Clustering Techniques in Machine Learning

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering/Information Technology

By

Vatsal Singh (191286)

Under the supervision of

Dr. Monika Bharti

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Certificate

I hereby declare that the work presented in this report entitled **" Clustering Techniques in Machine Learning"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from July 2022 to May 2023 under the supervision of **Dr. Monika Bharti** Assistant Professor (SG).
The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Vatsal Singh, 191286

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Monika Bharti
Assistant Ptofessor (SG)
Computer Science and Engineering/Information Technology
Dated:

Acknowledgement

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success.

First of all I wish to acknowledge the benevolence of omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds with the profound sense of gratitude and heartiest regard. I express my sincere feelings of indebtedness to my guide **Dr. Monika Bharti** for their positive attitude, excellent guidance, constant encouragement, keen interest, invaluable co-operation, generous attitude and above all their blessings. She has been a source of inspiration for me.

Last but not the least I would like to express my heartfelt thanks to my parents and my friends who with their thought provoking views, veracity and whole hearted cooperation helped in doing this project.

Vatsal Singh

(191286)

Abstract

In today's era data generated by scientific applications and corporate environment has grown rapidly not only in size but also in variety. This data collected is of huge amount and there is a difficulty in collecting and analyzing such big data. Data mining is the technique in which useful information and hidden relationship among data is extracted, but the traditional data mining approaches cannot be directly used for big data due to their inherent complexity.

Data Clustering is one of the most important issues in data mining and machine learning. Clustering is a task of discovering homogenous groups of the studied objects. Recently, many researchers have a significant interest in developing clustering algorithms. The most problem in clustering is that we do not have prior information knowledge about the given dataset. Moreover, the choice of input parameters such as the number of clusters, number of nearest neighbors and other factors in these algorithms make the clustering more challengeable topic. Thus any incorrect choice of these parameters yields bad clustering results. Furthermore, these algorithms suffer from unsatisfactory accuracy when the dataset contains clusters with different complex shapes, densities, sizes, noise, and outliers. In this project, we propose a new approach for unsupervised clustering task. Our approach consists of three phases of operations. In the first phase we use the Genetic algorithm for finding first initial cluster centroid. In genetic algorithm we use a crossover and mutation of the dataset. The second phase, takes these initial cluster centroid produced by genetic algorithm for finding clusters using K-means clustering. From the second phase we obtain a set of clusters of the given dataset. Hence, the third phase considers these clusters for evaluation of cluster based on Davies Bouldin Index. This new algorithm is named as Genetic K-means Algorithm (GKA). We present experiments that provide the strength of our new proposed algorithm in discovering clusters with different non-convex shapes, sizes, densities, noise, outliers and higher accuracy. These experiments show the superiority of our proposed algorithm when comparing with K-means algorithm.

Table of Contents

Certificate ………………………………………………………………………….. i

Acknowledgement……………………………………………………………. ii

Abstract…………………………………………………….…………………. iii

Table of Contents……………………………………………….……………….. iv

List of Figures……………………………………………….……………………... vi

List of Table……………………………………….……………………………….. vii

1. Introduction…………………………………………………………................ 1

1.1 Introduction to Machine Learning…………………………………… 1

1.2 Unsupervised Learning……………………………………………… 1

1.2.1 Clustering……………………………………………….. 2

1.3 Types of Clustering ………………………………………………….. 3

1.3.1 Partitioning Methods……………………………………. 3

1.3.2 Hierarchical Clustering…………………………………. 3

1.3.3 Fuzzy Clustering………………………………………… 4

1.3.4 Model Based Clustering………………………………… 4

1.3.5 Density Based Clustering……………………………….. 4

1.4 Comparison of Clusters………………………………………………. 4

1.4.1 Euclidian Distance………………………………………. 5

1.4.2 Manhattan Distance……………………………………… 5

1.4.3 Edit Distance…………………………………………….. 5

1.4.4 Hamming Distance………………………………………. 5

1.5 Techniques to find the optimum number of Clusters…………………. 5

1.5.1 Elbow method……………………………………………. 5

1.5.2 Average silhouette method………………………………. 6

1.5.3 Gap Statistical method…………………………………… 7

1.5.4 Davies Bouldin Index……………………………………. 8

1.5.5 Dunn Index………………………………………………. 8

1.6 Standard K-Means Algorithm………………………………………… 8

1.6.1 Flowchart of standard K-means…………………………. 9

1.6.2 Drawbacks of standard K means clustering……………… 9

1.6.3 Example of K-means…………………………………….. 10

1.7 Genetic Algorithm………………………………………………….. 11

1.7.1 Flow chart of genetic algorithm………………………… 14

2. Literature Survey………………………………………………………… 15

2.1 Clustering……………………………………………………….…… 15

2.2 Partitioning Clustering……………………………………………. 15

2.3 Unsupervised learning technique……………………………… 16

2.4 Conclusion………………………………………………………… 18

3. System Design and Development………………………………………… 19

3.1 Problem Statement………………………………………………… 19

3.2 Research Gap……………………………………………………… 19

3.3 Objectives………………………………………………………… 20

3.4 Research Methodology…………………………………………… 20

3.5 Proposed Hybrid Technique……………………………………… 20

3.6 Basic Genetic Algorithm………………………………………..… 21

3.6.1 Application of genetic algorithm…………………………… 23

3.6.2 Example of MaxOne using genetic algorithm……………… 23

3.7 Proposed Algorithm……………………………………...…………… 23

3.7.1 Example of proposed Algorithm……………………………. 27

4. Experiments and Result Analysis……………………………………………. 36

4.1 Implementation of Proposed Technique………………………………. 36

4.1.1 Iris Dataset for Implementation…………………………….. 36

4.2 Experimental Results…………………………………………………. 38

4.2.1 Confusion Matrix of K-means clustering…………………… 39

4.2.2 Confusion Matrix of Genetic K-means clustering………….. 40

4.2.3 Test for performance Accuracy…………………………….. 41

4.2.4 Calculation of Intra cluster distance………………………… 42

5. Conclusion and Future Scope………………………………………………... 45

5.1 Conclusion…………………………………………………………… 45

5.2 Limitations…………………………………………………………… 45

5.3 Future Scope………………………………………………………….. 46

References……………………………………………………………………… 46

List of Figures

Figure No. Description Page No.

1.1 Inter and Intra Similarities of Cluster 2

1.2 Evaluation Graph of Elbow Method 6

1.3 Evaluation Graph of Silhouette Method 7

1.4 Evaluation Graph of Gap Statistical Method 7

1.5 Flowchart of standard K-means Clustering 9

1.6 Clustering on Iris Dataset 11

1.7 Genetic Algorithm Chromosomes and Population 12

1.8 Execution Steps of Genetic Algorithm 14

3.1 Implementation Methodology for Clustering of dataset 21

3.2 Process of Genetic Algorithm 22

3.3 Flowchart of Proposed Algorithm 27

4.1 Code of K-means Clustering on Iris Dataset 37

4.2 Code of Genetic K-means Clustering on Iris dataset 38

4.3 Confusion Matrix obtained from K-means Algorithm 39

4.4 Confusion Matrix obtained from Genetic K-means Algorithm 40

List of Tables

Table No. **Description** **Page No.**

1.1 Crossover Operation on chromosome S1 and S3 13

1.2 Result of Crossover on Chromosome S1 and S3 13

1.3 Mutation Operation on Chromosome S1 and S3 13

1.4 Result of Mutation on Chromosome S1 and S3 13

3.1 Initialization of Chromosome for MaxOne Problem 23

3.2 Arrangement of Chromosome based on Fitness value 24

3.3 Crossover of chromosome S1 and S3 25

3.4 Crossover Result of chromosome S1 and S3 25

3.5 Crossover of chromosome S2 and S4 25

3.6 Crossover Result of chromosome S2 and S4 25

3.7 Crossover of chromosome S5 and S6 26

3.8 Crossover Result of chromosome S5 and S6 26

3.9 Mutation Result of chromosomes 26

3.10 Iris dataset for Genetic K-means Clustering 31

3.11 Normalized dataset for Genetic K-means Clustering 32

3.12 Selected Row Indices and Chromosomes 33

3.13 Calculated Distance and Assignment of cluster 34

3.14 Clusters obtained for Fifteen Records 34

4.1 Accuracy obtained from K-means and Genetic Algorithm 41

4.2 Intra Cluster distance using K-means algorithm 42

4.3 Intra Cluster distance using Proposed algorithm 42

4.4 Inter Cluster distance using K-means algorithm 43

4.5 Inter Cluster distance using Proposed algorithm 44

Chapter 1 Introduction

Raw data may be gathered in large quantities from many different disciplines, but this data is meaningless unless it is properly reasoned through to provide knowledge that is helpful. In this project, we concentrate on clustering, one of the key data mining techniques.

1.1 Introduction to Machine learning

Big data is defined as large-scale data, which can range from digital to health-related information. Big data consists of extensive and complicated data sets, making it difficult for typical data processing programmes to handle them. Every day, the information sector produces vast amounts of data [1]. Information explosion, the term used to describe the increasing rate in data volume, was first used to describe how data grew enormous seventy years ago. Unsupervised learning is a type of machine learning where the model learns patterns, structures, and relationships from unlabelled data without any specific guidance or predefined target values. Unlike supervised learning, which requires labelled data with known outcomes, unsupervised learning algorithms explore the data on their own to discover inherent patterns and make sense of the information.

In unsupervised learning, the goal is often to find hidden structures or clusters in the data or to uncover relationships between variables. By extracting meaningful patterns, unsupervised learning algorithms can help with tasks such as data exploration, data pre-processing, anomaly detection, and feature engineering. The most successful unsupervised learning approach is clustering, which is one of several strategies employed. The most used clustering method is K-means. K-means is sensitive to out-of-the-ordinary data points, and it occasionally creates empty clusters.We recommended a novel method to address this issue. K-means genetic algorithm. In this study, the Genetic K-means clustering approach is the main focus.

1.2 Unsupervised Learning

There are two main types of unsupervised learning algorithms:

1. Clustering: Clustering algorithms group similar data points together based on their

characteristics or proximity in the feature space. The objective is to identify natural clusters within the data without any prior knowledge of the class labels. Common clustering algorithms include k-means, hierarchical clustering, and DBSCAN.

2. Dimensionality reduction: Dimensionality reduction techniques aim to reduce the number of variables or features in the data while preserving important information. These methods are useful when dealing with high-dimensional data by simplifying it and allowing for easier visualization and analysis. Principal Component Analysis (PCA) and t-SNE (t-Distributed Stochastic Neighbor Embedding) are popular dimensionality reduction algorithms.

Unsupervised learning plays a crucial role in various fields, including data mining, pattern recognition, recommender systems, and exploratory data analysis, as it helps uncover underlying structures and insights in unlabelled datasets.

1.2.1 Clustering:

It is an unsupervised method that uses an automated method to group items with similar properties together [5]. Statistics professionals also refer to it as categorization. Clustering is grouping things based on similarities. The items are divided into two clusters: one for those with similar qualities and another for those with different properties. The Manhattan distance or the Euclidian distance are used to measure similarity. Getting low intra-cluster and the converse for inter-cluster similarity is the main goal of clustering.

