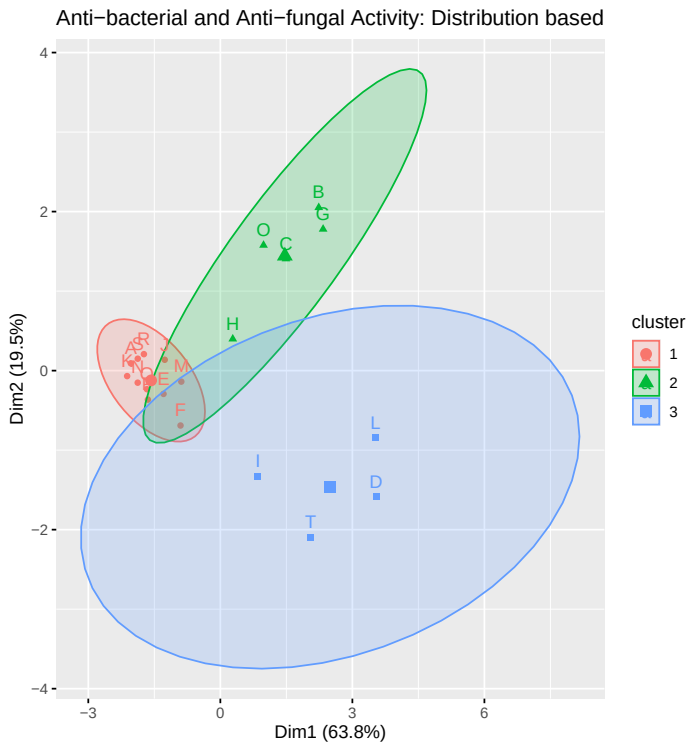


Figure 5.10: This figure represents that GMM clustering antibacterial and antifungal activity

The Figure 5.12 illustrates that a mixture of multivariate t distribution model clustering of antibacterial activity, elliptical shapes of clusters contain different chemical compositions of bacteria and fungus. The 1st cluster contains (B, I, and K) labels of a chemical compound, 2nd cluster contains those compounds that have the same chemical characteristics (A and C) and the 3rd cluster (J, M, P, Q, E, B, O, D, N, F, G, H, and L) have same chemical characteristics.

The best model (BIC of -248.12, ICL of -250.0513) is CICC with G=3 of antibacterial and antifungal activity.

The Figure 5.13 illustrates that a mixture of multivariate t distribution model clustering of antibacterial and antifungal activity, elliptical shapes of clusters contain a different chemical composition of bacteria and fungus. The 1st cluster contains (H, M, F, J, E, O, P, N, K, A, S, and R) labels of a chemical compound, 2nd cluster contains those compounds that have the same chemical characteristics (O, C, G, B, and L) and in the 3rd cluster (D, T and I) have

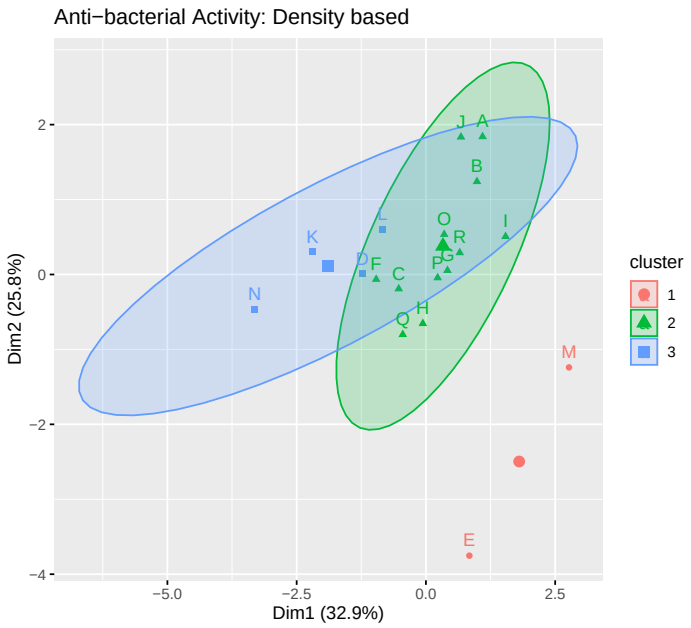


Figure 5.11: This figure represents that multivariate t distribution mixture model clusters antimicrobial evaluation of antibacterial

same chemical characteristics.

