Modified K-Means to improve clustering using Genetic algorithm

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
In
Computer Science and Engineering

Submitted By

Vandana Setia

(Roll No.801632054)

Under the supervision of:

Dr. Vinay Arora

Assistant Professor
Computer Science and Engineering Department

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA – 147004

June 2018

Certificate

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"Modified K-means to improve clustering using Genetic algorithm"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of ***Dr. Vinay Arora*** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University. In case of any disperancy, even after publication of thesis, I take the full responsibility.

Signature:

Vandana Setia

(Vandana Setia)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge. In case of any disperancy, even found after submission of thesis, only candidate will be responsible.

(Dr. Vinay Arora)
Assistant Professor
Computer Science and Engineering Department

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. Vinay Arora** for their positive attitude, excellent guidance, constant encouragement, keen interest, invaluable co-operation, generous attitude and above all their blessings. He has been a source of inspiration for me.

I am grateful to **Dr. Maninder Singh**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology for the motivation and inspiration for the completion of this thesis. I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia**, Senior Professor and Dean of Academics Affairs in the institute for making provisions of infrastructure such as Library facilities, Computer Lab equipped with internet facility immensely useful for the learners to equip themselves with latest in the field.

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 thesis.

Vandana Setia
(801632054)

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 thesis, 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

List of Figures

List of Tables

Chapter 1 Introduction

An extensive amount of raw data can be accumulated from various fields yet this data remains useless until proper reasoning is carried out to obtain useful information. In this thesis, we focus on one of the important techniques in data mining: Clustering.

1.1 Introduction to Big Data

The term big data is considered as data that is large in size which includes everything from digital data to health data. Big data contains voluminous and complex data sets that traditional data processing application software are inadequate to deal with large amount of data. Information industry generates large amount of data every day [1]. The concept of how data became large begin seventy years ago when the growth rate in volume of data was known as information explosion. In decade of 90s, IBM introduced the relational database concept in which data can be stored in tables and can be analyzed easily by using different analysis techniques. This large voluminous data is meaningless until it is transformed into some valuable information. Knowledge discovery in database or KDD is the process of discovering useful information from large amount of data. Data mining is the essential step of KDD [2]. Data mining is the process of examine large pre-existing database in order to generate new information. There are various techniques used for data mining among them clustering is the most effective data miming technique. K-means is most commonly used clustering technique. K-means is sensitive to anomalous data points and sometimes it form empty clusters .To overcome this problem we proposed a new technique Genetic K-means algorithm. This thesis work concentrate on Genetic K-means clustering technique.

1.2 Knowledge Discovery in database

One of the most commonly used data mining technique is Knowledge discovery in database or KDD. It is a process of determining valuable knowledge from large amount of data. The principle steps in KDD include [2]:

- *Data Cleaning*: It is the first step of KDD. Data coming from various sources is raw, inaccurate and inconsistent. KDD focuses on converting it into useful information which is accurate and consistent by removing or modifying the noisy data.
- *Data Integration*: This is the crucial step of KDD. Data obtained from different sources is heterogeneous is nature which has to be integrated. Data Integration and data warehousing are used for integration of data.
- *Data* Selection: In data selection data appropriate to analysis task are retrieved from database. Data mining tasks are used to transform the selected data.
- *Data Transformation:* Data after cleaning is not suitable for mining. So we need to reconstruct the data into forms suitable for mining.
- *Data Mining*: This is an essential process where intelligent methods are applied to extract data patterns. The mining and analysis of big data is a big challenge due to its heterogeneous nature. Various intelligent methods like clustering and association rule mining are used to essence data patterns.
- *Pattern Evaluation:* Pattern evaluation means to identify interesting patterns representing knowledge.
- *Knowledge Presentation*: In the last after data mining and pattern recognition visualization and knowledge representation techniques are used represent mined data to users.

Fig 1.1 shows the elements of data mining and knowledge discovery from database. The steps used to extract knowledge from data is explained in figure.

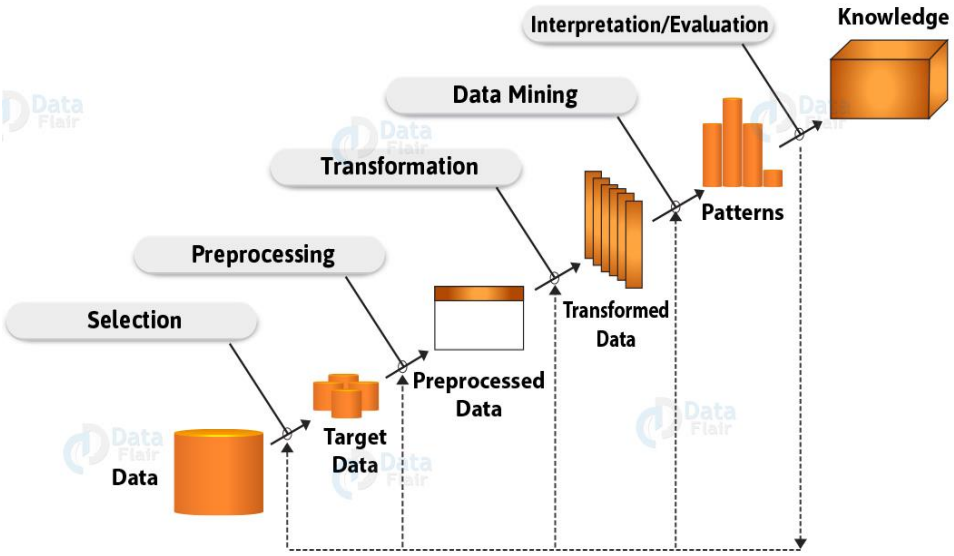

Figure 1.1: Elements of Data Mining and Knowledge discovery [2]

1.3 Data Mining

Data Mining is a process of auditing hidden patterns of data from different perspective and encapsulate it into useful information [3]. Data mining term first appeared in 1990's .Data mining includes effective data collection and warehousing as well as processing. This also recognized as Knowledge discovery in database (KDD).The agenda of data mining consist of three stages: (1) Exploration, (2) Model building or pattern identification and 3) deployment. Various techniques used for data mining includes.

- Association
- Classification
- Clustering

1.3.1Association Rule Mining:

In data mining, the main goal of association rule mining is to discover relationship between items from large database [3]. This technique is secondhand by retailers to analyze customer buying habits. For example, if a client is purchasing nutty spread from a shopping store, he /she is likewise prone to buy jam with that. To increase this possibility, nutty spread and jam can be

put together. Association rules are effective rules to determine interest of customers towards several products. The main utilization of association rule mining are:

- Basket Data Analysis
- Cross Marketing
- Catalog Design

1.3.2 Classification:

Classification is a data mining approach which pigeonhole the data items in predefined classes. In classification class labels are already defined [4] . It is also termed as supervised machine learning technique. In this we build a software using machine learning models that can learn how to classify the data items into different group. For example classification model can be used to identify employees as leave or stay in the company.

1.3.3 Clustering:

Clustering is an unsupervised technique that creates clusters of objects having similar features using automatic technique [5] . It is also termed as classification by statisticians. In clustering, grouping is done based on similarity of objects .The objects with similar properties are grouped in 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.

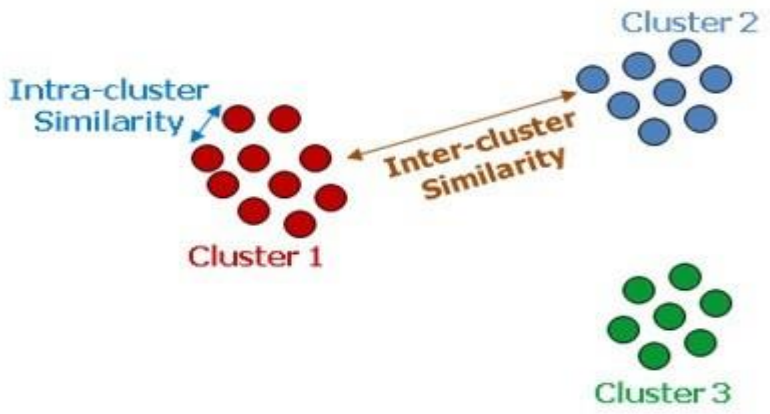

Figure 1.2 Inter and Intra similarities of cluster [5]

1.4 Types of Clustering

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

1.4.1 Partitioning methods

In Partitioning clustering the set of data objects are divided into non-overlapping clusters. Clusters of data objects are found by using greedy approaches [6]. Data points are separated into several subsets using data partitioning methods. Iterative process is applied in order to find most appropriate clusters. These algorithms provide high quality grouping of data points. For these algorithms, the number of clusters (k) should be known beforehand which are provided as an input to the algorithm. The heuristic methods for partitioning algorithm includes:

K-means clustering: K-means clustering was introduced by Macqueen in 1967.It decide the clusters depend on the value of k. The value of k is already defined or predefined by the analyst. It is an unsupervised machine learning technique [6]. In K-means clustering each data point is clustered based on the similarity measure. Although K-means clustering is the most powerful method but it is sensitive to anomalous data points and sometimes gives empty clusters for large datasets.