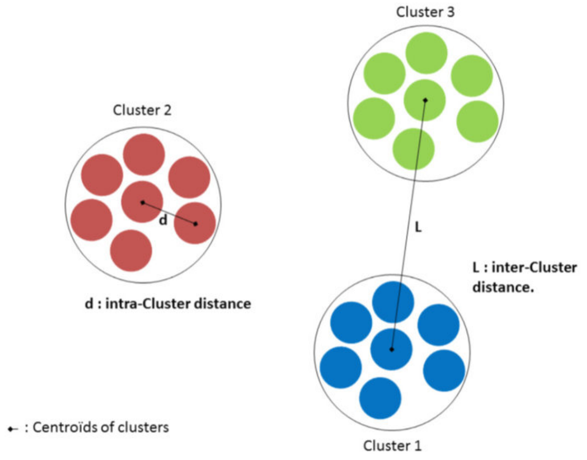

Cluster 3
Cluster 2
d
L
L : inter-Cluster distance.
d : intra-Cluster distance
• : Centroïds of clusters
Cluster 1

Figure 1.1 Inter and Intra similarities of cluster [5]

1.3 Types of Clustering

- Partitioning clustering
- Hierarchical clustering
- Fuzzy clustering
- Density Based clustering
- Model based clustering

1.3.1 Partitioning methods

The collection of data items is separated into non-overlapping groups in partitioning clustering. Greedy techniques are used to find groups of data items [6]. Data partitioning techniques divide data points into several subgroups. The use of an iterative method is used to identify the best suitable clusters. These algorithms offer high-quality data point grouping. These methods require that the algorithm's input, the number of partitions or clusters (k), be known in advance. The algorithm for partitioning uses the following heuristics:

K-means clustering: In 1967, Macqueen developed K-means clustering. Depending on the value of k, it determines the clusters. The analyst has previously determined or established what the value of k should be. It is a method of unsupervised machine learning [6]. Each data point is grouped according to the similarity metric in K-means clustering. Despite being one of the most effective technique, it can occasionally produce empty clusters for big datasets and is sensitive to anomalous data points.

K-medoids clustering: Kaufman and Rousseau proposed the K-medoids method, sometimes referred to by Partitioning around Medoids **(PAM),** in 1990.It is a method of clustering for dividing a dataset into k parts [6] that is comparable to K-means clustering. Since K-medoid is a reliable alternative to K-means, it is less sensitive to fluctuations and out of the ordinary data than K-means.

Clustering Large Applications: An extension of K-medoid is the clustering large application, or CLARA [7]. It was coined by Kaufman and Rousseau in 1990.It is utilized for data including several items.

1.3.2 Hierarchical Clustering:

Another name for hierarchical clustering is HCA. It creates a cluster by collecting related things. It is not necessary to pre-specify the value of k or the number of clusters [7]. The outcome of hierarchical clustering yields a tree-based structure. Its primarily of two types:

- Agglomerative Clustering
- Divisive Clustering

1.3.3 Fuzzy Clustering:

Soft clustering, commonly referred to as fuzzy clustering, allows data points to belong to many clusters [6]. Data points in hard clustering might belong to many clusters. In a fuzzy clustering, every item has a membership coefficient corresponding to that value, indicating which cluster it belongs to.

1.3.4 Model Based Clustering:

A technique to cluster analysis is model-based clustering. Data are considered in model-based clustering as originating from several probability distributions [7]. The amount of clusters is estimated using this clustering method, which also identifies outstanding model–data fit. The following methods are used for model-based clustering: Expectation reduction, conceptual grouping, and the use of neural networks

1.3.5 Density Based Clustering:

Ester first used the term "density-based spatial clustering" in 1996.DBSCAN is another name for it. Using this method, clusters of various sizes and forms may be produced.

1.4 Comparison of Clusters

Clusters are formed from data points based on how closely they are spaced. Items with comparable traits are gathered together into one cluster, and those with differing traits are clustered together into another [8]. Depending on the type of dataset, there are several approaches to determine how similar two clusters are. Creating a distance metric across data points is the most typical method of determining how similar two data points are. The following distances are used to determine how similar two clusters are:

1. Euclidian Distance
2. Manhattan Distance
3. Edit Distance

4. Hamming Distance

1.4.1 Euclidian Distance:

A straight line distance called a euclidian distance exists between two places. L2 standard is another name for the Euclidian separation [8]. The distance between two points on a plane with the coordinates (x, y) is known as the Euclidian separation. The formula to calculate Euclidian distance is:

$$d[|y_1, y_2 \dots\dots\dots.. y_n|], [|z_1, z_2 \dots\dots\dots. z_n|] = \sqrt{\sum (y_i - z_i)^2}$$

d- distance between the coordinates (x,y)

1.4.2 Manhattan Distance:

City block distance, or L1 norm are other names for Manhattan Distance. It is measured in right-angle axes [8]. Manhattan distance is the product of diagonal and horizontal distance, which is calculated using Pythagoras' theorem. The Manhattan Distance formula is as follows:

$$d[|x_1, x_2 \dots\dots\dots.. x_n|], [|y_1, y_2 \dots\dots\dots. y_n|] = \sqrt{\sum |x_i - y_i|}$$

d- distance between the coordinates (x,y)

1.4.3 Edit Distance:

When the dataset is available through form strings, Edit Distance is utilized for the purpose of Finding similarities between strings is done using the Edit separate approach. The amount of steps needed to switch from one string to another separates two strings [8] .It is used in bioinformatics to determine the similarities between DNA groups.

The operations used in Edit include Insert and Delete.

1.4.4 Hamming Distance:

Boolean numbers are separated by Hamming [9]. When the two strings have the same length, the hamming separation is calculated. Hamming distance's value is always positive. It might be used to the coding theory.

1.5 Techniques to find the optimum number of Clusters

1.5.1 Elbow method:

Finding the right number of clusters and interpreting and validating consistency inside the cluster are both aided by the Elbow approach. This approach focuses on variance %, which depends on the value of k [8]. The value of k should be set so that addition of more clusters does not significantly improve the modelling of the data.

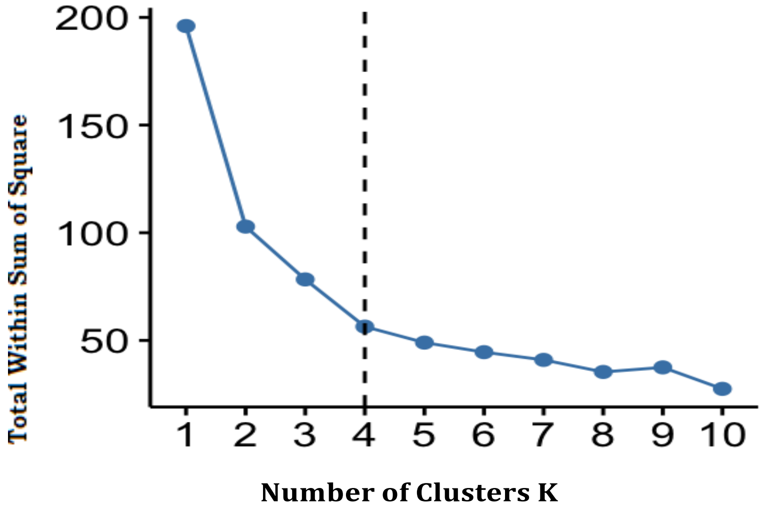


Figure 1.2 Evaluation graph of Elbow method [8]

1.5.2 Average silhouette method:

It is employed to determine the quality of clusters. The value range for the silhouette approach is [-1, 1]. The cluster is ideal [12] if the silhouette coefficient value is close to +1 A data point that has its Silhouette value set to 0 indicates that it belongs to a different cluster. Cluster is misclassified, as shown by the Silhouette value close to -1.

$$T(k) = \frac{n(k) - x(k)}{\max\{n(k), x(k)\}}$$

T(k) – silhouette index value

n(k) – distance between data points in same cluster

x(k) – distance between data points in different clusters

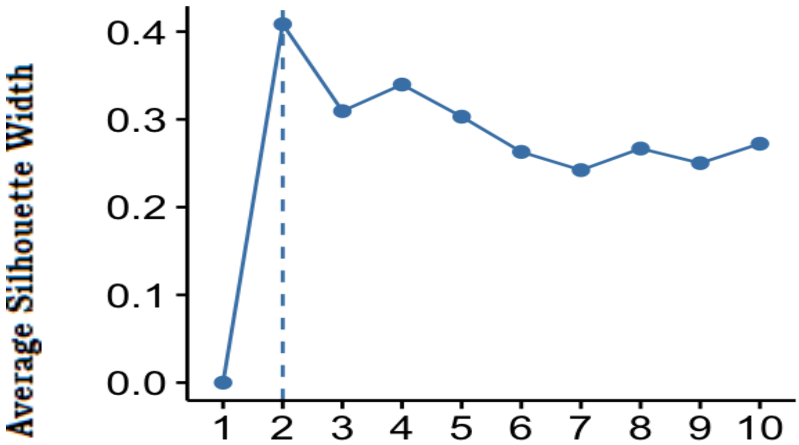


Figure 1.3 Evaluation Graph of Silhouette Method [8]

1.5.3 Gap Statistical method:

It contrasts the total inside intra-cluster variance for diff

erent values of k with the values that would be predicted under the data's null reference distribution. The ideal clusters will be formed by the values that maximise the gap statistic.

$$\text{Gap}_n\ (\text{k}) = \frac{1}{B} \sum_{b=1}^{\text{B}} \log(W_{kb}) - \log(W_k)$$

W_{kb} – Expected value

W_k – Observed value

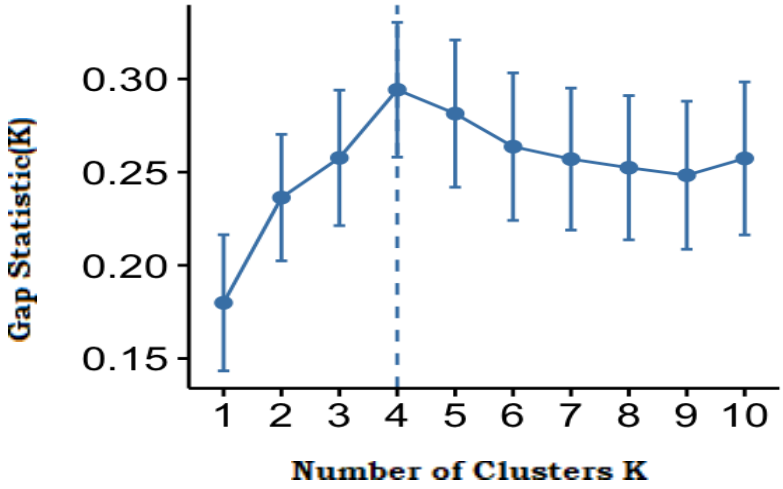


Figure 1.4 Evaluation Graph of Gap Statistic method [8]

1.5.4 Davies Bouldin Index:

It contrasts the total inside intra-cluster variance for different values of k with the values that would be predicted under the data's null reference distribution. The ideal clusters will be formed by the values that maximise the gap statistic.

$$D_j = \left(\frac{1}{T_j} \sum_{K=1}^{T_j} |S_j - M_i|^p\right)^{1/p}$$

M_i - centroid of cluster

$C_i, T_{j.}$ - the size of cluster

D_j -is the measure of validity of the cluster.

1.5.5 Dunn Index:

J.C Dunn introduced the Dunn Index in 1979. This is a metric for evaluating clusters. Dunn's index differentiates sets of clusters with similar characteristics that are grouped together [9]. A higher Dunn index value indicates better grouping. A primary disadvantage of the Dunn index is the computational cost, because the number of clusters and the dimensionality of the information increase the computation.

1.6 Standard K-Means Algorithm

Input: K: clusters to be formed.

D_n: dataset having n data points. $D_{n=}$ {d_1, d_2……d_n}

Output: Set of k clusters

Step 1: Select data points corresponds to the value of k (means).

Step 2: Calculate Centroid

- Calculate the centroid of data points using distance measures like Euclidean distance
- Data points are attached to nearest centroid based on the Euclidian distance.

Step 3: Recalculate the means.

- When all data points are assigned to the cluster than recalculate the mean.
- Repeat the step of calculating the distance and assigning the clusters.

Step 4: Repeat until the centers do not change.

- Repeat the steps 2 and 3 until the centers do not change.