Chemical composition against the microbiological evaluation of bacteria's the figure 5.14 shows 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 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 Figure 5.15 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.

Chemical composition against the microbiological evaluation of bacteria's the 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

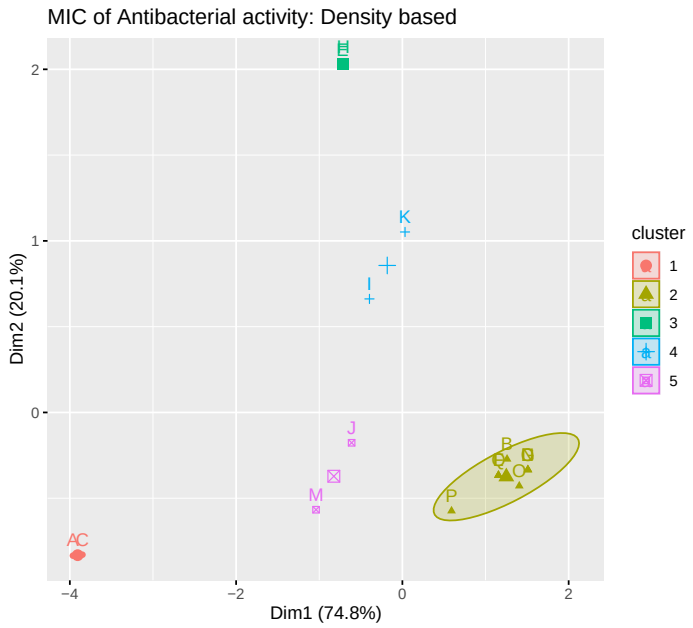


Figure 5.12: This figure represents that multivariate t distribution mixture model clusters MIC of antibacterial

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).

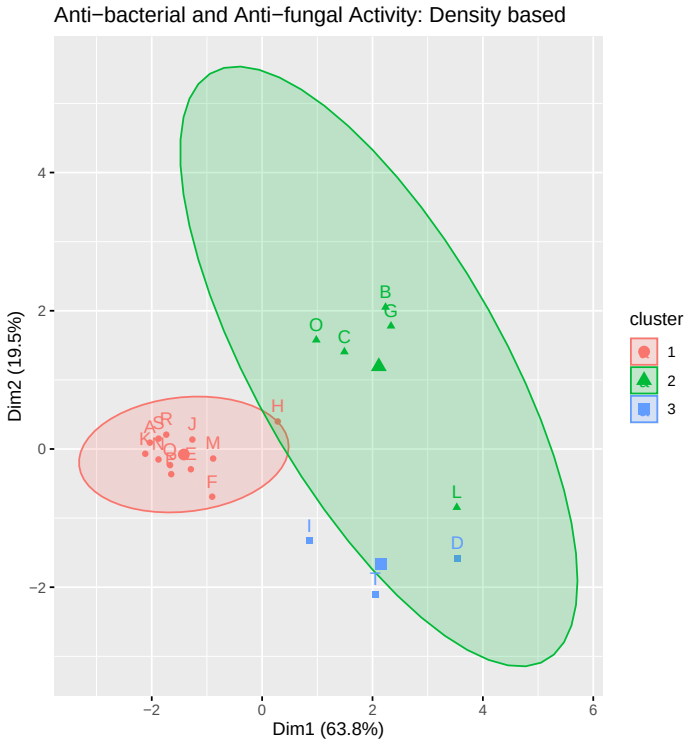


Figure 5.13: This figure represent that multivariate t distribution mixture model clusters antibacterial and antifungal activity