K-medoids clustering: K-medoids algorithm also known as partitioning around Medoids (**PAM**) was introduced by Kaufman and Rousseau in 1990.It is a clustering approach similar to K-means clustering for partitioning of dataset into k groups [6] . K-medoid is robust alternative to K-means which means K-medoid is less precise to noise and outliers as compared to K-means.

Clustering Large Applications: Clustering large application also termed as CLARA is an extension of K-medoid [7].It was coined by Kaufman and Rousseau in 1990.It is used for data consisting of large number of objects.

1.4.2. Hierarchical Clustering:

Hierarchical clustering is also termed as HCA. It groups similar objects into a cluster. It doesn't require the value of k or number of clusters to be pre-specified [7]. Tree based structure is obtained from the result of hierarchical clustering.

The two types of Hierarchical clustering are:

- Agglomerative Clustering.
- Divisive Clustering.

1.4.3. Fuzzy Clustering:

Soft clustering is also known as fuzzy clustering in which data point can be a member of more than one cluster [6]. In hard clustering data point can be a member of more than one cluster. Each item in a fuzzy clustering has a set of membership coefficient corresponding to that value, item is belonging to particular cluster.

1.4.4. Model Based Clustering:

Model based clustering is an approach for cluster analysis. In model based clustering data are viewed as data is coming from coming different probability distributions [7]. This approach of clustering estimates the number of clusters and also finds excellent fit of models to data. The approaches used for model based clustering are**:**

- Expectation minimization
- Conceptual clustering
- Neural Network Approach

1.4.5. Density Based Clustering:

Density based spatial clustering of algorithms was coined by Ester in 1996.It is also termed as DBSCAN. Clusters of various shapes and sizes are obtained by using this technique.

1.5 Similarity of clusters

On the basis of closeness of data points data points are grouped into clusters. The items which have similar characteristics are grouped into one cluster and with different grouped into another [8]. Similarity of clusters can be measured in different ways on the basis of type of dataset. The most common way to find similarity of data points is by defining a distance metric among data points. The distances used for finding similarity of clusters are:

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

1.5.1 Euclidian Distance

Euclidian distance is a straight line distance between two points. The Euclidian separation is otherwise called L2 standard [8]. The Euclidian separation is the distance between two points in plane with coordinates(x, y). The equation for finding Euclidian distance is:

$$d[|y_1, y_2 \dots \dots \dots . . y_n|], \left[\left|z_{1,} z_2 \dots \dots \dots . z_n\right|\right] = \sqrt{\sum (y_i - z_i)^2}$$

Where d is the distance between the coordinates (x, y).

1.5.2 Manhattan Distance

Manhattan Distance is also termed as city block distance, absolute value distance or L_1 norm. It is measured along axes of right angles [8]. Manhattan distance is the sum of horizontal and vertical distance and diagonal distance is computed by using Pythagoras theorem. The formula for Manhattan Distance is:

$$d[|x_1, x_2 \dots \dots \dots . . x_n|], \left[\left|y_{1,} y_2 \dots \dots \dots . y_n\right|\right] = \sqrt{\sum |x_i - y_i|}$$

1.5.3 Edit Distance

Edit Distance is used when the dataset is accessible in the form strings. The Edit separate is the method for discovering how one string is like each other. The separation between two strings is number of activities required to change over one string to another [8]. It is utilized as a part of bioinformatics to discover the similitude between the DNA groupings. The Operation utilized as a part of Edit remove are:

- Insertion
- Deletion

1.5.4 Hamming Distance

Hamming separation is utilized for Boolean numbers [9]. Hamming separation is figured when the two strings are of equivalent length. The value of hamming distance can never be negative. It can be utilized as a part of coding hypothesis.

1.6 Methods to find ideal count of Clusters

1. Elbow Method
2. Average silhouette method
3. Gap statistical method
4. Davies Bouldin index method
5. Dunn Index methods

1.6.1 Elbow Method:

The Elbow method helps to interpret and validate consistency within the cluster and to find the appropriate number of clusters in a dataset. This method focuses on percentage of variance which is a function of the number of clusters [8]. The number of clusters should be chosen in such a way that adding another cluster doesn't give much better modeling of data.

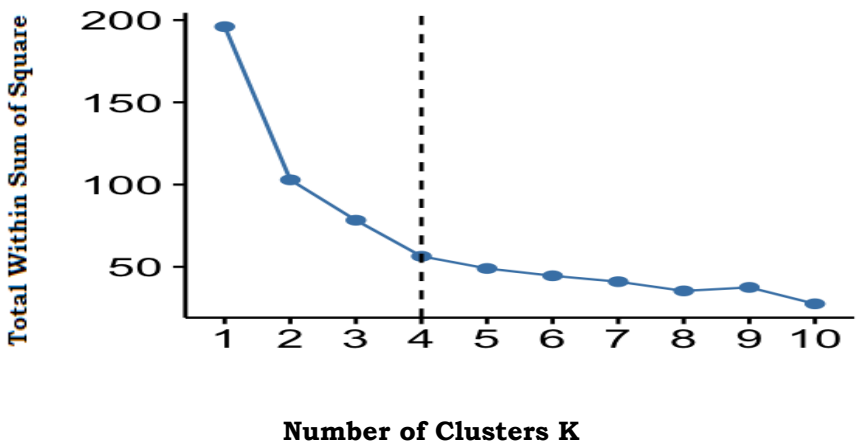

Figure 1.3 Evaluation graph of Elbow method [8]

1.6.2 Average silhouette method

It is used for finding quality of clusters. The range of silhouette method lies between the value of [-1, 1].Silhouette coefficient value near to +1 indicate that the cluster is well clustered [12] A Silhouette value is going to 0 shows that the data point is assigned to another cluster. The Silhouette value near - 1 shows cluster is misclassified.

The formula for finding silhouette index value is:

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

where T(k) is the silhouette index value and n(k) is the distance between data points in same cluster and x(k) is the distance between data points in different cluster and max{n(k),x(k)} is the maximum distance between data points in the neighboring clusters.

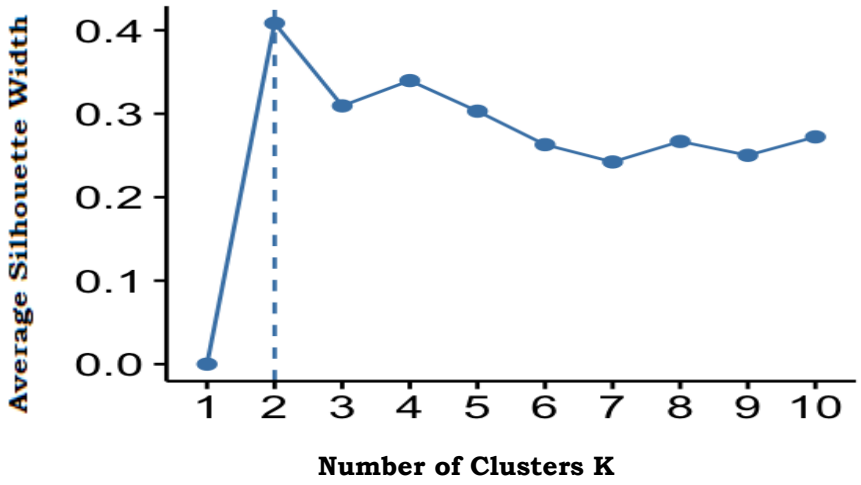

Figure 1.4 Evaluation Graph of Silhouette Method [8]

1.6.3. Gap Statistical method:

It compares the total within intra - cluster variation for distinctive values of k with their expected values under null reference distribution of the data. The value that maximize the gap statistic will form the optimal clusters.

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

Where W_{kb} is the expected value and W_k is observed value under null hypothesis.