1.6.1 Flowchart of standard K-means:

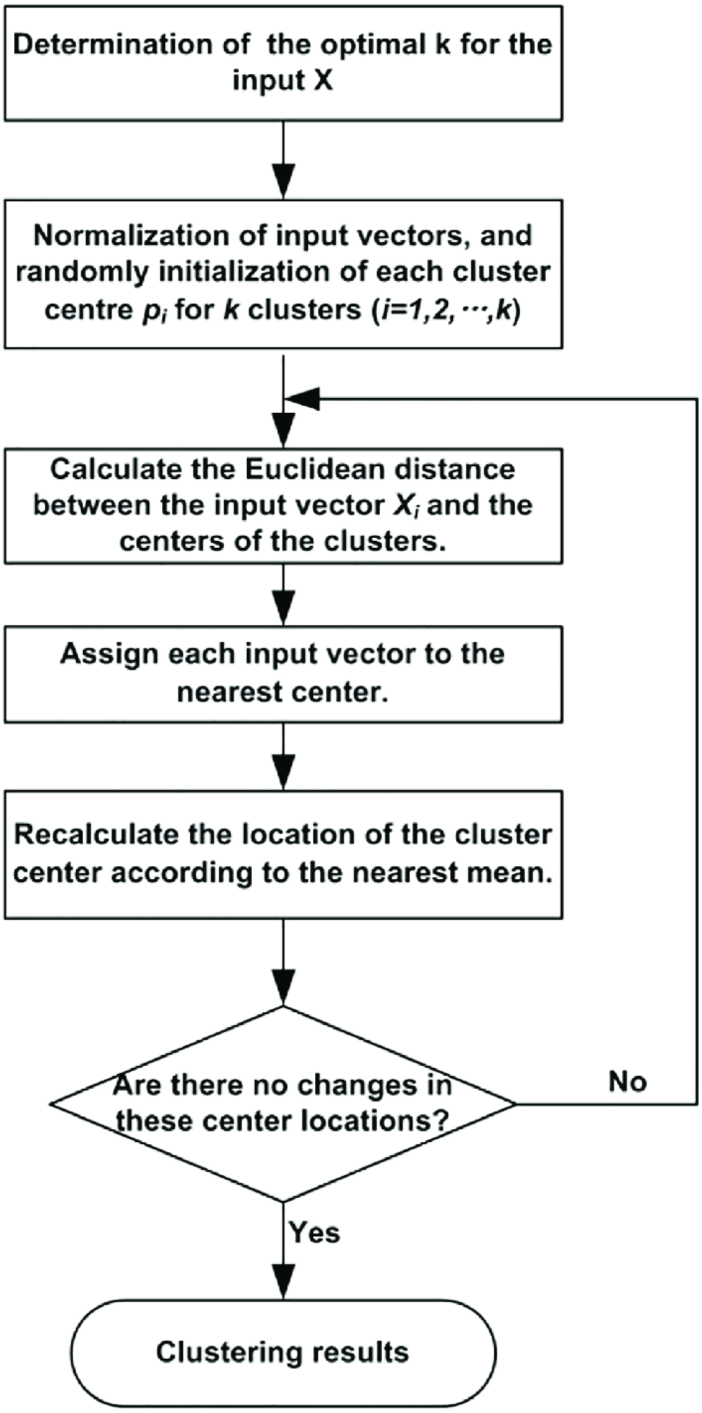

Determination of the optimal k for the input X
Normalization of input vectors, and randomly initialization of each cluster centre p_i for k clusters ($i=1,2,\cdots,k$)
Calculate the Euclidean distance between the input vector X_i and the centers of the clusters.
Assign each input vector to the nearest center.
Recalculate the location of the cluster center according to the nearest mean.
Are there no changes in these center locations?
No
Yes
Clustering results

Fig 1.5 Flowchart of standard K-means

1.6.2 Drawbacks of standard K-means clustering:

The limitations of K-Means algorithm have been listed as follows:

- It requires advance knowledge of the data and a domain expert is needed to choose the requisite value for k.
- Final result is affected more by initial selection and less by outliers.
- Reconfiguring the data yields a different result for the same data
- Only convex shaped clusters are formed

1.6.3 Example of K-means :

Dataset {2, 3, 4, 10, 11, 12, 20, 25, 30} and K=2

Step 1: Randomly select two K values or means :{ 4, 12}

Step 2: Calculate the distance between chosen values and whole dataset.

Step 3: After finding the distance of all data points the assignment is done based on minimum distance.

Step 4: After finding Euclidean distance cluster obtained are given below:

C1**:**{2,3,4}   C2:{10,11,12,20,25,30}

Step 5: Take a mean of the cluster and repeat the steps 1 to 4 until we get the similar cluster

Iteration 1: Mean of clusters

- M1=(2+3+4)/3=3   M2=(10+11+12+20+25+30/)6=18
- Again find the cluster centroid using Euclidean distance:
- Cluster obtained are:C1{2,3,4,10} C2{11,12,20,25,30}

Iteration 2: Take a mean of the cluster and again find the Euclidian distance

- M1=(2+3+4+10)/4=4.75.   M2=11+12+20+25+30/5=19.6 = 20
- Find the cluster centroid and again assign it to cluster
- Cluster obtained are C1= {2, 3, 4, 10, 11, 12} C2= {20, 25, 30}

Iteration 3:

- Find the mean: M1:2+3+4+10+11+12/6 = 7 M2= 20+25+30/3=25
- Find the cluster centroid C1 ={ 2,3,4,10,11,12} C2={20,25,30}
- As we get the same cluster with the previous cluster so we have to stop the clustering
- Final Cluster Obtained are:C1 :{2 ,3,4,10,11,12} C2{ 20,25,30}

1.6.4 K-means Clusters on the Iris dataset:

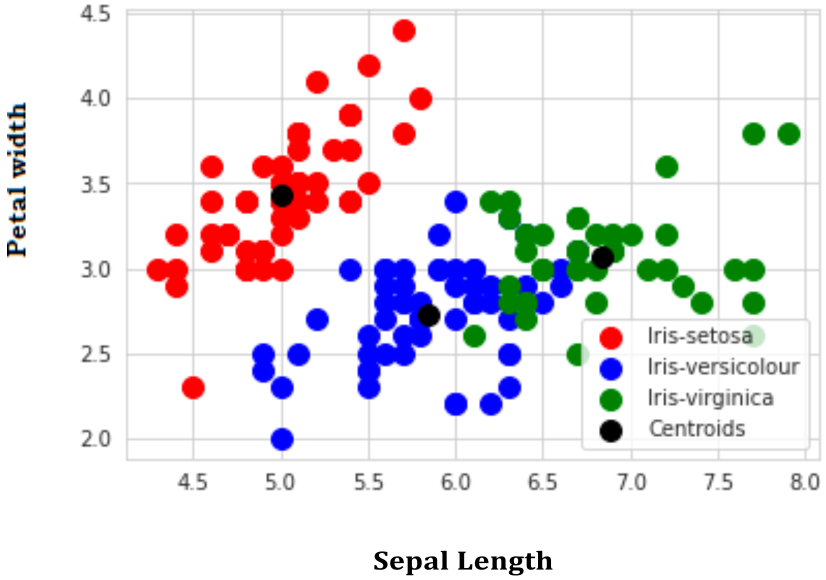


Fig 1.6 Clustering on Iris Dataset

1.7 Genetic Algorithm

It is a technique that comes from biological evolution and natural selection to tackle limited and unconstrained optimization issue [10]. The genetic algorithm repeatedly adjusts a population of individual solutions. At every step, the genetic algorithm randomly chooses individuals from the current population to be parents which are used to create the offspring for the next generation. Genetic algorithm may be used to tackle different optimization challenges. Steps are as follows:

1. Initialization of Population
2. Finding Fitness value
3. Selection of chromosomes
4. Crossover
5. Mutation

1. Initialization:

In initialization step, population is defined as set of individuals. An individual is defining by a set of variables known as **Genes**. In a genetic algorithm, strings are used to describe the set of genes of an individual. To encode the genes in a chromosome binary values are used.

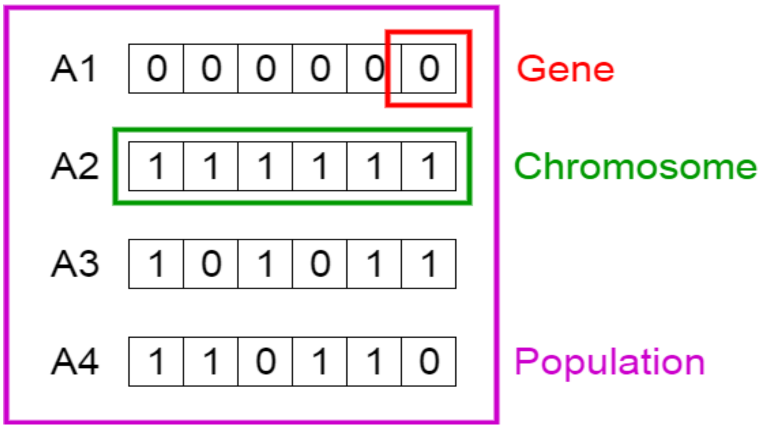


Figure 1.7 Genetic Algorithm Chromosome and Population [10]

2. Fitness Function: The fitness function calculate the ability of individual to compete with other individuals. It present a fitness score to each individual. The probability that an individualwill be selected for reproduction is depend on its fitness value. The formula for fitness value depend on the type of problem .The most commonly used formula for finding fitness value is

$$S_{max} = \max(F(S_{INTER})/F(S_{INTRA}))$$

Where S_{max} is the fitness function obtained by dividing the total inter cluster distance ($F(S_{INTER})/$) and total intra cluster distance ($F(S_{INTRA})$).

3. Selection: The concept of selection phase is to prefer the fittest individuals and their genes move onward to the next generation .Pairs of an individuals are selected according to their fitnessscores. Individuals with large fitness value have more chance to be selected for reproduction.

4. Crossover: Crossover is the most powerful phase in a genetic algorithm. Crossover is used toproduce new offspring's. Crossover point is chosen randomly from genes. The Crossover of chromosomes S1 and S3 is shown in table 1.1 and result of crossover is shown in table 1.2.

Table 1.1 Crossover operation on S1 and S3 chromosome.

S1	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**
S3	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**

Table 1.2 Result of Crossover on S1 and S3 chromosome.

S1	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**
S3	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**

Table 1.2 and 1.3 shows the crossover operation and result of chromosome S1 and S3. Thecrossover operator is applied to produce new offspring's with higher probability.

5. Mutation: The mutation is a random tweak in a chromosomes to form new results. Themutation of S1 and S3 is shown in table 1.3 and 1.4

Table 1.3 Mutation operation on Chromosome S1 and S3

S1	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**
S3	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**

Table 1.4 Result of Mutation on Chromosome S1 and S3

S1	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**
S3	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**

Table 1.3 and 1.4 shows the mutation of chromosome S1 and S3. In mutation the bits with lowerprobability is flipped and increase the fitness of particular chromosome.