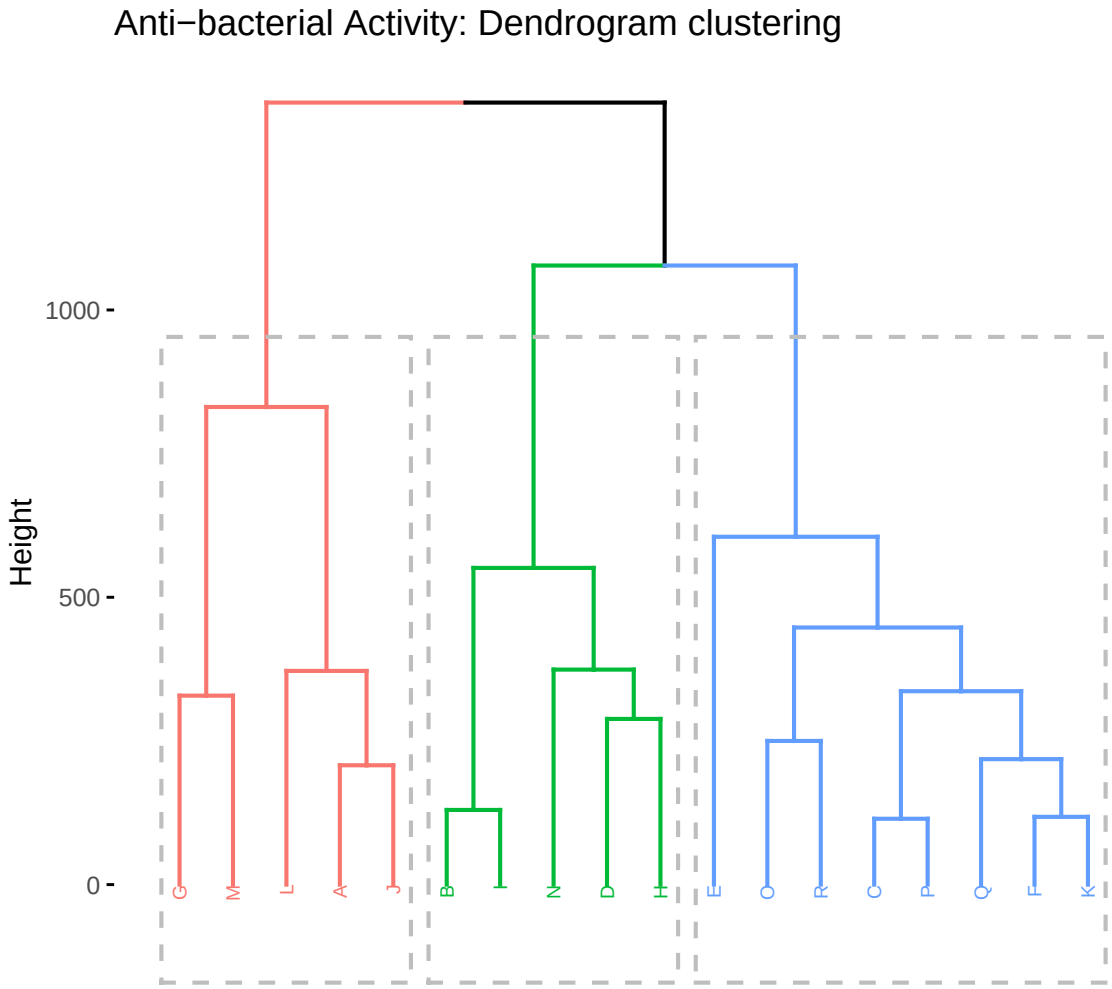


Figure 5.14: This diagram describe that dendrogram of antimicrobial evaluation

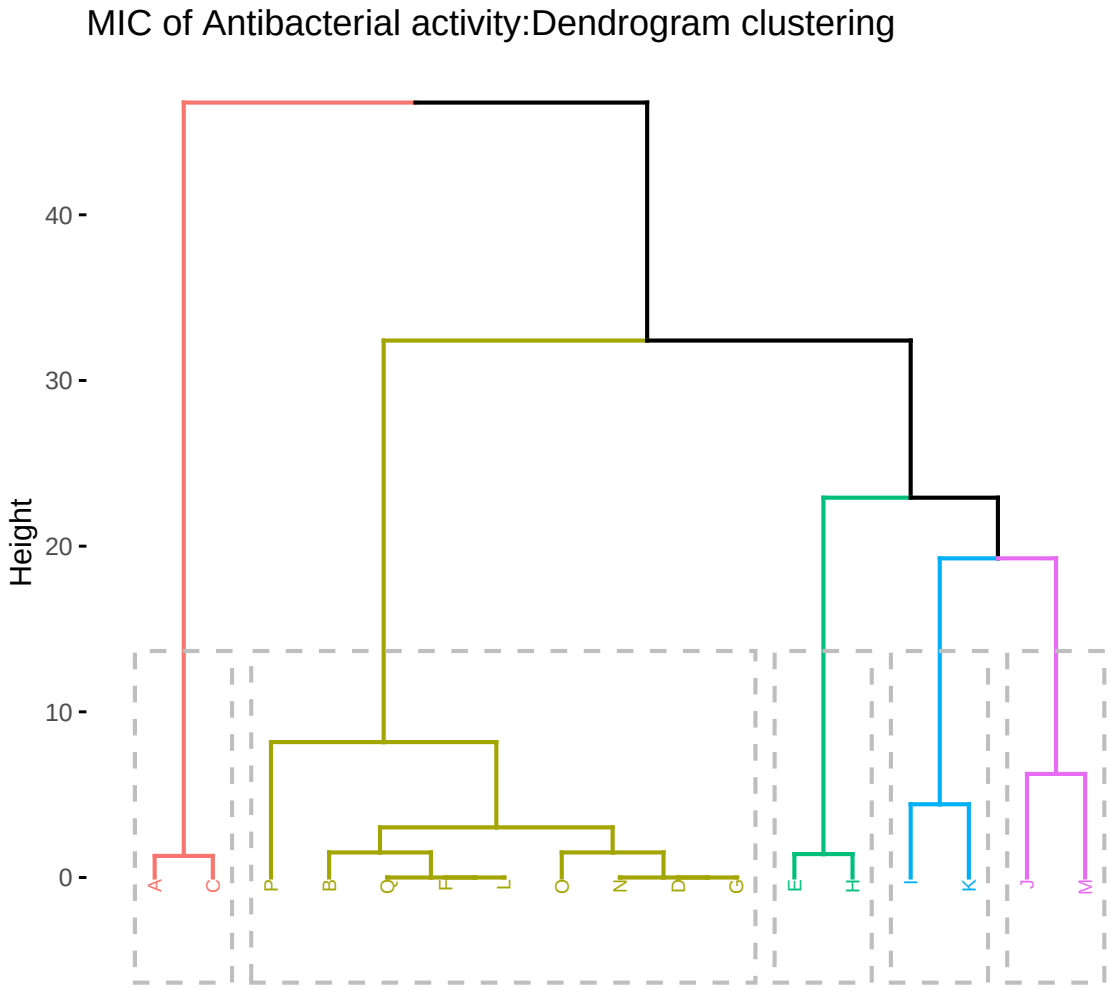


Figure 5.15: This diagram describes that dendrogram of MIC

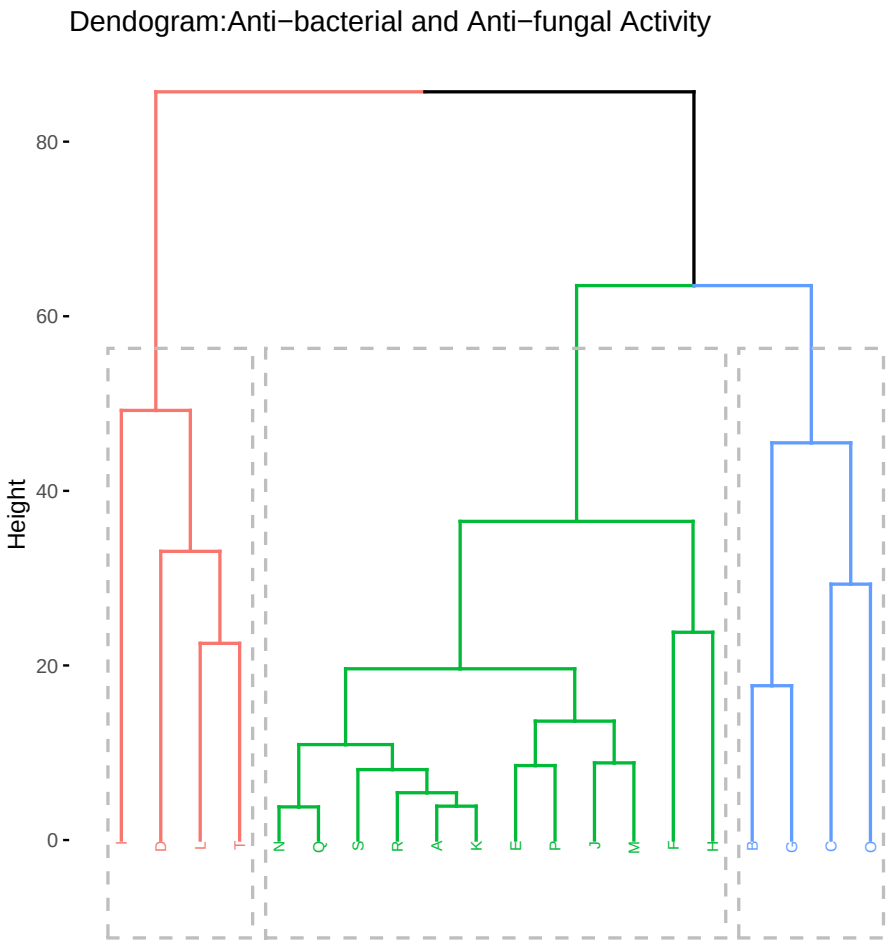


Figure 5.16: This diagram describes the dendrogram of antibacterial and antifungal activity

Discussion

In this section, we will compare our thesis results with already existing literature of comparison of clustering algorithms, which clustering algorithms perform better.

This paper examines the cluster representativeness of K-means and GMM in the Cloud workloads data set. K-means provide significantly abstracted information and take less time for computation as compared to GMM. The performance of K-means is better than GMM [59].

In this paper, the authors compare K-means and GMM methods on the forest fire data set and evaluate that K-means gives better results as compared to GMM [60].