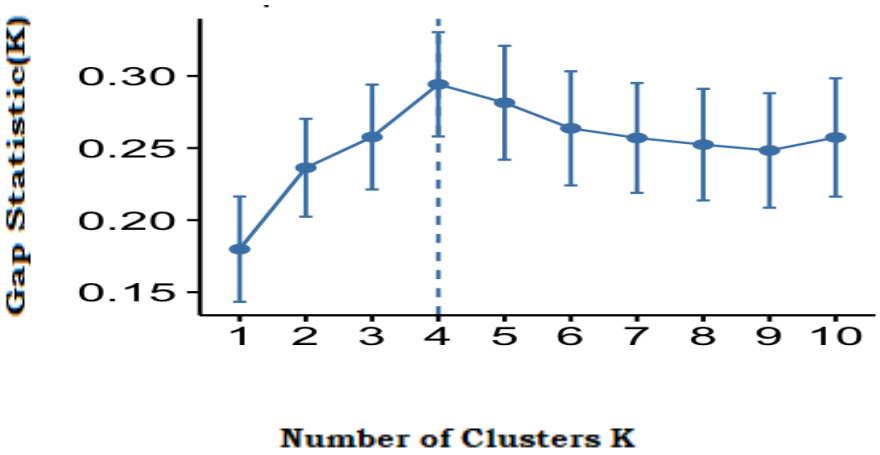

Figure 1.5 Evaluation Graph of Gap Statistic method [8]

1.6.4. Davies Bouldin Index:

It was coined by David and Donald in 1979. It is another technique for evaluation the clusters. Lower the value of Davies Bouldin index indicates good clustering [9]. The principle inconvenience of Davies Bouldin index is that good value reported by this technique doesn't suggest best information retrieval. The formula for finding Davies Bouldin index is

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

Where M_i the centroid of cluster is C_i, T_j is the size of cluster and D_j is the measure of validity of the cluster.

1.6.5. Dunn Index:

Dunn Index was introduced by J.C Dunn in 1979.It is a metric technique for evaluation of clusters. The Dunn index distinguish sets of clusters that will be that have similar properties and that datasets are assemble in one group [9]. Higher Dunn index value shows better clustering. The fundamental disadvantage of Dunn index is the computational cost as when the quantity of clusters and dimensionality of the information increment the calculation additionally increments.

1.7 Basic K-Means Algorithm

Input: K: number of clusters to be formed.

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

Output: Set of k clusters

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

Step 2: Calculate Centroid

- Calculate the centroid of data points using distance measures either by using Euclidean distance or hamming distance.
- Data points are assigned 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.

Formula used to calculate Euclidian distance:

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

Where d is the distance between coordinates(y, z)

1.7.1 Flowchart of K -means clustering.

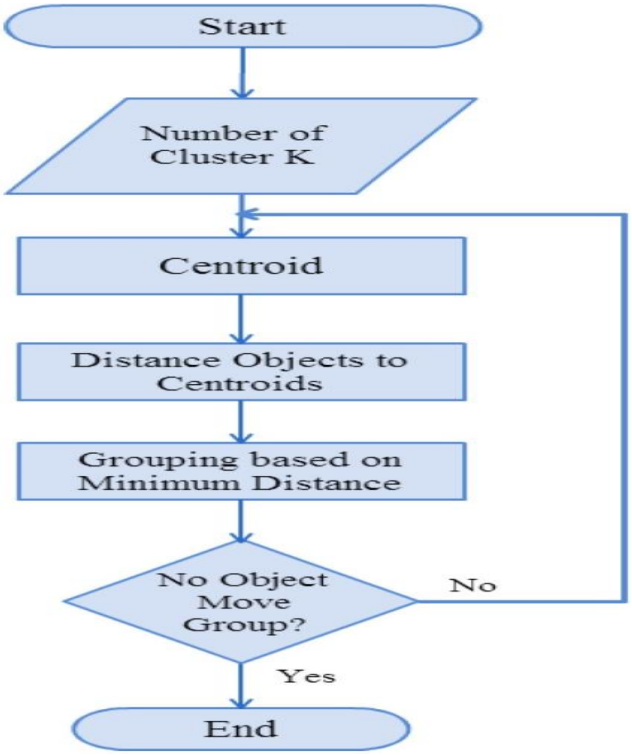

Fig 1.6 Flowchart of K-means Clustering [9]

Fig 1.6 represents the flowchart of K-means clustering. First we have to select the value of K or number of clusters to be formed. After deciding the value of k we have to calculate the distance

between all points and grouping of clusters based on minimum distance. The termination condition of the K-means clustering is when we get the same clusters in two or more iteration than that is the final clusters

1.7.2 Clusters of K-means on Iris dataset.

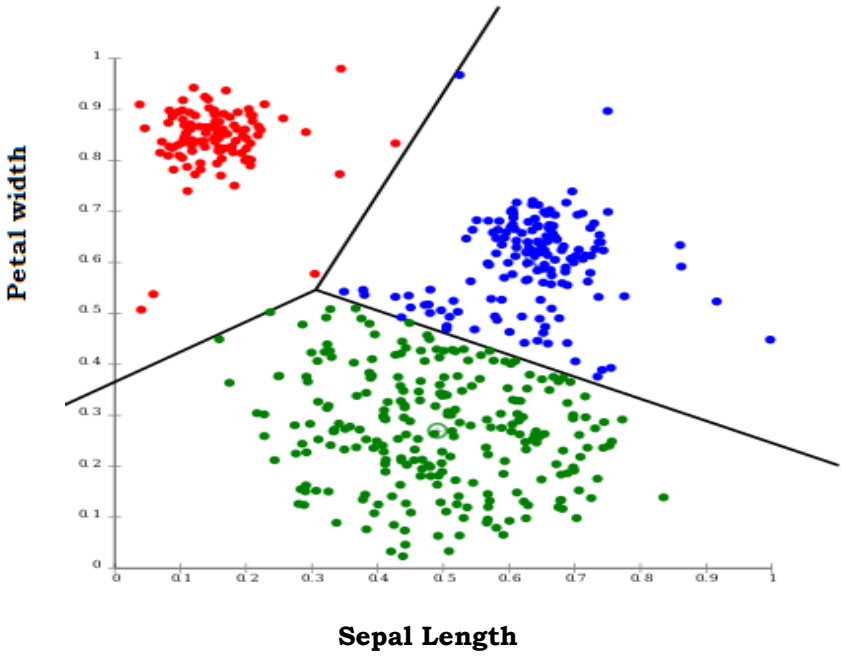

Figure1.7 Clustering on Iris Dataset [9]

Fig 1.7 shows the results of K-means clustering on Iris dataset. The value of K used is 3 and red color indicates the clusters of Iris setsoa, blue color indicates the clusters of iris versicolor and green color indicate the clusters of Iris virginica.

1.7.3Limitations:

The K-Means algorithm has the following limitations.

i) K means clustering involves prior knowledge of data and require the analyst to choose the appropriate number of k points in advance.

ii) The final result obtained is sensitive to initial random selection and is less sensitive to outliers.

iii) Rearranging the data results in different in different result every time you change the ordering of data.

iv) The clusters formed by this algorithm is fixed in shape that is only convex shape clusters are formed

1.7.4 Possible Solutions:

The Possible Solutions of K-means algorithm are listed below:

- Compute the K means for a range of K values or applying the Meta heuristic techniques (Genetic Algorithm, PSO) for initializing the first cluster centroid.
- The final result obtained is sensitive to initial random selection of cluster centroid. On every different run different initial centroid is selected that will give different clustering results every time.

1.7.5 Example of K-Means:

Dataset for K means {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}

1.8 Genetic Algorithm

It is a process that derives from biological evolution and natural selection to solve constrained and unconstrained optimization problem [10]. The genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the current population to be parents which are used to produce the children for the next generation. Genetic algorithm can be used to solve numerous optimization problems. Following are the main steps for genetic algorithm:

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.