1.7.1 Flow chart of genetic algorithm:

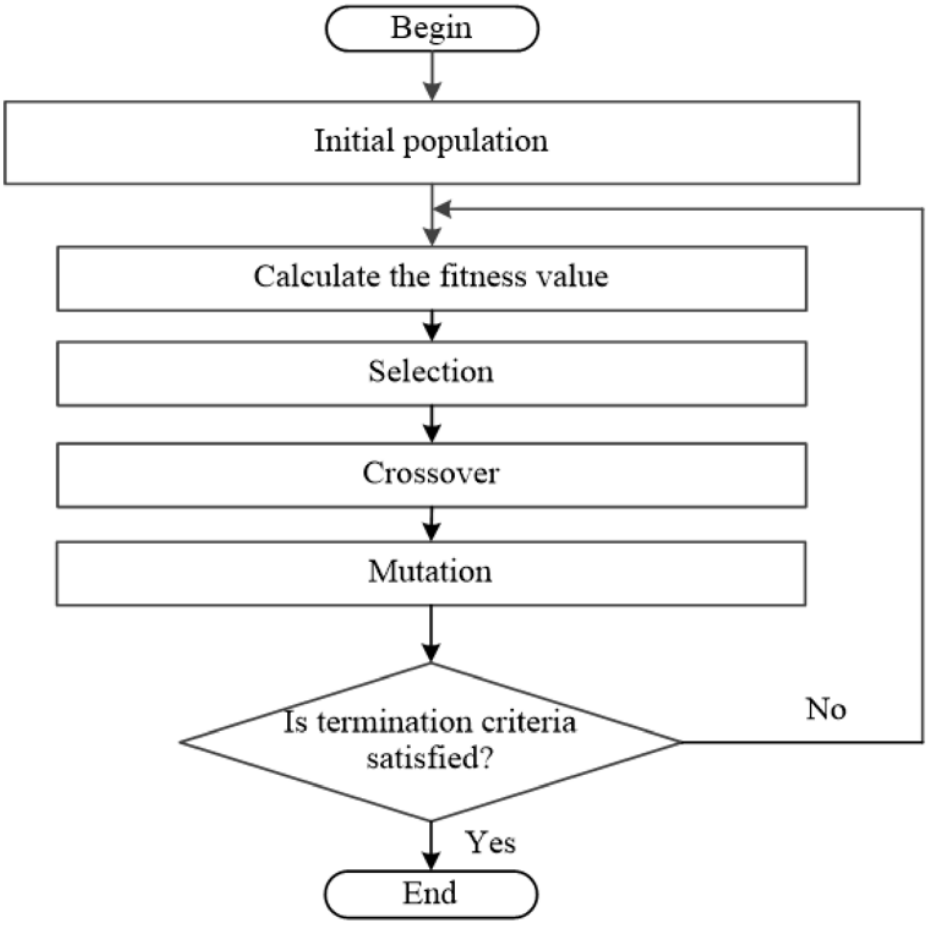

Begin
Initial population
Calculate the fitness value
Selection
Crossover
Mutation
Is termination criteria satisfied?
No
Yes
End

Figure 1.8 Execution Steps of genetic algorithm [10]

Chapter 2 Literature Survey

2.1 Clustering

Clustering is an unsupervised technique that creates clusters of objects having similar featuresusing automatic technique [1]. It is also termed as classification by statisticians. In clustering, grouping is done based on similarity of objects .The objects with similar properties are groupedin one cluster and with dissimilar properties into another. Similarity is evaluated by using Euclidian distance or Manhattan distance. The main aim of clustering is to obtain low intra- cluster similarity and high inter-cluster similarity. There are different types of clustering techniques includes partitioning clustering, hierarchical clustering, fuzzy clustering and densitybased clustering .This chapter presents literature survey of clustering techniques for analyzingbig data and tabular comparison of these techniques is presented.

2.2 Partitioning Clustering

Partioning clustering means division of the datasets into non-overlapping clusters based on similarity measure by using Euclidian distance or any distance [3]. In partitioning methods numbers of clusters is randomly selected or predefined by the analyst. Partioning means division of datasets into k groups where k is randomly selected or predefined by the analyst.

Each cluster satisfy the following conditions.

- Each group of cluster at least contain one object.
- Each object belong to one group.

There are distinctive kinds of partitioning clustering techniques. The most mainstream Partioning clustering strategy is the K-means clustering presented by Macqueen in1967. In K-means clustering every data point is grouped based on similarity. The K-means strategy is sensitive to outliers and in some cases it frames empty clusters with large datasets. In this we proposed another method (Hybrid K-means with Genetic algorithm) to evacuate the downsideof k means clustering [11].

2.3 Unsupervised learning technique: Clustering

There is huge amount of data generated from information industry and other repositories. This data is useless until it's processed and extract useful knowledge from it. The researchers analyze two elementary objectives of data mining: description and prediction. There are different techniques used for extracting useful information from large amount of data. Clustering is the most commonly used partitioning method for mining large datasets. In partitioning methods K-means clustering is most efficient method butit is sensitive to outliers and sometimes produce empty clusters with large dataset. Clustering using optimization problem techniques: genetic algorithm, decision trees, neural network remove the problem of partitioning methods [1].

Researchers identify two elementary objectives of data mining: description and prediction. Prediction utilizes various existing variables in the database in order to predict the future valuesof interest and description mainly focuses on finding various patterns that describes the data and the subsequent presentation for individual interpretation. The relative prominence of both description and prediction differ with respect to fundamental technique and the application. There are several data mining techniques fulfilling these objectives: classification mining, association rule mining and clustering using the techniques such as genetic algorithms, decisiontree, neural networks and machine learning [2].

S.Bandyopadhyay et al. [4] designed a genetic clustering algorithm called as GKA that classifythe pixels of satellite image. This KGA clustering algorithm is applied when number of clustersis known a priori and crisp in nature. In this paper Genetic algorithm is used to search clusterscenters. Floating point representation of chromosomes is used because it is more natural and appropriate form for coding the cluster centers. The major drawback of this paper is that the algorithm is applied only to those dataset whose k is already known.

M.Jain. et al. [7] proposed a K-means with genetic algorithm for enhancing stock prediction. Inthis paper they enhance a stock prediction or market analysis using k means with genetic algorithm for finding the cluster centroids. Chi square similarity is used for determining the accuracy and the result obtained has highest accuracy then k means .The drawback of proposedalgorithm is that they are only applied to matrix represented dataset.

C. Ordonez. [9] Presented two different approaches of the K-means clustering algorithm to cluster the binary data streams. The variants used by them are incremental K-means and scalable K-means. A proposed variant gives high quality clusters and less sensitive as compared to k-means. The proposed incremental k-means and scalable K-means is compared with existing k-means in terms of accuracy, confusion matrix and error rate. The incremental and scalable k-means gives higher accuracy than existing k-means.

Mor et al. [11] proposed a genetic algorithm approach for clustering and compared the results with k means algorithm. In this proposed approach the fitness is calculated on the basis of intracluster and inter cluster similarity measures. The proposed algorithm has low intra cluster distance and high inter cluster distance and also remove the drawback of local optima. The drawback of this algorithm is that the GA algorithm not find the value of K (number of clusters)as K is randomly chosen in this algorithm.

K.Dharmendra et al. [12] presented the efficient K-means clustering algorithm. The main objective of K-means clustering algorithm is to divide the dataset into K number of clusters where k is predefined or randomly selected by the analyst. The main aim of K-means clusteringused in this paper is to minimize the within sum of square.

K.Shahroudi et al. [13] proposed an algorithm for variable selection in clustering of market segmentation using genetic algorithm. The objective of this proposed algorithm to identify thevariables that are optimal and remove the irrelevant variables using genetic algorithm. Finally the result obtained have efficiently improved the outcomes based on the most relevant techniqueof segmentation.

E.O.Hartono et al. [14] proposed an algorithm for determining a cluster centroid using genetic algorithm. The determining an initial value of cluster centroid using genetic algorithm providebetter results than random numbers. Fitness value is calculated using MSE (Mean Square Error).Fitness value is 1 divided by the MSE. Lower the MSE higher the performance. The drawback of proposed algorithm is that the evaluation of clusters is not done and also the valueof K is not known Apriori.

P.Vats et al. [17] had discussed a comparative analysis of various clustering technique using genetic and K-means algorithm. It uses the sample Iris dataset to perform the differentclustering techniques i.e. K-means algorithm, Incremental K-means and Fuzzy C-means. The code is implemented using matlab and Weka.In this paper they have discussed Fuzzy C-meansgives better results as compared to K-means algorithm.

K.Kim et al. [22] proposed a recommender system using Genetic algorithm and K-means for online shopping market. In this proposed technique the initial seed is optimized by Genetic algorithm called GA K-means for online shopping recommender system. The proposed algorithm results is compared with existing algorithms and it shows that proposed algorithm improves the segmentation performance compared to existing algorithms.

2.4 Conclusion

In this chapter literature review and comparative analysis of the recent techniques for clusteringbig data is done. The various clustering techniques for analyzing big data are compared and theirmerits and demerits are presented.

Chapter 3 System Design and Development

3.1 Problem Statement

In today's world big data has become a buzz in the market. Among various challenges in analyzing big data the major issue is to design and develop the new techniques for clustering. Clustering techniques are used for analyzing big data in which cluster of similar objects are formed that is helpful for business world, weather forecasting etc. The problem in clustering is non-availability of prior information knowledge about the given dataset. Moreover, the choice of input parameters such as the number of clusters, number of nearest neighbors and other factorsin these algorithms make the clustering more challengeable topic. Thus any incorrect choice of these parameters results into bad clustering results.

Moreover, these algorithms suffer from unsatisfactory accuracy when the dataset contains clusters with different complex shapes, densities, sizes, noise and outliers. The Partitioning algorithm: K-means is the most powerful partitioning algorithm. K-means is sensitive to outliersand works well when number of clusters is known in advance. However, it gives empty clusters for large dataset. These issues can be addressed by using K-means in combination with Geneticalgorithm. In this research work a hybrid of Genetic and K means is applied in order to mitigatethe drawbacks of k-means clustering.

3.2 Research Gap

In the present period corporate and scientific environment produce large amount of data. To accumulate and examine enormous data is a difficult task as data is increasing not in amount only but in complexity. Based on literature survey, there are various techniques which are usedto analyze large datasets but these techniques are not effective as they don't give global solutions.Some of them are good but they had to compromise with the quality of clusters and vice versa. There has been lot of work done to improve efficiency of K-Means and Genetic algorithms to determine good quality clusters in less computation time but there are some shortcomings in both these techniques.

The k-means algorithm executes fast but it cannot handle non arbitrary shape, sensitive to outliers and form empty clusters for large datasets. The hybrid Genetic K-means algorithm can handle noise as well as non-arbitrary shape but it takes more computation time and is more complex than K-Means.

- The K-means algorithm is applied with cuckoo search algorithm to remove the shortcomings of existing K-means algorithm, but it doesn't handle large dataset.
- In another technique k-means is modified to incremental K-means which generates better result than k-means for numeric datasets.
- Another approach of optimized k-means clustering based on genetic algorithm. Genetic algorithm is used for finding the optimal value of k but the computation time is more as compares to another algorithms.

3.3 Objectives

The objectives of the project are as follows:

- To study the existing K-means clustering technique in combination with Genetic algorithm.
- To introduce an efficient clustering technique to remove the drawback of existing clusteringalgorithm.
- To implement and validate the proposed technique on Iris dataset.

3.4 Research Methodology

In our research work we will use python language, which is an excellent scripting language formanipulating text. The dataset required for analysis will be extracted from UCI machine learning repository. Iris dataset, Wine quality dataset is used for analyzing proposed algorithm.In First phase normalization of data is done using python and then normalized data is used for clustering. In second phase simple K-means algorithm is tested on sample dataset and predict the accuracy using confusion matrix. In the third phase genetic K-means algorithm or proposedalgorithm is tested on sample dataset and result of genetic is passed to K- means to predict accuracy. In the last phase Davies Bouldin index is used for evaluation of clusters.

3.5 Proposed Hybrid Technique

The proposed method is a hybrid technique based on K-Means and Genetic Algorithm that combines the benefits of both K-Means and Genetic algorithms [20]. The benefit of genetic algorithm is to determine the first initial cluster centroid using genetic algorithm and after applyingcrossover and mutation the result is passed to K-Means for clustering .In last after clustering DaviesBouldin index is used for Evaluation of clusters.

Fig 4.1 shows the procedure of proposed hybrid technique

Step 1: In the first step collection of data from UCI machine learning repository [42] for formationof clusters.

Step 2: In Second step normalization of data is done using python IDLE as normalized data is easyto handle.

Step 3: In the Third step Genetic algorithm for finding initial cluster centroid and after crossoverand mutation, the result is passed to K-means for formation of clusters.

Step 4: In the last steps after formation of clusters Davies Bouldin Index is used Evaluation ofcluster.