This paper investigates the comparison of two unsupervised clustering algorithms of K-means and GMM. It was discovered that the GMM technique can appropriately categorize the machining context, however, K-means is ineffective[61]. The clustering algorithm comparison of high dimensional data set on a sales of video games in the world.The authors conclude that K-means performance is faster than GMM [62].

In this thesis, we conclude that K-means and mixtures of the multivariate t distribution perform better as compared to GMM in chemical compounds of antibacterial data sets. K-means and mixtures of multivariate t distribution equally satisfy the CVI and give the same results.

Chapter 6

Conclusion

The Tables 5.4, 5.5, and 5.6 represent that K-means clustering algorithm and a mixture of the multivariate t distribution with 3 optimal number of clusters gives maximum Silhouette Width is 0.22, minimum connectivity is 15.31 and maximum Dunn index is 0.45 of Anti-microbial evaluation of the antibacterial activity. For minimum inhibitory concentration of antibacterial activity, K-means clustering algorithm and multivariate t distribution clustering algorithm give almost the same results minimum sum square of K-means Silhouette Width is 0.79 at 5 optimal number of clusters, The minimum connectivity of K-means and a mixture of the multivariate t distribution is 3.85 at 2 number of clusters. The Dunn Index (DI) is 1.446 maximum at optimal 5 number of clusters of K-means and multivariate t distribution. Antibacterial activity of chemical compounds distributed the K-means gives minimum within sum square is 15.19 at 6 optimal clusters and maximum silhouette width of K-means and Gaussian mixture model is 0.52 at 3 optimal clusters. The connectivity 7.14 is a minimum of K-means cluster at 2 optimal clusters. The Dunn index is 0.57 maximum at 3 optimal clusters.