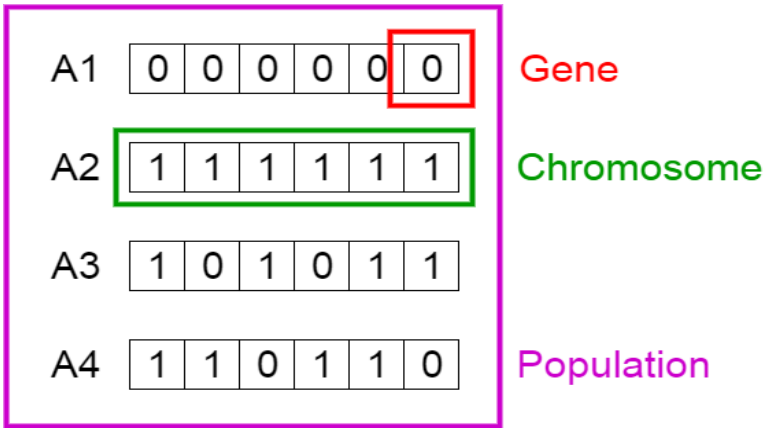

Figure 1.8 Genetic Algorithm Chromosome and Population [10]

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 individual will 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 fitness scores. 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 to produce 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.The crossover 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. The mutation 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**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**F=6**
S3	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**F=8**

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

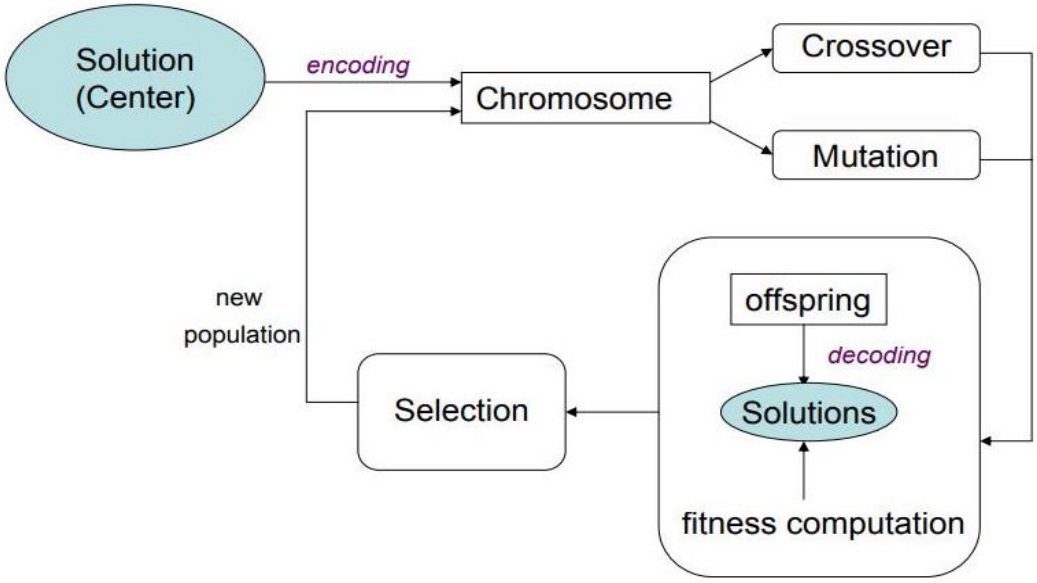

Figure 1.9 **Execution Steps of genetic algorithm [10]**

1.9 Organization of Thesis

The rest of the thesis is organized as follows:

Chapter 2- This chapter contains exhaustive description of literature survey done to study the concept of clustering of big data, recent clustering techniques for analyzing big data and comparison of these techniques.

Chapter 3- This chapter presents the problem statement along with the objectives of this research work.

Chapter 4- This chapter describes proposed algorithm to solve the stated problem.

Chapter 5- This chapter focuses on related concepts that are cardinal in our proposed method followed by implementation details and experimental results.

Chapter 6- In this chapter conclusion, followed by possible future research work is discussed.

Chapter 2 Literature Survey

2.1 Clustering

Clustering is an unsupervised technique that creates clusters of objects having similar features using 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 grouped in 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 density based clustering .This chapter presents literature survey of clustering techniques for analyzing big 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 downside of k means clustering [11]. The Clustering of dataset using K-means is shown in figure 2.1

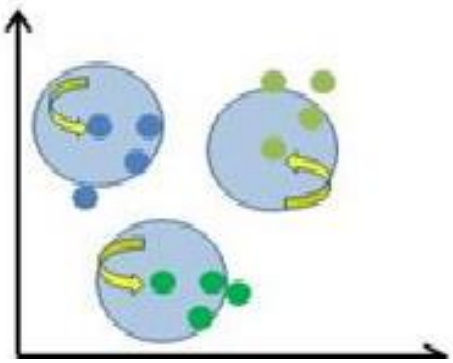

Fig 2.1 (a) Selction of seed

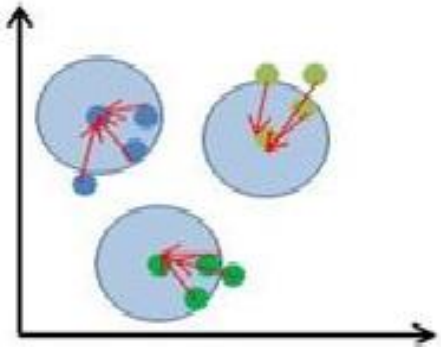

Fig 2.1(b) Assigment of Document

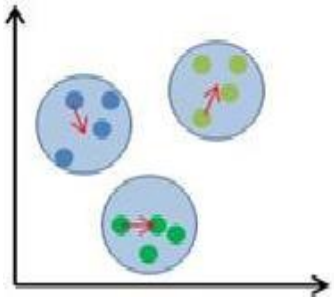

Fig 2.1(c) Recomputation of Centroids

Fig 2.1 Clustering of Scattered Documents

2.3 Data mining 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. Data Mining, also known as knowledge discovery in databases (KDD).The researchers analyze two elementary objectives of data mining: description and prediction. There are different techniques used for mining 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 but it 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 values of 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, decision tree, neural networks and machine learning [2].

P. vats and M.mandot [3] presented a comparative analysis of various Distributed Data Mining algorithms and its applications. They presented the related research of data mining techniques and discussed various data distribution scenarios. Then they reviewed the various data mining algorithms i.e. agglomerative for hierarchical and K-means and fuzzy c-means for partitioning clustering. Subsequently, they discussed various issues in data mining techniques.

S.Bandyopadhyay et al. [4] designed a genetic clustering algorithm called as GKA that classify the pixels of satellite image. This KGA clustering algorithm is applied when number of clusters is known a priori and crisp in nature. In this paper Genetic algorithm is used to search clusters centers. 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. In this 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 proposed algorithm 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 intra cluster 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 clustering used 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 the variables 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 technique of 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 provide better 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 value of K is not known Apriori.

D.X.Chang et al. [15] presented an algorithm for gene arrangement for k means clustering using genetic algorithm. In this paper Genetic algorithm with Gene arrangement (GAGR) is proposed.

Performance of GAGR is compared with k means algorithm. Proposed algorithm is applied for pattern recognition and image segmentation. GAGR clustering is applied to remote sensing image for clustering the pixels into classes but the evaluation of clusters is not done.

R.. Lletı et al. [16] presented a technique for selecting variables for k-means cluster analysis by using genetic algorithm that optimize the silhouettes value . Cluster silhouettes is used as fitness function for genetic algorithm. The use of cluster silhouettes to qualify the cluster or evaluating the cluster validity based on the silhouettes index value. Silhouette value measures the quality of cluster and selection of chromosomes is also based on silhouette value.

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 different clustering 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-means gives better results as compared to K-means algorithm.

I.B Saida and Omar [18] presented a new metaheuristic approach for clustering of data based on Cuckoo Search Optimization (CSO) to remove drawback of K-means. The basic functionality of cuckoo search is easy to implement and has good computational efficiency performance. The proposed algorithm using cuckoo search gives better results than k-means clustering .The experiment are carried on Iris sample dataset and breast cancer dataset.

A.Likas et al. [21] had given a genetic K-means clustering to find a globally optimal partition of data into specified number of clusters. It uses a Markov chain theory to provide a global optimum function. It is implemented using python.

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 compares to existing algorithms. It is implemented using Recommender system.

G. B. Phanendra et al. [23] introduced a genetic algorithm for initial seed selection of K-means using genetic algorithm. The proposed scheme is implemented using python and matlab and perform on different data sets. The proposed algorithm doesn't give empty clusters for large datasets and reduce the problem of global minimum. The proposed algorithm also give the highest accuracy as compared to existing algorithm techniques.

M. Celebi et al. [25] Introduced an efficient initialization methods for k means clustering. In this various initialization methods for initializing the value of k (forgy method, jancey method) is presented and perform on different datasets and compared the computational efficiency of all the algorithm. In all the algorithms K is initialized randomly and doesn't give optimal solution.