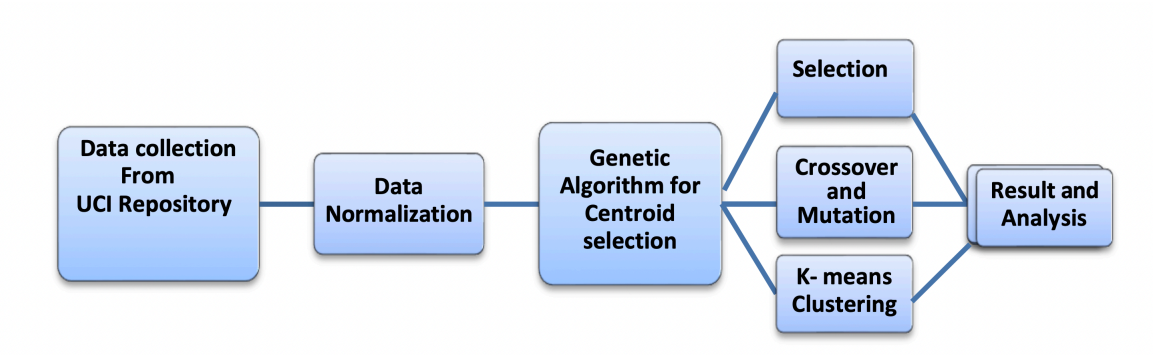


Fig 3.1 Implementation methodology for Clustering of dataset

3.6 Basic Genetic Algorithm

The Genetic Algorithm executes in three stages.

Input: Initialization of chromosomes in the form of binary strings (0's and 1's).Output: Optimal number of clusters that are less sensitive to the outliers.

Step 1: Initialization

The first step is Initialization of population or chromosomes based on our dataset.

Step 2: Fitness function

Calculate the fitness of each chromosome based on fitness function. The fitness function is mainlydepend on our problem. The main aim of clustering is low intra cluster distance and high inter cluster distance, so the most commonly used fitness function is:

$$F_{max} = \max(S(D_{INTER})/S(D_{INTRA}))$$

Where F_{max} is the fitness function obtained by dividing the total inter cluster distance and total intra cluster distance.

Step 3: Selection Phase

The selection of chromosomes is done based on the fitness value .The chromosomes are selected using roulette wheel selection method or rank based selection .The chromosomes with high fitnessvalue have high probability to be selected first.

Step 4: Crossover

Crossover is applied to chromosomes to produce new off springs. The selected chromosomes are randomly selected to produce new off springs. Crossover is simply reproduction of new chromosomes.

Step 5: Mutation

It is used to maintain a genetic diversity of population After Crossover Mutation is applied which is random tweak to particular chromosome. There are various methods for mutation includes:

- Bit Manipulation
- Random Resetting
- Swap Mutation
- Scramble Mutation
- Inversion Mutation

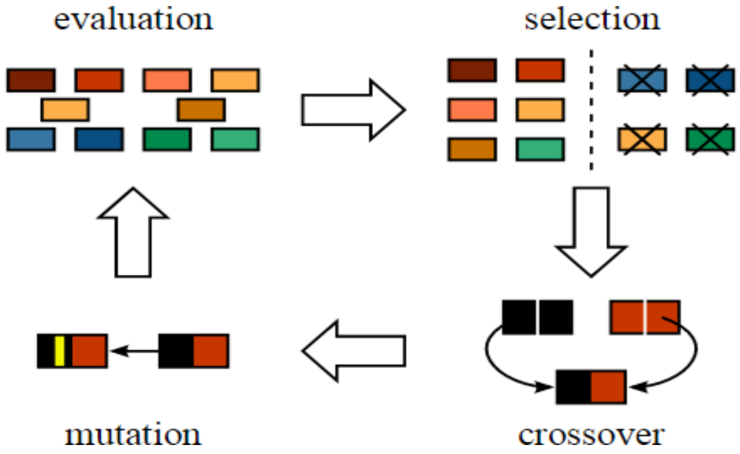


Figure 3.2 Process of Genetic Algorithm [21]

Fig 3.2 shows the process of Genetic algorithm. First evaluation of chromosome using fitness function and select the chromosomes based on Roulette wheel selection .After selection crossover is applied and lastly the result of crossover passed to mutation to produce new offspring’s.

3.6.1 Application of genetic algorithm

Genetic Algorithm is mainly used in optimization problem wherein we have to minimize ormaximize the given objective function under given set of constraints. Genetic algorithm canalso be used in various applications includes:

1. Vehicle Routing Problems
2. Neural Networks
3. Machine Learning
4. Travelling Salesman Problem
5. DNA Analysis

3.6.2 Example of MaxOne using genetic algorithm

Problem Statement: Suppose we want to maximize the number of one’s in a string of binary digits.

Step 1: Initialization and calculate the fitness value

Divide it into 10 slots and number of iterations is 6 *i.e.* from S1 to S6 and calculate the fitnessvalue F *i.e.* no of one’s in particular row

Table 3.1 Initialization of chromosomes of MaxOne problem.

S1	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**
S2	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**F=5**
S3	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**
S4	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**F=4**
S5	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**F=8**
S6	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**F=3**

Table 3.1 shows the initialization of chromosomes of MaxOne problem and fitness value of allthe chromosomes. The fitness value is calculated based on number of one’s occur in particulariteration. After calculate the fitness value of all chromosomes the chromosomes is arranged according to fitness value and selection is done based on roulette wheel

selection. The total number of one occur is 34 out of 60. Our aim is to increase the number of one's so we use the genetic algorithm for optimization of MaxOne problem.

Step 2: Selection of chromosomes using roulette wheel selection

In step 2 Selection is done based on roulette wheel selection method. In Roulette wheel selectionmethod the chromosomes are selected based on fitness value of chromosomes. The Chromosomes with high fitness value have probability to be selected first. The selected chromosome can be used for crossover and mutation to produce new population. Fig 4.3 showsthe pie chart representation of roulette wheel selection based on fitness value.

2.1 Arrange according to the fitness function:

Table 3.2 Arrangement of Chromosomes based on fitness value

S1	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**
S3	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**
S5	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**F=8**
S2	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**F=5**
S4	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**F=4**
S6	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**F=3**

Table 3.2 represent the arrangement of chromosomes based on fitness value. We randomly select the chromosomes using roulette wheel selection for crossover and mutation.

2.2 Randomly select chromosomes for crossover and mutation to produce new off springs.

Table 3.3 Crossover of S1 and S3 chromosome.

S1	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**
S3	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**

Table 3.3 represent the chromosomes S1 and S3 for crossover. The two point crossover isperformed on chromosomes S1 and S3 to increase the probability of number of one's.

Table 3.4 Crossover result of chromosome S1 and S3

S1	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=7**
S3	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=7**

Table 4.4 represent the crossover result of S1 and S3.After applying crossover on S1and S3 thenumber of one's increases.

2.3 Perform crossover on S2 and S4 chromosome.

Table 3.5 Crossover of S2 and S4 chromosome

S2	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**F=5**
S4	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**F=4**

Table 3.5 represent the chromosomes S2 and S4 for crossover. The two point crossover isapplied to increase the probability of number of one's.

Table 3.6 Crossover Result of S2 and S4 chromosome

S2	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**F=4**
S4	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**F=5**

Table 3.6 represents the crossover result of S2 and S4.After applying crossover on S2 and S4 the number of one's increases.

2.4 Perform crossover on S5 and S2 chromosome.

Table 3.7 Crossover of S5 and S6

S5	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**F=8**
S6	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**F=3**

Table 3.7 represent the chromosomes S5 and S6 for crossover. The two point crossover isapplied to increase the probability of number of one's.

Table 3.8 Crossover result of S5 andS6

S5	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**F=4**
S6	**0**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**F=8**

Table 3.8 represent the crossover result of S5 and S6.After applying crossover on S5 and S6 thenumber of one's increases.

Step 3: Perform mutation on the crossover results so that we obtain a better result.

3.1 Mutation means to change the particular bit of the chromosomes to increase the number ofone's.

Table 3.9 Mutation result of chromosomes

S1	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**F=8**
S3	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=8**
S5	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**F=5**
S2	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**F=5**
S4	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**F=5**
S6	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**F=8**

Table 3.9 shows the result of mutation of chromosomes of S1 to S6.After applying mutation onS1 to S6 the probability of number of one's increases from 34 to 39.

Conclusion: The optimization algorithm give better results when number of iteration is more. So we have to increase the number of iteration to get better results.

3.7 Proposed Algorithm (Genetic K-means Algorithm)

The design of proposed algorithms is as follows:

This algorithm focus on random selection of initial seed problem of k means clustering. Geneticalgorithm is used to select the optimum initial centroid for better results. Steps of Genetic K- means (GAKM) is as follows:

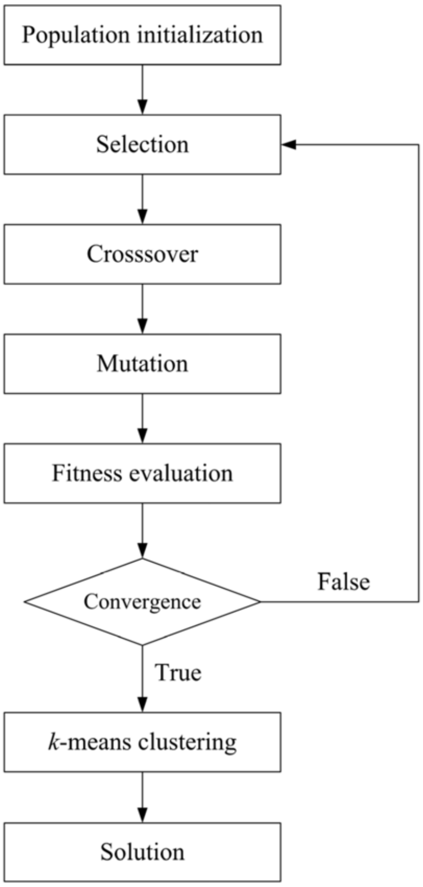


Figure 3.3 Flowchart of Proposed Algorithm

Fig 3.3 shows the flowchart of proposed Genetic K means clustering. First initialize the population in the form of strings and select the chromosomes based on roulette wheel selection.After selection of chromosomes the result is passed to crossover .By applying two point crossover to produce new offspring's the result is passed to mutation .After applying mutation on particular chromosome the final chromosomes is passed to K-means clustering to perform clustering. After clustering the evaluation of clusters is done based on Davies Bouldin index value. Higher the value of Davies Bouldin index indicates the good clustering.

Steps of Proposed Algorithm:

i) Initialization:

In the initialization step of Genetic K-means clustering the parameter of the dataset is prearranged in the strings (called chromosomes).The formation of chromosomes is the crucial step in selecting the initial centroid. Initially K centroids are selected for K random clusters. Initial population corresponds to Z where Z is the number of centroids that are randomly selectedfrom normalized dataset. Z is equal to (pop_size*K) where K is the quantity of groups to be framed.

ii) Chromosome Length:

Chromosome length in population is equal to (K*mv) where K is the number of clusters and mvis the number of variables or attributes in the dataset.

iii) Initial Population Size:

Initial population measure relates pop_size (no of rows) K*mv (no of attribute) and the pop_size *K (number of centroids) are actually chosen for initial population.

iv) Fitness Function:

Determining a fitness function is the crucial step. The objects are clustered based on Euclidian distance, each object belong to cluster whose centroid to object Euclidian distance is minimum.The objective is to maximize the inter-cluster distance and minimize the intra cluster distance, so the fitness function is evaluated by dividing the sum of intra cluster distance and inter-clusterdistance.

Formula for finding the Euclidean distance:

$$d(p, q) = \sqrt{\sum (q_i - p_i)^2}$$

Where p and q are the points the dataset and d is the distance between the data points.