In this thesis, we conclude results on the bases of the majority score clustering algorithm, K-means and a mixture of multivariate t distribution satisfy the maximum and the Gaussian mixture model satisfies the minimum cluster validation techniques. The K-means clustering algorithm and a mixture of multivariate t distribution clustering algorithm gives 3 optimal number of clusters in an anti-microbial evaluation of antibacterial activity data set and 5 number of optimal clusters in minimum inhibitory concentration (MIC) of anti bacteria's

data set. K-means a mixture of multivariate t distribution and Gaussian mixture model gives 3 maximum number of clusters in antibacterial and antifungal activity data set.

Those chemical compounds having alike antibacterial activity are grouped into the same clusters or components. The maximum number of clusters for chemical compounds of antimicrobial evaluation of antibacterial activity is three, 5 optimal number of clusters for MIC of antibacterial activity, and 3 maximum number of clusters for antibacterial and antifungal activity. In the end, the K-means and multivariate t distribution clustering algorithm give the best results for clusters because they satisfy the most of cluster validation indices.

Bibliography

[1] Yalin Başţanlar and Mustafa Özuysal. Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, pages 105–128, 2014.

[2] Taiwo Oladipupo Ayodele. Machine learning overview. *New Advances in Machine Learning*, 2:9–18, 2010.

[3] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. An overview of machine learning. *Machine learning*, pages 3–23, 1983.

[4] Amparo Albalate and Wolfgang Minker. *Semi-supervised and unsupervised machine learning: novel strategies*. John Wiley & Sons, 2013.

[5] Robert Gentleman and Vincent J Carey. Unsupervised machine learning. In *Bioconductor case studies*, pages 137–157. Springer, 2008.

[6] Charu C Aggarwal. An introduction to cluster analysis. In *Data clustering*, pages 1–28. Chapman and Hall/CRC, 2018.

[7] V Moertini. Introduction to five dataclustering algorithms clustering algorithm. *Integral*, 7(2), 2002.

[8] Iawei Han, Micheline Kamber, and Jian Pei. Data mining: Concepts and techniques third edition. elsevier. 2012.

[9] Nidhi Grover. A study of various fuzzy clustering algorithms. *International Journal of Engineering Research*, 3(3):177–181, 2014.

[10] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[11] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203, 1984.

[12] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012.

[13] Frank Nielsen. Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*, pages 195–211. Springer, 2016.

[14] Jeffrey L Andrews, Jaymeson R Wickins, Nicholas M Boers, and Paul D McNicholas. teigen: An r package for model-based clustering and classification via the multivariate t distribution. *Journal of Statistical Software*, 83:1–32, 2018.

[15] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(3):231–240, 2011.

[16] Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi, and Elvia M Quiroz. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34, 2011.

[17] Eréndira Rendón, Itzel M Abundez, Citlalih Gutierrez, Sergio Díaz Zagal, Alejandra Arizmendi, Elvia M Quiroz, and H Elsa Arzate. A comparison of internal and external cluster validation indexes. In *Proceedings of the 2011 American Conference, San Francisco, CA, USA*, volume 29, pages 1–10, 2011.

[18] Julia Handl, Joshua Knowles, and Douglas B Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[19] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[20] Kheyreddine Djouzi and Kadda Beghdad-Bey. A review of clustering algorithms for big data. In *2019 International Conference on Networking and Advanced Systems (ICNAS)*, pages 1–6. IEEE, 2019.

[21] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[22] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

[23] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[24] Leon Bottou and Yoshua Bengio. Advances in neural information processing systems. *Neural Information Processing Systems Foundation (NIPS)*, pages 161–168, 1995.

[25] KA Abdul Nazeer and MP Sebastian. Improving the accuracy and efficiency of the k-means clustering algorithm. In *Proceedings of the world congress on engineering*, volume 1, pages 1–3. Association of Engineers London London, UK, 2009.

[26] Soumi Ghosh and Sanjay Kumar Dubey. Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications*, 4(4), 2013.

[27] Ahamed Shafeeq and KS Hareesha. Dynamic clustering of data with modified k-means algorithm. In *Proceedings of the 2012 conference on information and computer networks*, pages 221–225, 2012.

[28] Juntao Wang and Xiaolong Su. An improved k-means clustering algorithm. In *2011 IEEE 3rd international conference on communication software and networks*, pages 44–46. IEEE, 2011.