B. H. Park and H. Kargupta [31] presented a brief overview of the Distributed Data Mining algorithms, applications, systems and the emerging research areas. They first presented the related research of distributed data mining and illustrated various data distribution scenarios. Then they reviewed the various distributed data mining algorithms. Subsequently, they discussed various architectural issues in distributed data mining systems.

J. Han et al. [34] designed a spatial data mining system prototype that is called GeoMiner. The spatial data mining power of GeoMiner includes mining mainly three kinds of rules: comparison rules, characteristic rules and association rules, in geo-spatial datasets, with a designed extension to include mining clustering rules and classification rules. GeoMiner includes the spatial on-line analytical processing (OLAP) section, spatial data cube construction section, and spatial data mining section. A spatial data mining language, Geo-Mining Query Language, is designed by them and implemented as an extension to Spatial SQL, for spatial data mining. Moreover, they constructed a user- friendly, interactive data mining interface and implemented the tools for visualization of discovered spatial knowledge.

2.4 Conclusion

In this chapter literature review and comparative analysis of the recent techniques for clustering big data is done. The various clustering techniques for analyzing big data are compared and their merits and demerits are presented. The next chapter provides the gap between these algorithms and the problem statement along with the objectives of this thesis work.

Chapter 3 Research Problem

3.1Problem 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 factors in 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 outliers and 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 Genetic algorithm. In this research work a hybrid of Genetic and K means is applied in order to mitigate the 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 used to 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 thesis 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 clustering algorithm.
- To implement and validate the proposed technique on Iris and Breast cancer dataset.

3.4 Research Methodology

In our research work we will use python language, which is an excellent scripting language for manipulating text. The dataset required for analysis will be extracted from UCI machine learning repository. Iris dataset, Wine quality dataset and breast cancer dataset is used for analyzing proposed algorithm. We will use Weka data mining tool for visualization of clusters. 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 proposed algorithm 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.

Chapter 4 Proposed Hybrid Clustering Technique

The problem stated in previous chapter is solved by proposed Clustering technique (Genetic K-means). It takes advantages of both algorithms and analyze the data in an efficient way and create a cluster that are less sensitive to the outliers. The proposed technique generate clusters with maximum accuracy and low intra-cluster distance.

4.1 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 applying crossover and mutation the result is passed to K-Means for clustering .In last after clustering Davies Bouldin 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 formation of clusters.

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

Step 3: In the Third step Genetic algorithm for finding initial cluster centroid and after crossover and 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 of cluster.

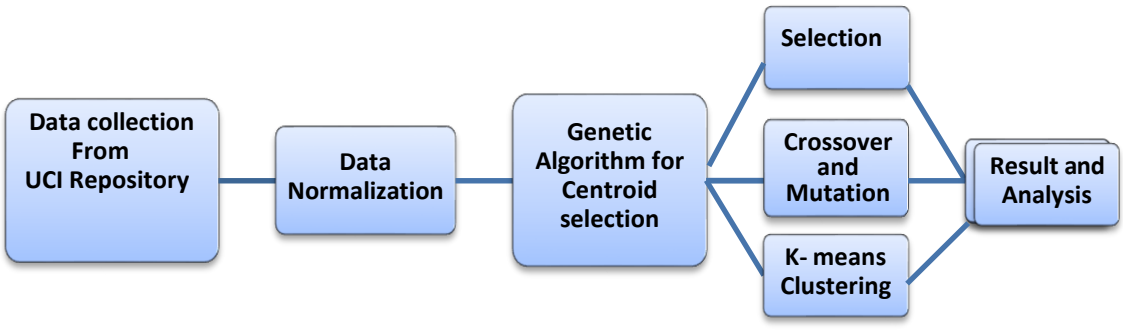

Fig 4.1 Implementation methodology for Clustering of dataset

4.2 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 mainly depend 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 fitness value 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

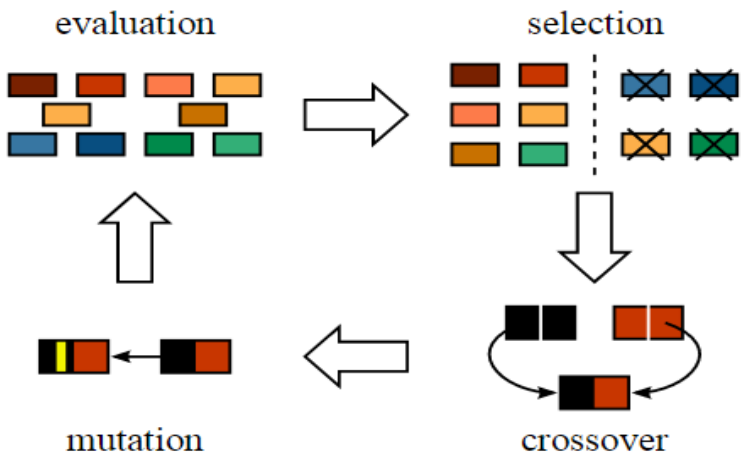

Figure 4.2 Process of Genetic Algorithm [21]

Fig 4.3 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.

4.2.1 Application of Genetic Algorithm

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

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

4.2.2 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 fitness value F *i.e.* no of one's in particular row

Table 4.1 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 4.1 shows the initialization of chromosomes of MaxOne problem and fitness value of all the chromosomes. The fitness value is calculated based on number of one's occur in particular iteration. 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 selection method 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 shows the pie chart representation of roulette wheel selection based on fitness value.

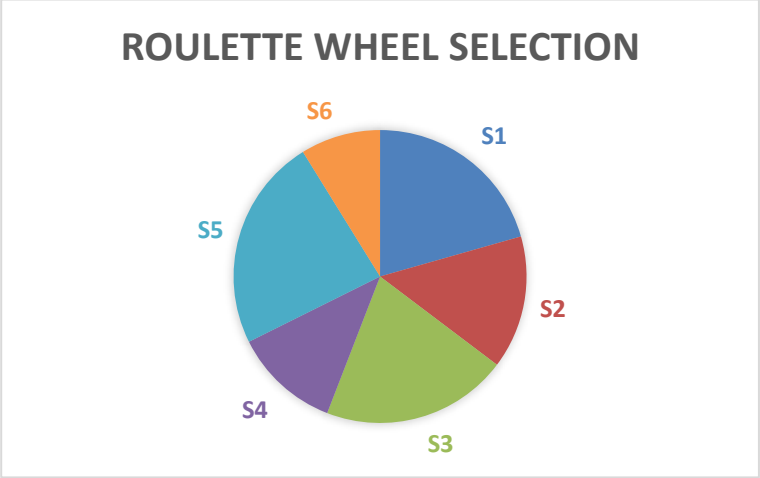

Figure 4.3 Roulette Wheel selection of Genetic Algorithm

Fig 4.3 shows the roulette wheel selection based on fitness value of chromosomes. Higher the value of particular chromosomes highest the probability to be selected first for crossover and mutation.

2.1 Arrange according to the fitness function:

Table 4.2 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 4.2 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 4.3 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 4.3 represent the chromosomes S1 and S3 for crossover. The two point crossover is performed on chromosomes S1 and S3 to increase the probability of number of one's.

Table 4.4 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 the number of one's increases.

2.3 Perform crossover on S2 and S4 chromosome.

Table 4.5 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 4.5 represent the chromosomes S2 and S4 for crossover. The two point crossover is applied to increase the probability of number of one's.

Table 4.6 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 4.6 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 4.7 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 4.7 represent the chromosomes S5 and S6 for crossover. The two point crossover is applied to increase the probability of number of one's.

Table 4.8 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 4.8 represent the crossover result of S5 and S6.After applying crossover on S5 and S6 the number of one's increases.

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

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

Table 4.9 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 4.9 shows the result of mutation of chromosomes of S1 to S6.After applying mutation on S1 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.

4.3 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. Genetic algorithm is used to select the optimum initial centroid for better results. Steps of Genetic K-means (GAKM) is as follows:

Fig 4.4 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.