Formula for finding Intra Cluster distance:

$$D_{\text{INTRA}}^{q}(x_i, x_j) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} \left(x_{i-}x_j\right)^2} /(m * m)$$

The total Intra cluster distance is:

$$S(\mathrm{D}_{\text{INTRA}})=\sum_{q=1}^{k}(\mathrm{D}_{\text{INTRA}})$$

Where $S(\mathrm{D}_{\text{INTRA}})$ is the sum of Intra-cluster distance. The intra-cluster distance is obtained bycalculating the distance between data points in the same cluster.

Formula for finding Inter Cluster distance:

$$D_{\text{INTER}}^{q.r}(x_i, x_j) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \left(x_{i-}x_j\right)^2 /m * n}$$

Where $D^{q.r}$ is the distance between the neighboring clusters and xi and xj are the points of the clusters m and n are the data points of the neighboring clusters.

Total inter cluster distance is:

$$\mathrm{S}(D_{\text{INTER}}) = \sum_{q=1}^{k-1} \sum_{r=q+1}^{k} (D_{\text{INTER}}^{q,r})$$

Formula for finding Fitness value

$$\mathrm{F}_{\text{max}} = \max(\mathrm{S}(\mathrm{D}_{\text{INTER}})/\mathrm{S}(\mathrm{D}_{\text{INTRA}}))$$

Where F_{max} is the fitness function obtained by dividing the total inter cluster distance and total intra cluster distance.

v) Crossover Operator:

After the selection using rank based selection the next step is to produce off springs. The mainlyused solution is crossover. Different types of crossover are single point crossover, two point crossover, multiple point crossover .Single point crossover gives better result for Integer and real datasets. Crossover operators are applied to maintain the genetic diversity. Genetic diversityis the crucial step for the process of evolution. Crossover operator is applied to the one of the genetic operator to maintain the genetic diversity. Different types crossover operator are used like one point crossover, two point crossover and multiple point crossover based on the requirement.

Formula for crossover is:

$$\text{Offspring } 1 = (\alpha * \text{parent } 1) + ((1-\alpha) * \text{parent}2)$$

$$\text{Offspring } 2 = ((1-\alpha) * \text{parent } 1) + (\alpha * \text{parent } 2)$$

vi) Mutation:

Mutation is the genetic operator used to preserve the genetic diversity from one generation to next generation. There were different genetic operator like bitwise operator, uniform operator, on Uniform operator and Gaussian operator. Uniform mutation operator is used for real and integer dataset as it gives better results. Mutation is applied so as to obtained the better accuracyand it is applied to (Pm*pop size* u) where Pm is the probability of population and u is the chromosome length.

vii) K means Algorithm:

After initialization of first centroid using genetic algorithm and applying crossover and mutationfor better result .The resulting chromosomes will pass to the K-means clustering algorithm to find the optimal no of clusters that are less sensitive to outliers and with highest accuracy as compared to k means clustering.

viii) Evaluation of cluster using Davies Bouldin index.

After k means clustering when clusters are formed we have to check the value of Davies Bouldinindex for evaluation of clusters. Davies Bouldin index is used for evaluation of clusters whetherthe clusters formed are optimal or not. The lower value indicate that clustering is good.

Formula for Davies Bouldin:

$$S_i=(\frac{1}{T_i}\sum_{j=1}^{T_i}|X_j-A_i|^p)^{1/p}$$

Where $A_i$the centroid of cluster isC_i, T_i is the size of cluster and $S_i$is the measure of validity of cluster.

3.7.1 Example of Proposed Algorithm (Genetic K-means)

Step 1: The sample Iris dataset is used for performing Genetic K-means algorithm. The irisdataset consist of three species Setsoa, versicolor and Virginica and four dimensions sepal length, sepal width, petal length and petal width.

Table 3.10 Iris dataset for Genetic K-means clustering

Sepal Length	Sepal Width	Petal Width
10	20	10
12	18	8
11	21	11
9	20	9
10	17	11
40	50	60
42	48	58
41	51	59
38	47	60
40	52	57
80	100	120
81	101	119
78	98	118
80	100	121
82	102	120

Table 3.10 shows the iris dataset used for Genetic K-means clustering. The iris dataset consist of three dimensions sepal length, sepal width and petal length. Genetic Algorithm can be appliedto iris dataset to find accuracy, precision, recall and sensitivity. Davies

Bouldin index is used forevaluation of clusters. The first 15 rows of iris dataset is collected to perform genetic K-means clustering.

Step 2: Normalization of iris dataset to obtain maximum accuracy.

In this step normalization is applied to iris dataset to get accurate results. Unnormalized data doesn't give accurate result. Conversion of data from Unnormalized to normalize gives accurateresult and promotes easiness of data handling. Table 3.11 shows the normalized data to performcomputation.

Table 3.11 Normalized dataset for Genetic K-means Clustering

Sepal length	Sepal Width	Petal Width
0.222222	0.625	0.067797
0.166667	0.416667	0.067797
0.111111	0.5	0.050847
0.083333	0.458333	0.084746
0.194444	0.666667	0.067797
0.305556	0.791667	0.118644
0.083333	0.583333	0.067797
0.194444	0.583333	0.084746
0.027778	0.375	0.067797
0.166667	0.458333	0.084746
0.305556	0.708333	0.084746
0.138889	0.583333	0.101695
0.138889	0.416667	0.067797
0	0.416667	0.016949
0.416667	0.833333	0.033898

Table 3.11 represents the normalized iris dataset obtained by applying normalization on unnormalized dataset. The normalized dataset give accurate results and easiness of handling.

Step 3: After normalization of dataset:

Let assume pop size=4, k=3

Rows actually selected are X= (pop size *K) i.e. 4*3 rows are actually selected from dataset. Let Y be the 12 indices returned from the dataset

Y= [4, 12, 14, 7, 9, 1, 2, 3, 5]

The rows correspond to first 3 indices represent first chromosome and the first index represent first Centroid

Step 4: Selected row indices and chromosomes of the dataset.

Table 3.12: Selected Row indices and chromosomes

Column Number	Selected row indices	Chromosomes

1	{4,12,14}	{0.083,0.45,0.084}
		{0.138,0.583,0.101}
		{0,0.416,0.169}

2	{7,9,1}	{0.083,0.583,0.0677}
		{0.0277,0.375,0.0677}
		{0.2222,0.625,0.0677}

3	{2,3,5}	{0.1666,0.41666,0.0677}
		{0.1111,0.5,0.050}
		{0.19444,0.666,0.0677}

4	{13,11,15}	{0.1388,0.41666,0.0677}
		{0.3055,0.7083,0.084}
		{0.41666,0.8333,0.0338}

Table 3.12 shows the selected row indices and chromosomes of the dataset. In column No. 1where, Ist Centroid (C1): 0.083, 0.45, 0.084 represent Ist cluster. 2nd Centroid (C2): 0.138, 0.583, 0.101 represent 2nd cluster and 3rd Centroid (C3): 0, 0.0416, and 0.169 represents 3rd cluster.

Step 5: Calculate the Euclidean distance of point 1 {0.2222, 0.625, and 0.0677} from dataset to cluster 1{0.083, 0.45, and 0.084}

Distance from cluster 1 to object 1=0.0163
Distance from cluster 2 to object 1 =1.50
Distance from cluster 3 to object 1=1.66

Table 3.13: Calculated distance and Assignment of Clusters

Object dataset	Distance to cluster 1	Distance to cluster 2	Distance to cluster 3	Assignment of clusters
1	**0.00163**	**1.5**	**1.66**	**1**
2	**1.2**	**0.0163**	**1.18**	**2**

Table 3.13 represents the calculated distance obtained from cluster 1, cluster 2 and cluster 3 andalso the assignment of clusters.

Step 6: After calculating all the distance of clusters assignment of cluster of is shown in table

Table 3.14: Clusters obtained for fifteen records

1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**3**	**2**	**3**	**3**	**2**

Step 7: Crossover: Crossover is applied to two parents to produce new offspring's.

Formula for finding crossover is

Offspring 1= (α *parent 1) + ((1- α)*parent2)

Offspring 2 = ((1-α)*parent 1) + (α*parent 2)

Where α is crossover rate to produce new off springs. The crossover rate α is taken as 0.6.

- **Crossover of Parent 1 and Parent 2 chromosomes**

Parent 1: {0.083, 0.583, 0.0677} {0.0277, 0.375, 0.0677} {0.2222, 0, 0.625, 0.0677}

Parent 2: {0.1666, 0.4166, 0.677} {0.1111, 0.5, 0.050} {0.1944, 0.6666, 0.0677}

Parent 1 and Parent 2 are selected from selected row indices and chromosomes to produce newoff springs. The crossover rate α is taken as 0.6.The off springs produces are given below.

- **Generation of off springs**

  Offspring 1= (α *parent 1) + ((1- α)*parent2)

  (0.083*0.6 + (1-0.6)*0.1666) =0.02055

  Offspring 2 = ((1-α)*parent 1) + (α*parent 2)

  ((1-0.6)*0.083 + (0.6 *0.1666)) =0.2877

  Offspring 1:0.02055, 0.1530, 0.1770, 0.1754, 0.1694, 0.20000, 0.01374, 0.0131, 0.0230

  Offspring 2:0.02877, 0.2235, 0.2655, 0.2493, 0.2306, 0.2867, 0.0137, 0.0212, 0.0212

  The offspring 1 and offspring 2 are obtained by applying crossover on selected row indices 2and 3.

Step 8. Mutation: Apply mutation on 5th element of 1st chromosome

Parent 3: 0, 0.0353, 0.0088, 0.9863, 0.9882, 0.9726, 0.9765, 1.0000

Offspring 3: 0, 0.0353, 0.0088, 0.9863, 0.9882, 0.7982, 09763, 1.0000

Parent 3 is selected from selected from selected row indices and chromosome table to maintain the genetic diversity. Mutation is simply the random tweak to a particular chromosome. It is applied to 5th chromosome of parent 3.After apply mutation on 5th chromosome the value of chromosome changed from 0.9726 to 0.7982

Chapter 4 Experiments and Result Analysis

This chapter of project report will focus on implementation of the proposed hybrid algorithm Genetic K-means using python and MATLAB.

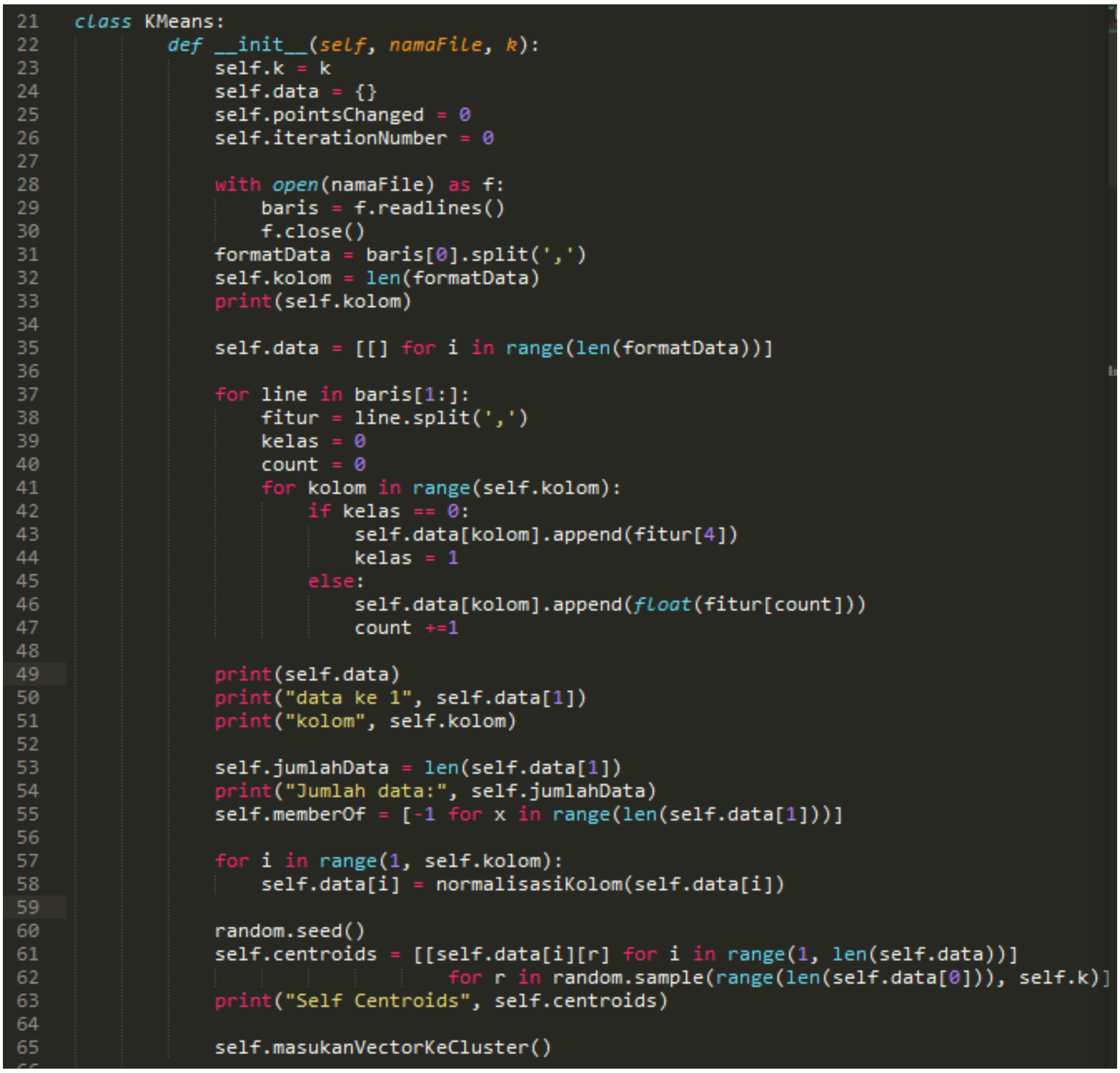
4.1 Implementation of Proposed Technique