[29] Shi Na, Liu Xumin, and Guan Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*, pages 63–67. Ieee, 2010.

[30] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA, 2009.

[31] Xiao-Li Meng and David van Dyk. The em algorithm–an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):511–567, 1997.

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[33] Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, 142(5):1114–1127, 2012.

[34] David Peel and Geoffrey J McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

[35] David A Langan, James W Modestino, and Jun Zhang. Cluster validation for unsupervised stochastic model-based image segmentation. *IEEE Transactions on Image Processing*, 7(2):180–195, 1998.

[36] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2):107–145, 2001.

[37] Brian S Everitt. Cluster analysis, 3rd edn., edward arnold, 1993.

[38] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[39] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[40] Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75, 2002.

[41] B Mukunthan and N Nagaveni. Identification of unique repeated patterns, location of mutation in dna finger printing using artificial intelligence technique. *International Journal of Bioinformatics Research and Applications*, 10(2):157–176, 2014.

[42] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[43] E Vorhees. *The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval*. PhD thesis, PhD thesis, Department of Computer Science, Cornell University, UK, 1985. 97.

[44] Sanghamitra Bandyopadhyay and Ujjwal Maulik. Nonparametric genetic clustering: comparison of validity indices. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(1):120–125, 2001.

[45] IH Osman and S Salhi. Local search strategies for the vehicle fleet mix problem. vj rayward-smith, ih osman, cr reeves, gd smith, eds. modern heuristic search methods, 1996.

[46] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part i. *ACM Sigmod Record*, 31(2):40–45, 2002.

[47] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[48] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[49] Hugo Steinhaus et al. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1(804):801, 1956.

[50] Youguo Li and Haiyan Wu. A clustering method based on k-means algorithm. *Physics Procedia*, 25:1104–1109, 2012. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.

[51] Shy Shoham. Robust clustering by deterministic agglomeration em of mixtures of multivariate t-distributions. *Pattern Recognition*, 35(5):1127–1142, 2002.

[52] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE international conference on data mining*, pages 911–916. IEEE, 2010.

[53] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. Understanding and enhancement of internal clustering validation measures. *IEEE transactions on cybernetics*, 43(3):982–994, 2013.

[54] Anupriya Vysala, Dr Gomes, et al. Evaluating and validating cluster results. *arXiv preprint arXiv:2007.08034*, 2020.

[55] Tanvi Gupta and Supriya P Panda. Clustering validation of clara and k-means using silhouette & dunn measures on iris dataset. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 10–13. IEEE, 2019.

[56] Juan Carlos Rojas-Thomas and Matilde Santos. New internal clustering validation measure for contiguous arbitrary-shape clusters. *International Journal of Intelligent Systems*, 36(10):5506–5529, 2021.

[57] Jacques Bertin. *Sémiologie graphique: les diagrammes, les réseaux, les cartes*. De Gruyter Mouton, 1973.

[58] Gap: A graphical environment for matrix visualization and cluster analysis. *Computational Statistics & Data Analysis*, 54(3):767–778, 2010. Second Special Issue on Statistical Algorithms and Software.

[59] Eva Patel and Dharmender Singh Kushwaha. Clustering cloud workloads: K-means vs gaussian mixture model. *Procedia Computer Science*, 171:158–167, 2020.

[60] Sarvesh Kumar. Comparative analysis in between the k-means algorithm, k-means using with gaussian mixture model and fuzzy c means algorithm. 02 2017.

[61] Zhiqiang Wang, Catherine Da Cunha, Mathieu Ritou, and Benoît Furet. Comparison of k-means and gmm methods for contextual clustering in hsm. *Procedia Manufacturing*, 28:154–159, 2019.

[62] Saadaldeen Rashid Ahmed Ahmed, Israa Al Barazanchi, Zahraa A Jaaz, and Haider Rasheed Abdulshaheed. Clustering algorithms subjected to k-mean and gaussian mixture model on multidimensional data set. *Periodicals of Engineering and Natural Sciences (PEN)*, 7(2):448–457, 2019.