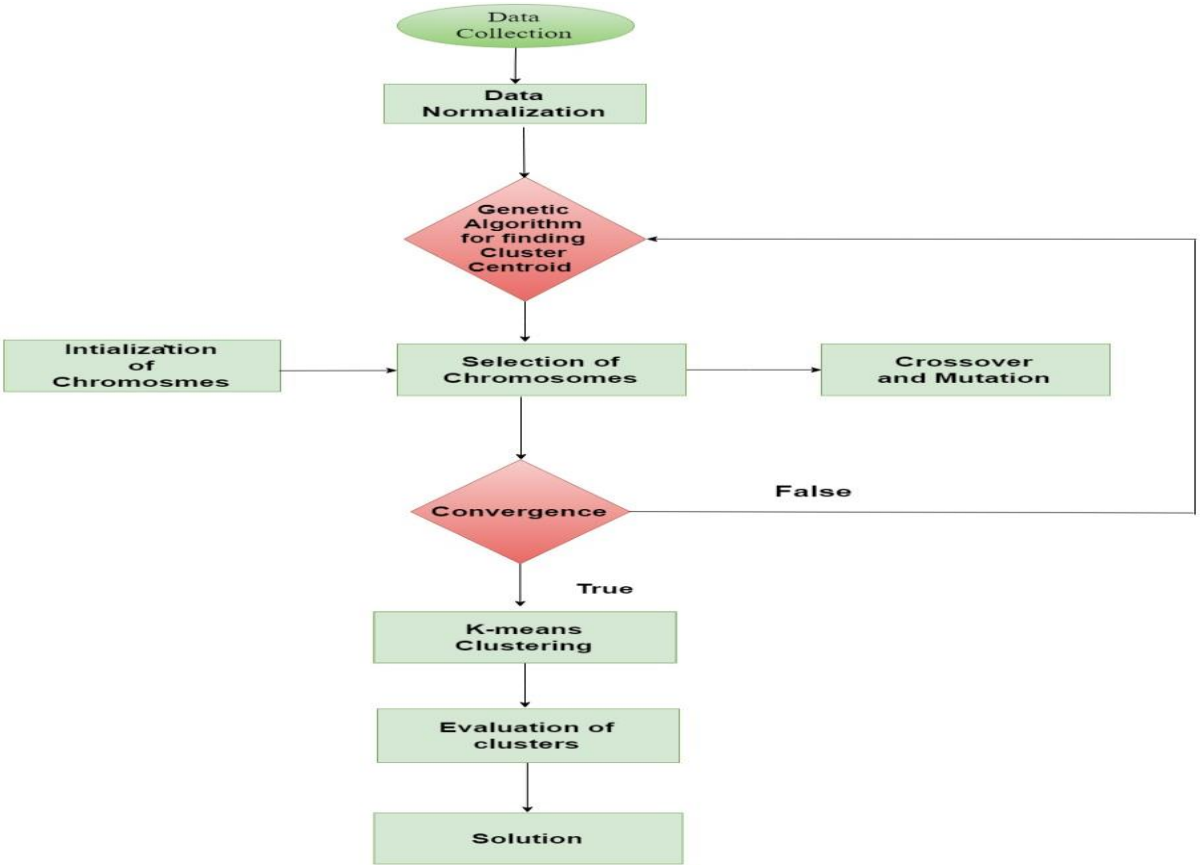

Figure 4.4 Flowchart of Proposed Algorithm

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 selected from 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 mv is 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-cluster distance.

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:

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

Where D^{q}_{INTRA} is the distance of the qth cluster, x_i and x_j are the points of the clusters and m is the no. elements in the cluster.

The total Intra cluster distance is:

$$S(\mathrm{D_{INTRA}})=\sum_{q=1}^{k}(\mathrm{D_{INTRA}})$$

Where $S(\mathrm{D_{INTRA}})$ is the sum of Intra-cluster distance. The intra-cluster distance is obtained by calculating the distance between data points in the same cluster.

Formula for finding Inter Cluster distance:

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

Where $D_{\mathrm{INTER}}^{q.r}$ is the distance between the neighboring clusters and x_i and x_j are the points of the clusters.m and n are the data points of the neighboring clusters.

The total inter cluster distance is:

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

Where $\mathrm{S}(D_{\mathrm{INTER}})$ is the total sum of inter-Cluster distance. The Intra-cluster distance is obtained by calculating the distance the distance between data points in same cluster.

Formula for finding Fitness value

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

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

v) Crossover Operator:

After the selection using rank based selection the next step is to produce off springs. The mainly used 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 diversity is 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:

$$\text{Offspring 1} = (\alpha * \text{parent 1}) + ((1- \alpha) * \text{parent2})$$

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

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 accuracy and it is applied to (P_m*pop size* u) where P_m 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 mutation for 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 Bouldin index for evaluation of clusters. Davies Bouldin index is used for evaluation of clusters whether the clusters formed are optimal or not. The lower value indicate that clustering is good.

Formula for Davies Bouldin:

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

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

4.3.1 Example of Proposed Algorithm (Genetic K-means)

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

Table 4.10 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 applied to iris dataset to find accuracy, precision, recall and sensitivity. Davies Bouldin index is used for evaluation of clusters. The first 15 rows of iris dataset is collected to perform genetic K-means clustering.

Table 4.10 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

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 accurate result and promotes easiness of data handling. Table 4.11 shows the normalized data to perform computation.

Table 4.11 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 4.11 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 corresponds 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 4.12: 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 4.12 shows the selected row indices and chromosomes of the dataset. In column No. 1 where, 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 4.13: 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 4.13 represents the calculated distance obtained from cluster 1, cluster 2 and cluster 3 and also the assignment of clusters.

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

Table 4.14: 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

$$\text{Offspring 1} = (\alpha * \text{parent 1}) + ((1 - \alpha) * \text{parent2})$$

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

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 new off 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 2 and 3.

Step 8. Mutation: Apply mutation on 5^{th} element of 1^{st} 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 5^{th} chromosome of parent 3.After apply mutation on 5^{th} chromosome the value of chromosome changed from 0.9726 to 0.7982

4.4 Conclusion

In this chapter the design approach of proposed algorithm was presented. The detailed explanation of proposed method was given along with its execution diagram and flow chart. In the next chapter the implementation of this proposed approach is conducted on different datasets and results of the experiment would be gathered.

Chapter 5 Implementation and Experimental results

This chapter of thesis work will focus on implementation of the proposed hybrid algorithm Genetic K -means (GAKM) using python and matlab.

5.1 Implementation of Proposed Technique

To implement the proposed approach four different datasets are used. These datasets are collected from different repositories. The dataset is collected from UCI Machine Learning Repository [42] .The dataset used are iris dataset, Wine quality dataset, Cancer dataset and sales dataset. All these datasets have different dimensionality. The proposed approach is implemented by Python using Python IDLE and matlab.

5.1.1 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 5.1 and figure 5.2.

I. Code of K-means Clustering:

kmeans-test.py - C:\Users\vandseti\Desktop\vandsna\Genetic-Algorithm-on-K-Means-Clustering-master\kmeans-test.py (3.6.5)

File Edit Format Run Options Window Help

```
import matplotlib.pyplot as plt
from matplotlib import style
style.use('ggplot')
import numpy as np
from sklearn.cluster import KMeans
from sklearn import preprocessing
from scipy.spatial import distance
import pandas as pd

# correct = 0
# for i in range(len(x)):
#     predict_me = np.array(x[i].astype(float))
#     predict_me = predict_me.reshape(-1, len(predict_me))
#     prediction = clf.predict(predict_me)
#     if prediction[0] == y[i]:
#         correct+=1


def main():
    data = pd.read_csv('data/iris.csv', header=None)
    clf = KMeans(n_clusters=3)
    clf.fit_predict(data)
    print(clf.cluster_centers_)
    print(clf.labels_)

    centroids = clf.cluster_centers_
    # 10 clusters
    labels = clf.labels_
    correct_answer = 0
    for i in range(0,50):
        if labels[i] == 0:
            correct_answer+=1
    for i in range(50,100):
        if labels[i] ==1:
            correct_answer+=1
    for i in range(100,150):
        if labels[i] ==2:
            correct_answer+=1

    accuracy = (correct_answer/150)*100
    print(accuracy)
```

Figure 5.1 Code of K-means Clustering on Iris Dataset.

Fig 5.1 Shows the code of K-means clustering on Iris datset.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 %.

II. Result of K-means Clustering.

Figure 5.2 Result of K-means on Iris dataset

Fig 5.2 shows the result of K-means clustering on Iris dataset. Comparison is made between the actual and predicted cluster using K-means. The number if iterations performed is 200 and the accuracy obtained from K-means is 45%.