To implement the proposed approach iris dataset is used. The dataset is collected from UCI Machine Learning Repository [42]. The proposed approach is implemented by Python using google collab.

4.1.1 Iris Dataset for Implementation:

The Iris dataset is a multi-dimensional dataset consisting of three classes (setsoa, versicolor and virginica) and four dimensions (sepal length, petal length, sepal width and petal width).This dataset consists of integer data points. K-means clustering and Genetic K-means are implemented by using this dataset to find optimal clusters. Confusion matrix, Accuracy, Recall and Precision are calculated.Davies Bouldin index is used for evaluation of each cluster. The snapshots below show the implementation of K-means and Genetic K-means on this dataset and also the calculated value of Davies Bouldin Index for evaluation of each cluster. The Snapshots for code and results of K- means clustering is shown in figure 4.1 and figure 4.2.

I. Code of K-means Clustering:

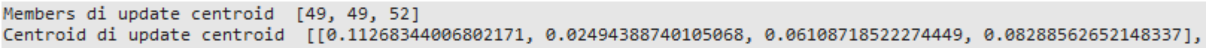


```python
class KMeans:
    def __init__(self, namaFile, k):
        self.k = k
        self.data = {}
        self.pointsChanged = 0
        self.iterationNumber = 0

        with open(namaFile) as f:
            baris = f.readlines()
            f.close()
        formatData = baris[0].split(',')
        self.kolom = len(formatData)
        print(self.kolom)

        self.data = [[] for i in range(len(formatData))]

        for line in baris[1:]:
            fitur = line.split(',')
            kelas = 0
            count = 0
            for kolom in range(self.kolom):
                if kelas == 0:
                    self.data[kolom].append(fitur[4])
                    kelas = 1
                else:
                    self.data[kolom].append(float(fitur[count]))
                    count +=1

        print(self.data)
        print("data ke 1", self.data[1])
        print("kolom", self.kolom)

        self.jumlahData = len(self.data[1])
        print("Jumlah data:", self.jumlahData)
        self.memberOf = [-1 for x in range(len(self.data[1]))]

        for i in range(1, self.kolom):
            self.data[i] = normalisasiKolom(self.data[i])

        random.seed()
        self.centroids = [[self.data[i][r] for i in range(1, len(self.data))]
                            for r in random.sample(range(len(self.data[0])), self.k)]
        print("Self Centroids", self.centroids)

        self.masukanVectorKeCluster()
```

Figure 4.1 Code of K-means Clustering on Iris Dataset.

Fig 4.1 Shows the code of K-means clustering on Iris dataset. It shows implementation of K-means clustering and uses the iris data and value of K used is 3. This code predicts the accuracy of K- means. The predicted clusters are compared with actual class clusters and accuracy of iris dataset is evaluated .The accuracy obtained from K-means clustering is 45 %.

Centroid Update Results:



```
Members di update centroid  [49, 49, 52]
Centroid di update centroid  [[0.11268344006802171, 0.02494388740105068, 0.06108718522274449, 0.08288562652148337],
```

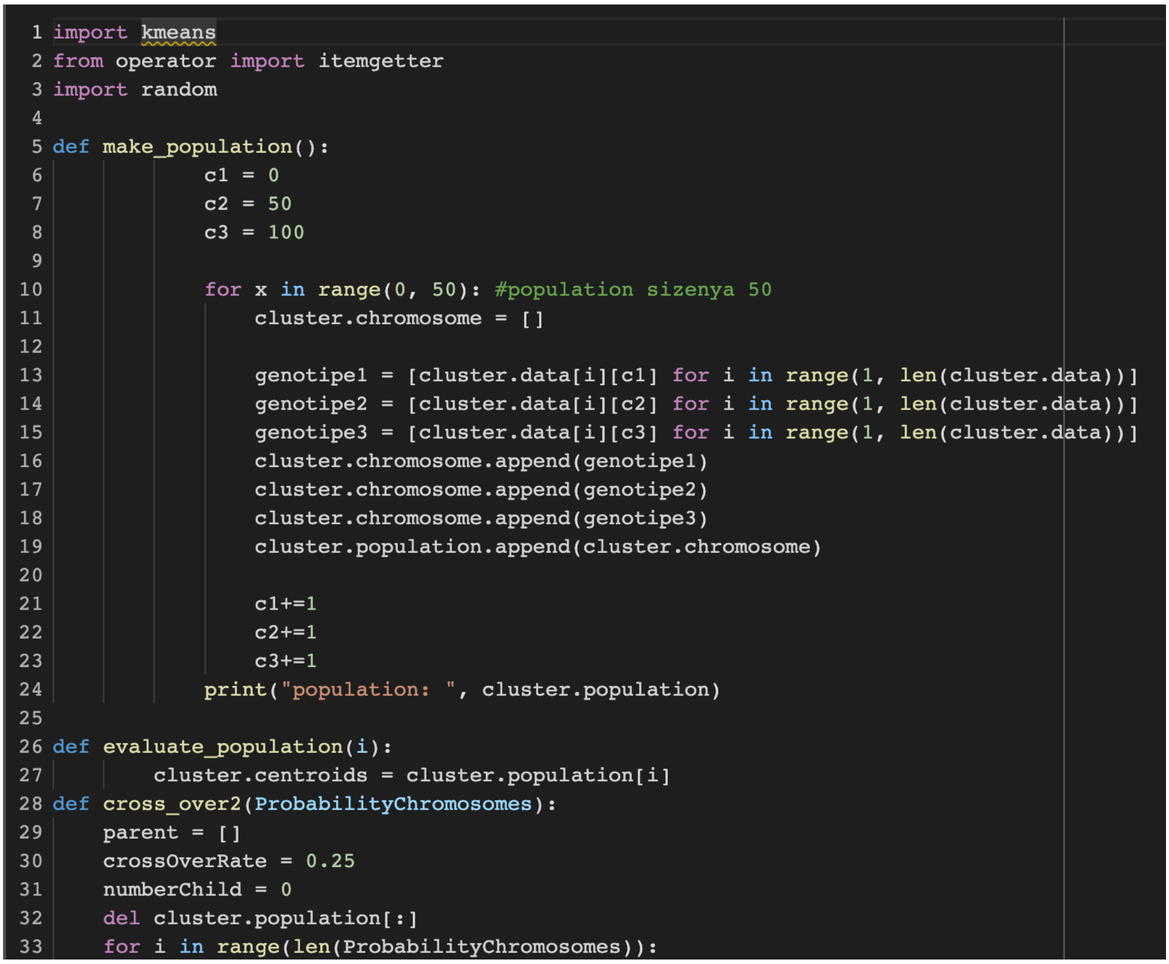
This means that in each cluster there are as many members→[49,49,52]

Update Centroid as given in fig 5.1 generates a new centroid for cluster 1

```
[[0.11268344006802171, 0.02494388740105068, 0.06108718522274449, 0.08288562652148337],
```

Data is in the same range after normalization

II. Code of Genetic K-means clustering on Iris dataset



```python
import kmeans
from operator import itemgetter
import random

def make_population():
            c1 = 0
            c2 = 50
            c3 = 100

            for x in range(0, 50): #population sizenya 50
                cluster.chromosome = []

                genotipe1 = [cluster.data[i][c1] for i in range(1, len(cluster.data))]
                genotipe2 = [cluster.data[i][c2] for i in range(1, len(cluster.data))]
                genotipe3 = [cluster.data[i][c3] for i in range(1, len(cluster.data))]
                cluster.chromosome.append(genotipe1)
                cluster.chromosome.append(genotipe2)
                cluster.chromosome.append(genotipe3)
                cluster.population.append(cluster.chromosome)

                c1+=1
                c2+=1
                c3+=1
            print("population: ", cluster.population)

def evaluate_population(i):
        cluster.centroids = cluster.population[i]
def cross_over2(ProbabilityChromosomes):
    parent = []
    crossOverRate = 0.25
    numberChild = 0
    del cluster.population[:]
    for i in range(len(ProbabilityChromosomes)):
```

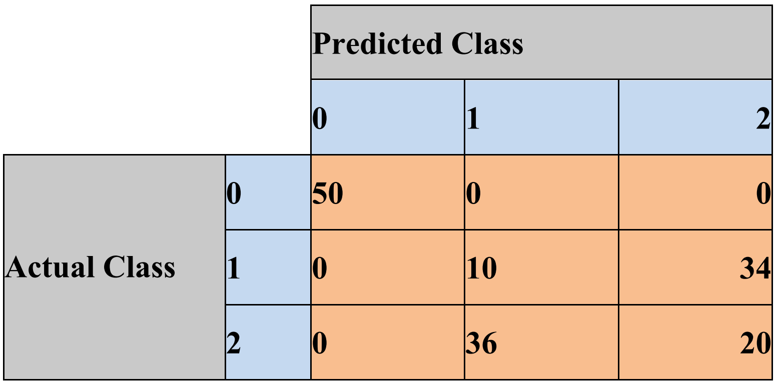
Figure 4.2 Code of Genetic K-means Clustering on Iris Dataset.

4.2 Experimental Results

The experiments performed on these algorithms are on the basis of many parameters that is confusion matrix, Davies Bouldin index value ,accuracy , Recall, precision, Sensitivity and Specificity.

4.2.1. Confusion Matrix of K-means Clustering

Both the algorithms that is K-Means and Genetic K-means algorithm are tested on the iris datasets to calculate accuracy. Figure 4.3 shows the comparison of the confusion matrix obtained by each algorithm on Iris dataset.



		Predicted Class		
		0	1	2
Actual Class	0	50	0	0
	1	0	10	34
	2	0	36	20

Fig 4.3:Confusion Matrix obtained from K means Algorithm

Fig 4.3 shows the Confusion matrix obtained from K-means clustering.Accuracy Recall, Precision, Sensitivity and Specificity are calculated by comparing the actual and predicted results.