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

main.py - C:\Users\vandseti\Desktop\vandsna\Genetic-Algorithm-on-K-Means-Clustering-master_main_.py (3.6.5)

File Edit Format Run Options Window Help

```
import configparser
import numpy as np
import pandas as pd

from cluster import Clustering
from genetic import Genetic
from generation import Generation

NORMALIZATION = True


def readVars(config_file):
    config = configparser.ConfigParser()
    config.read(config_file)
    budget = int(config.get("vars", "budget"))
    kmax = int(config.get("vars", "kmax"))  # Maximum number of Clusters
    numOfInd = int(config.get("vars", "numOfInd"))  # number of individual
    Ps = float(config.get("vars", "Ps"))
    Pm = float(config.get("vars", "Pm"))
    Pc = float(config.get("vars", "Pc"))

    return budget, kmax, Ps, Pm, Pc, numOfInd


# minmax normalization
def minmax(data):
    normData = data
    data = data.astype(float)
    normData = normData.astype(float)
    for i in range(0, data.shape[1]):
        tmp = data.iloc[:, i]
        # max of each column
        maxElement = np.amax(tmp)
        # min of each column
        minElement = np.amin(tmp)

        # norm_dat.shape[0] : size of row
        for j in range(0, normData.shape[0]):
            normData[i][j] = float(
                data[i][j] - minElement) / (maxElement - minElement)
```

Figure 5.3 Genetic K-means clustering on Iris dataset

Fig 5.3 Shows the code of Genetic K-means clustering on Iris datset.It shows implementation of Genetic K-means clustering and uses the iris data and value of K used is 3. This code predicts the accuracy of Genetic 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 85 %.

IV. Result of Genetic K-means on Iris dataset:

Figure 5.4 Result of Genetic K-means on Iris dataset.

Fig 5.4 shows the result of Genetic K-means clustering on Iris dataset. Comparison is made between the actual and predicted cluster using Genetic K-means. The number if iterations performed is 200 and the accuracy obtained from Genetic K-means is 45%.

5.1.2 Wine Quality Dataset: The Wine quality dataset is collected form UCI machine learning repository dataset. This dataset contains 11 attributes and 1599 rows and last attribute contains the quality of wine (0 indicate low quality and 1 indicate good quality). Confusion matrix, Accuracy, Recall and Precision are calculated. Davies Bouldin index is used for evaluation of each cluster. Following figure shows the code of K-means and Genetic K-means using Wine Quality dataset.

I. Code of K-means on Wine Dataset:

kmeans-test.py - C:\Users\vandseti\Desktop\vandsna\Genetic-Algorithm-on-K-Means-Clustering-master\kmeans-test.py (3.6.5)

File Edit Format Run Options Window Help

```
import matplotlib.pyplot as plt
from matplotlib import style
style.use('ggplot')
import numpy as np
from sklearn.cluster import KMeans
from sklearn import preprocessing
from scipy.spatial import distance
import pandas as pd

# correct = 0
# for i in range(len(x)):
#     predict_me = np.array(x[i].astype(float))
#     predict_me = predict_me.reshape(-1, len(predict_me))
#     prediction = clf.predict(predict_me)
#     if prediction[0] == y[i]:
#         correct+=1


def main():
    data = pd.read_csv('data/wine.csv', header=None)
    clf = KMeans(n_clusters=3)
    clf.fit_predict(data)
    print(clf.cluster_centers_)
    print(clf.labels_)

    centroids = clf.cluster_centers_
    # 10 clusters
    labels = clf.labels_
    correct_answer = 0
    for i in range(0,50):
        if labels[i] == 0:
            correct_answer+=1
    for i in range(50,100):
        if labels[i] ==1:
            correct_answer+=1
    for i in range(100,150):
        if labels[i] ==2:
            correct_answer+=1

    accuracy = (correct_answer/150)*100
    print(accuracy)
```

Figure 5.5 K means algorithm on wine dataset

Fig 5.5 Shows the code of K-means clustering on Wine datset.It shows implementation of K-means clustering and uses the Wine data and value of K used is 2. 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 50 %.

II. Result of K-means clustering on Wine quality dataset

Figure 5.6 Result of K-means algorithm on wine quality dataset

Fig 5.6 shows the result of Genetic K-means clustering on Wine dataset. Comparison is made between the actual and predicted cluster using Genetic K-means. The number if iterations performed is 200 and the accuracy obtained from Genetic K-means is 50%.

III. Genetic K-means on wine quality dataset

main.py - C:\Users\vandseti\Desktop\vandsna\Genetic-Algorithm-on-K-Means-Clusterin...

File Edit Format Run Options Window Help

```
import pandas as pd

from cluster import Clustering
from genetic import Genetic
from generation import Generation

NORMALIZATION = True


def readVars(config_file):
    config = configparser.ConfigParser()
    config.read(config_file)
    budget = int(config.get("vars", "budget"))
    kmax = int(config.get("vars", "kmax"))  # Maximum number of Clusters
    numOfInd = int(config.get("vars", "numOfInd"))  # number of individual
    Ps = float(config.get("vars", "Ps"))
    Pm = float(config.get("vars", "Pm"))
    Pc = float(config.get("vars", "Pc"))

    return budget, kmax, Ps, Pm, Pc, numOfInd


# minmax normalization
def minmax(data):
    normData = data
    data = data.astype(float)
    normData = normData.astype(float)
    for i in range(0, data.shape[1]):
        tmp = data.iloc[:, i]
        # max of each column
        maxElement = np.amax(tmp)
        # min of each column
        minElement = np.amin(tmp)

        # norm_dat.shape[0] : size of row
        for j in range(0, normData.shape[0]):
            normData[i][j] = float(
                data[i][j] - minElement) / (maxElement - minElement)

    normData.to_csv('result/norm_data.csv', index=None, header=None)
```

Figure 5.7 Genetic K-means on wine quality dataset.

Fig 5.7 Shows the code of Genetic K-means clustering on Wine quality datset.It shows implementation of Genetic K-means clustering and uses the wine quality and value of K used is 2. This code predicts the accuracy of Genetic 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 88 %.

IV.Result of Genetic K-means on Wine quality dataset

Figure 5.8 Result of Genetic K-means on wine quality dataset

Fig 5.8 shows the result of Genetic K-means clustering on wine quality dataset. Comparison is made between the actual and predicted cluster using Genetic K-means. The number if iterations performed is 200 and the accuracy obtained from Genetic K-means is 85%.

5.2Experimental 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.

5.2.1. Confusion Matrix of K-means Clustering

All the two algorithms that is K-Means and Genetic K-means algorithm are tested on the iris and wine quality datasets to calculate accuracy. Table 5.6 shows the comparison of the confusion matrix obtained by each algorithm on two different datasets that includes Iris and Wine quality dataset.

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

Fig 5.9:Confusion Matrix obtained from K means Algorithm

Fig5.9 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%**
- **Missclassification rate or Error rate=FP+FN/Total =0+0+0+34+0+36/150= 0.46 or 46%**
- **Precision:** When its predicts cluster 1,how often it is correct.**TP/Predicted 1=10/46=0.21 or 21%**
- **Recall:**When its actual 1 ,how often its predicts correct. **TP/Actual 1=10/44=0.22 or 22%**
- **Sensitivity:**Sensitivity is also known as Recall =**TP/Actual 1=10/44=0.22 or 22%**
- **Specificity:**Specificity is also known as precision =**TP/Predicted 1=10/46=0.21 or 21%**

5.2.2 Confusion Matrix of Genetic K-means algorithm.

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

Fig 5.10:Confusion Matrix obtained from Genetic K-means Algorithm.

Fig5.10 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= 0.10 or 10%**
- **Precision:** When its predicts cluster 1,how often it is correct.**TP/Predicted 1=48/62=0.77 or 77%**
- **Recall:** When its actual 1 ,how often its predicts correct. **TP/Actual 1=48/50=0.96 or 96%**
- **Sensitivity:** Sensitivity is also known as Recall =**TP/Actual 1=48/50=0.96 or 96%**
- **Specificity:** Specificity is also known as Precision.**TP/Predicted 1=48/62=0.77 or 77%**

5.2.3. Test for Performance of Accuracy

All the two algorithms that is K-Means algorithm and proposed approach are tested on four datasets that includes iris dataset, wine dataset, Cancer Dataset and Sales 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 the total 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 gives better 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 5.1

Table 5.1: Accuracy obtained from K means and Proposed Algorithm

Dataset	K-means	Proposed Algorithm
Iris Dataset	**66%**	**83%**
Wine Dataset	**53%**	**80%**
Cancer Dataset	**50%**	**77%**
Sales Dataset	**63%**	**75%**

Table 5.1 Shows the table of accuracy obtained from K-means and Genetic K-means Algorithm. Accuracy is computed on four datasets (Iris dataset, Wine dataset, Cancer dataset, Sales dataset) using K-means and Genetic Algorithm. The table shows that Genetic algorithm gives higher accuracy than K-means algorithm.

5.2.4. 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-means and proposed algorithm is to minimize the intra cluster distance. The formula for finding intra cluster distance 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}_{INTRA} is the distance between data points in same cluster.

The Total sum of Intra cluster distance is:

$$S(\mathrm{D_{INTRA}})=\sum_{q=1}^{k}(\mathrm{D_{INTRA}})$$

Where $S(\mathrm{D_{INTRA}})$ is the sum of intra-cluster distance between all data points. The total intra cluster distance is obtained by calculating the distance between data points in the similar cluster.

The Intra cluster distance obtained from k means and proposed algorithm is shown in table 5.2 and table 5.3.

Table 5.2: 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 5.2 shows the table of intra-cluster distance using K-means algorithm.The intra-cluster distance is 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 5.3 :Intra Cluster distance using proposed algorithm

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

Table 5.3 shows the table of intra-cluster distance using proposed algorithm.The intra-cluster distance is obtained by calculating the distance between data points in the same cluster.The distance obtained by proposed algorithm is less than K-means algorithm.

5.2.5. 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 objective of K-means and proposed algorithm is to maximize the inter cluster distance. The inter cluster distance obtained 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 in clusters and $D_{\text{INTER}}^{q.r}$ is the inter cluster distance between data points.

The Total Inter Cluster distance is :

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

Where $S(D_{\text{INTER}})$ is the total sum of inter-cluster distance .The total sum is computed by calculating the inter-cluster distnace between all data points from one cluster to another.

Table 5.4: 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 5.4 shows the table of Inter-Cluster distance using K-means Algorithm. The table shows the value of distance between data points of one cluster to another cluster. The calculated value is lesser than proposed algorithm.

Table 5.5: 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 5.5 shows the table of Inter-Cluster distance using Genetic K-means Algorithm. The table shows the value of distance between data points of one cluster to another cluster. The calculated value is higher than proposed algorithm.

5.3 Conclusion

This chapter gives the details of implementation with the experimental results. The results obtained shows that proposed algorithm is better than K-Means .In the next chapter the conclusion and future directions of this work is presented with thesis contribution.

Chapter 6 Conclusion and Future Scope

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

6.1 Conclusion

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

6.2 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 value we 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.

6.3 Thesis Contribution

In this thesis different current clustering techniques for analyzing big data have been analyzed and are compared according to some parameters.

- A hybrid approach based on Genetic K-Means has been designed.
- The proposed design is implemented and deployed Python framework by using Python IDE.
- Experimental results demonstrate that the proposed technique is more precise than K-Means clustering algorithm.
- Proposed technique has higher accuracy then k means and also intra cluster distance is less and inter cluster distance is more.

6.4 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 issues involved 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] "Knowledge Discovery in Database" [Online].Available: https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/KDD3.htm.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] "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]

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

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

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

[12] K.Dharmendra and K. Sharma. "Genetic k-Means clustering algorithm for mixed numeric and categorical data sets." *International Journal of Artificial Intelligence & Applications* 1.2 (2010): 2328.

[13] 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.

[14] 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).

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

[16] 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.

[17] 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), 2014 International Conference on*. IEEE, 2014.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

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

[25] 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.

[26] A.K. Jain, M.N. Murty, and P.J. Flynn. "Data clustering: a review." *ACM computing surveys (CSUR)* 31.3 (1999): 264-323.

[27] Xiao, Jing, et al. "A quantum-inspired genetic algorithm for k-means clustering." *Expert Systems with Applications* 37.7 (2010): 4966-4973.

[28] S.S. Khan, and A. Ahmad. "Cluster center initialization algorithm for K-means clustering." *Pattern recognition letters* 25.11 (2004): 1293-1302.

[29] L. Bottou, Y. Bengio, "Convergence Properties of the k-Means Algorithms", *Advances in Neural Information Processing Systems 7*, pp. 585-592, 1995.

[30] K. Krishna, M. N. Murty, "Genetic k-Means Algorithm", *IEEE Trans. Systems Man and Cybernetics*, vol. 29, no. 3, pp. 433-439, June 1999.

[31] B. H. Park and H. Kargupta, "*Distributed Data Mining: Algorithms, Systems, and Applications*", In Proceedings of International Conference on Distributed Data Mining, pp. 341-358, 2002.

[32] S.Z. Selim, and M.A. Ismail. "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality." *IEEE Transactions on pattern analysis and machine intelligence* 1 (1984): 81-87.

[33] K. R. Vazquez., C. M Fonseca, P. J. Fleming, "Identifying the Structure of Nonlinear Dynamic Systems Using Multiobjective Genetic Programming" in IEEE Transactions on Systems Man and Cybernetics, Part A-Systems and Humans, vol. 34, 2004, pp. 531-545.

[34] J. Han, K. Koperski and N. Stefanovic, "*GeoMiner: a system prototype for Spatial Data Mining*", In Proceedings of the ACM SIGMOD International Conference on Management of data, pp. 553-556, 1997.

[35] F. Han and C.Yang "A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information." *IEEE/ACM transactions on computational biology and bioinformatics* 14.1 (2017): 85-96.

[36] N. Trivedi and S. Kanungo. "Performance enhancement of K-means clustering algorithm for gene expression data using entropy-based centroid selection." *Computing, Communication and Automation (ICCCA), 2017 International Conference on*. IEEE, 2017.

[37] R.M. Suresh and K. Dinakaran, and P. Valarmathie. "Model based modified k-means clustering for microarray data." *Information Management and Engineering, 2009. ICIME'09. International Conference on*. IEEE, 2009.

[38] M. Yang, J. Song, G. I. Ji, "FCM_FS: A Simultaneous Clustering and Feature Selection Model for Classification", *Computer Science and Information Engineering 2009 WRI World Congress*, pp. 250-255, 2009.

[39] S.Dutta and S. Ghatak “A graph based clustering technique for tweet summarization." *Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015 4th International Conference on*. IEEE, 2015.

[40] “Evolution of Big Data and Significant Growth of Data” [online]. Available: http://www.atkearney.com/strategic-it/article/-/asset_publisher/content/big-data- business-models.

[41] “Evolution of Big Data” [online]. Available: http://www.forbes.com/sites/gilpress/a- very-short-history-of-big-data. [3 June 2014].

[42] “UCI Machine Repository” [online]. Available: https://archive.ics.uci.edu/ml/dataset. [13 March 2014].

ME THESIS(VANDANA SETIA)

ORIGINALITY REPORT

10%	4%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.codeproject.com Internet Source	<1%
2	www.cc.gatech.edu Internet Source	<1%
3	Min Keng Tan, Helen Sin Ee Chuo, Heng Jin Tham, Kenneth Tze Kin Teo. "Exothermic Batch Process Optimisation via Multivariable Genetic Algorithm", 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT), 2012 Publication	<1%
4	nzcsrsc08.canterbury.ac.nz Internet Source	<1%
5	Submitted to University of Pretoria Student Paper	<1%
6	Submitted to London School of Economics and Political Science Student Paper	<1%
7	Wellington Simbarashe Manjoro, Mradul	

	Dhakar, Brijesh Kumar Chaurasia. "Operational analysis of k-medoids and k-means algorithms on noisy data", 2016 International Conference on Communication and Signal Processing (ICCSP), 2016 Publication	<1%
8	"Collaborative Networks: Reference Modeling", Springer Nature America, Inc, 2008 Publication	<1%
9	Submitted to University of Bridgeport Student Paper	<1%
10	Ravindra R. Rathod, Rahul Dev Garg. "Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data", International Journal of Energy Sector Management, 2017 Publication	<1%
11	M. Indulska, M. E. Orlowska. "Gravity based spatial clustering", Proceedings of the tenth ACM international symposium on Advances in geographic information systems - GIS '02, 2002 Publication	<1%
12	Submitted to National Economics University Student Paper	<1%
13	Submitted to Maulana Azad National Institute of Technology Bhopal	<1%