- **Accuracy: (TP+TN)/TP+TN+FP+FN = 50+10+20/50+10+20+36+34=80%**
- **Misclassification rate =(FP+FN)/Total =0+0+0+34+0+36/150= 0.46 or 46%**
- **Precision:.TP/Predicted 1=10/46=0.21 or 21%**
- **Recall: TP/Actual 1=10/44=0.22 or 22%**
- **Sensitivity:**Sensitivity is also known as Recall =**TP/Actual 1=10/44= 22%**
- **Specificity:**Specificity is also known as precision =**TP/Predicted 1=10/46= 21%**

4.2.2 Confusion Matrix of Genetic K-means Clustering

Both the two algorithms that is K-Means and Genetic K-means algorithm are tested on the iris datasets to calculate accuracy. Fig. 4.4 shows the comparison of the confusion matrix obtained by each algorithm on Iris dataset.

		Predicted Class		
		0	**1**	**2**
Actual Class	**0**	**50**	**0**	**0**
	1	**0**	**48**	**2**
	2	**0**	**14**	**36**

Fig 4.4: Confusion Matrix obtained from Genetic K means Algorithm

Fig 4.4 shows the Confusion matrix obtained from Genetic K-means clustering.Accuracy Recall, Precision, Sensitivity and Specificity are calculated by comparing the actual and predicted results.

- **Accuracy:TP+TN/TP+TN+FP+FN = 50+48+36/50+48+36+14+2= 89%**
- **Missclassification rate or Error rate=FP+FN/Total =0+0+0+2+14/150= 10%**
- **Precision:.TP/Predicted 1=48/62=77%**
- **Recall: TP/Actual 1=48/50=0.96 or 96%**
- **Sensitivity: TP/Actual 1=48/50=0.96 or 96%**
- **Specificity:.TP/Predicted 1=48/62=0.77 or 77%**

4.2.3 Test for Performance of Accuracy

Both the algorithms that is K-Means algorithm and proposed approach are tested on iris dataset to calculate the accuracy performance. Accuracy evaluation is needed to maintain the quality of clusters.

The formula for calculating accuracy

$$\text{Accuracy}=\frac{Tp+TN}{TP+TN+FP+FN}$$

Where accuracy is obtained by dividing the sum of TP (True Positive) and TN (True Negative) by thetotal sum. *i.e.* TP (True Positive), TN(True Negative),FP(False Positive) and FN(False Negative).

The accuracy obtained from K-means and Genetic K-means shows that the proposed algorithm givesbetter accuracy than K-means clustering .The accuracy obtained from all the four datasets is using K-means and Genetic K-means is shown is table 4.1

Table 4.1: Accuracy obtained from K means and Proposed Algorithm

Dataset	K-means	Proposed Algorithm
Iris Dataset	**66%**	**83%**

Table 4.1 Shows the table of accuracy obtained from K-means and Genetic K-means Algorithm. The table shows that Genetic algorithm gives higher accuracy thanK-means algorithm.

4.2.4 Calculation of Intra cluster distance using K means and proposed algorithm.

Intra cluster distance is the distance between data points in the same cluster. The objective of k-meansand proposed algorithm is to minimize the intra cluster distance. The formula for finding intra clusterdistance is

$$D^q_{\text{INTRA}}(x_i, x_j) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} \left(x_{i-}x_j\right)^2 /m * m}$$

Where m is the no of elements, x_i and x_j are data points of clusters and D^q is the distance between data points in same cluster.

The Intra cluster distance obtained from k means and proposed algorithm is shown in table 4.2 and table 4.3.

Table 4.2: Intra cluster distance using K-means algorithm

Cluster No	Intra Cluster Distance Using K-means
1	5.927
2	1.0245
3	1.086

Table 4.2 shows the table of intra-cluster distance using K-means algorithm.The intra-cluster distanceis obtained by calculating the distance between data points in the same cluster. The distance obtained by K-means algorithm is more than Genetic Algorithm.

Table 4.3 :Intra Cluster distance using proposed algorithm

Cluster No	Intra Cluster Distance Using Proposed algorithm
1	4.9275
2	0.0284
3	0.0861

Table 4.3 shows the table of intra-cluster distance using proposed algorithm. The intra-cluster distanceis obtained by calculating the distance between data points in the same cluster. The distance obtained by proposed algorithm is less than K-means algorithm.

4.2.5 Calculation of Inter cluster distance using K means and proposed algorithm.

Inter cluster distance is the distance between data points of corresponding cluster. The main objectiveof K-means and proposed algorithm is to maximize the inter cluster distance. The inter cluster distanceobtained from k means and proposed algorithm is shown in table.

Formula for finding Inter cluster distance:

$$D_{\text{INTER}}^{q.r}(x_i, x_j) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \left(x_{i-}x_j\right)^2} \, /(m * n)$$

Where m, n are the number of elements in the $q^{th}$and r^{th} cluster, x_i and x_j are the elements inclusters and $D_{\text{INTER}}^{q.r}$ is the inter cluster distance between data points.

The Total Inter Cluster distance is:

$$\text{S}(D_{\text{INTER}}) = \sum_{q=1}^{k-1} \sum_{r=q+1}^{k} (D_{\text{INTER}}^{q,r})$$

Table 4.4: Inter Cluster distance using K means clustering

Cluster 1-Cluster 2	Inter Cluster Distance Using K-means
1	1.3585
2	1.3203
3	0.3323

Table 4.4 shows the table of Inter-Cluster distance using K-means Algorithm. The table shows thevalue of distance between data points of one cluster to another cluster. The calculated value is lesser than proposed algorithm.

Table 4.5: Inter Cluster distance using proposed algorithm

Cluster 1 – Cluster 2	Inter Cluster Distance Using Proposed algorithm
1	1.585
2	1.4505
3	0.5823

Table 4.5 shows the table of Inter-Cluster distance using proposed Algorithm. The table shows thevalue of distance between data points of one cluster to another cluster. The calculated value is higher than the proposed algorithm.

Chapter 5 Conclusion and Future Scope

This chapter discusses the conclusion of the work done in the project and ends with a clear vision of future direction which can be taken further.

5.1 Conclusion

This project introduces big data and provides background of various clustering techniques used to analyzebig data. In this work comparative analysis of these techniques is done. A hybrid approach based on Genetic K-Means is effective grouping of enormous information is proposed. This approach is developed in python. The experimental results have been gathered which shows that the proposed approach is more accurate as compared to K-Means when tested on dataset.

5.2 Limitations

The proposed approach still has some of the following limitations.

- The proposed approach still requires the value of K, but after finding the Davies Bouldin index valuewe will find the number of initial or desired clusters as input, though data points has been distributed.
- The proposed approach can be applied only for those data sets which have numerical values or attributes

5.3 Future Scope

- The proposed approach shows that initial value of clusters is needed as input, in future the approach can be enhanced by finding the optimal no of cluster using best technique so that automatic number of desired clusters is formed.
- In future, this approach can be applied for a particular real time application area by addressing issuesinvolved and can be applied for data sets with categorical attributes.

References

[1] E. Ferrara, P. D. Meo, G. Fiumara and R. Baumgartner, “*Web Data Extraction, Applications and Techniques: A Survey*”, Knowledge Based Systems, Vol. 70, No. 3, pp. 301-323, 2014.

[2] P.Vats, M.Mandot and A.Gosain, (2014, January). A Comparative Analysis of Various ClusterDetection Techniques for Data Mining. In *Electronic Systems, Signal Processing and Computing Technologies (ICESC), 2014 International Conference on* (pp. 356-361). IEEE.

[3] S.Bandyopadhyay and U.Maulik. "An evolutionary technique based on K-means algorithm for optimalclustering in RN." *Information Sciences* 146.1-4 (2002): 221-237.

[4] AG.Picciano “The Evolution of Big Data and Learning Analytics in American Higher Education,” *Journal of Asynchronous Learning Networks,* Vol. 16, No.3, pp. 9-20, 2012.

[5] T. Hu, H. Chen, L. Huang and X. Zhu “A survey of mass data mining based on cloud-computing,”

Anti-Counterfeiting, Security and Identification (ASID), 2012 International Conference, 2012.

[6] M.Jain, K.Anil, M. N. Murty and P.J. Flynn, “Data clustering: a review,” *ACM computing surveys(CSUR)*, Vol. 31, No.3, pp. 264-323, 1999.

[7] “Methods for finding optimal number of clusters” [online] Available: http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal- number-of-clusters-3-must-know-methods/.[14 July 2014]

[8] C. Ordonez, “*Clustering Binary Data streams with K-Means*”, In Proceedings of the 8th ACMSIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 12-19, 2003.

[9] “Genetic Algorithm ”[Online] Available .https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3.[13 Jan 2016]

[10] M.Mor, P.Gupta and P.Sharma "A Genetic Algorithm Approach for Clustering." *InternationalJournal of Engineering & Computer Science* 3.6 (2014).

[11] K.Dharmendra and K. Sharma. "Genetic k-Means clustering algorithm for mixed numeric andcategorical data sets." *International Journal of Artificial Intelligence &*

Applications 1.2 (2010): 2328.

[12] K.Shahroudi and S.Biabani "Variable selection in clustering for market segmentation using genetic algorithms." *Interdisciplinary Journal of Contemporary Research in Business* 3.6 (2011): 333-341.

[13] E.O.Hartono and D. Abdullah. "Determining a Cluster Centroid of K-Means Clustering Using Genetic Algorithm." *International Journal of Computer Science and Software Engineering (IJCSSE)* 4.6 (2015).

[14] D.X.Chang, XD. Zhang, and CW. Zheng. "A genetic algorithm with gene rearrangement for K-meansclustering." *Pattern Recognition* 42.7 (2009): 1210-1222.

[15] R.Lletı, M.C.Ortiz, L.A.Sarabia, & M.S.Sánchez, (2004). "Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes". *Analytica Chimica Acta*, *515*(1), 87-100.

[16] P.Vats, M. Mandot and A. Gosain, "A Comparative Analysis of Various Cluster Detection Techniques for Data Mining," *Electronic Systems, Signal Processing and Computing Technologies (ICESC), 2014International Conference on*. IEEE, 2014.

[17] I.B. Saida, K. Nadjet and B. Omar "A new algorithm for data clustering based on cuckoo search optimization," *Genetic and Evolutionary Computing*, pp. 55-64. Springer, 2014.

[18] A.K. Jain, MN. Murty and P.J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, Vol. 31, No.3, pp. 264-323, 1999.

[19] Lu, Yi, et al. "FGKA: A fast genetic k-means clustering algorithm." *Proceedings of the 2004 ACMsymposium on applied computing*. ACM, 2004.

[20] A.Likas, N.Vlassis and J.J. Verbeek. "The global k-means clustering algorithm." *Pattern recognition* 36.2 (2003): 451-461.

[21] K. Kim and H. Ahn. "A recommender system using GA K-means clustering in an online shopping market." *Expert systems with applications* 34.2 (2008): 1200-1209.

[22] G.P. Babu and M.N. Murty. "A near-optimal initial seed value selection in k-means means algorithm using a genetic algorithm." *Pattern Recognition Letters* 14.10 (1993): 763-769.

[23] D.K. Roy and L. K. Sharma. "Genetic k-Means clustering algorithm for mixed numeric and categoricaldata sets." *International Journal of Artificial Intelligence & Applications* 1.2 (2010): 23-28.

[24] M.E. Celebi, H.A. Kingravi, and P. A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." *Expert systems with applications*40.1 (2013): 200-210.

Vatsal_Updated_Plag

ORIGINALITY REPORT

6%	2%	2%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Thapar University, Patiala Student Paper	4%
2	tudr.thapar.edu:8080 Internet Source	1%
3	"Computational Intelligence in Data Mining", Springer Science and Business Media LLC, 2020 Publication	<1%
4	pt.scribd.com Internet Source	<1%
5	Sione Paea, Christos Katsanos, Gabiriele Bulivou. "Information Architecture: Using K-Means Clustering and the Best Merge Method for Open Card Sorting Data Analysis", Interacting with Computers, 2022 Publication	<1%

Exclude quotes	On	Exclude matches	< 20 words
Exclude bibliography	On		