DESIGN OF AN EFFICIENT FRAMEWORK TO ENHANCE THE CLUSTERING PERFORMANCE IN DATA MINING

A Thesis Submitted
In Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY
In
COMPUTER SCIENCE & ENGINEERING

By

Muhammad Kalamuddin Ahamad
(Enrolment No.: MUIT0117038001)

Under the Supervision of
Dr. Ajay Kumar Bharti
MUIT

To the
Faculty School of Engineering
MAHARISHI UNIVERSITY OF INFORMATION TECHNOLOGY
LUCKNOW, U.P.

August, 2021

DECLARATION

I hereby declare that the work presented in this report entitled "**Design of an Efficient Framework to Enhance the Clustering Performance in Data Mining** ", was carried out by me. I have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and manipulation of the experiments and results, I shall be fully responsible and answerable.

Name: Muhammad. Kalamuddin Ahamad

Enroll. No. : MUIT0117038001

Field: Computer Science & Engineering

(Signature of the Research Scholar)

Date:

CERTIFICATE

Certified that **Muhammad Kalamuddin Ahamad** (Enrollment No.:MUIT0117038001) has carried out the research work presented in this thesis entitled **"Design of an Efficient Framework to Enhance the Clustering Performance in Data Mining"** for the award of Doctor of Philosophy from Maharishi University of Information Technology (MUIT), Lucknow under my/our **Dr. Ajay Kumar Bharti** supervision. The thesis embodies results of original work, and studies are carried out by the student himself/herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature:

(Dr. Ajay Kumar Bharti)
Professor
Maharishi University of Information Technology (MUIT), Lucknow
Date:

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor Prof. (Dr.) Ajay Kumar Bharti for their constant support, guidance, and motivation. It would never have been possible for me to take this work to completion without their incredible support and encouragement. I benefited greatly from many fruitful discussions with Professor A. A. Zilli (Retd. Prof. Govt. Autonomous Engineering College, Kamla Nehru Institute of Technology Sultanpur U.P.). I cannot forget the valuable help and motivation of Honorable Chancellor S. W. Akhtar Integral University Lucknow.

I would like to thank the faculty member of Integral University Lucknow especially Dr. Manish Madhav Tripathi, Dr. Mohd. Haroon, Anwar sheikh for discussed the concepts on clustering, and Dr. Mohd. Muqeem for statistical analysis (SPSS tool) and neural network tool (NNT). I would like to thank Mr. Tabrez Khan his help in writing the thesis. Sincere thanks to Professor Dr. Santos MUIT Lucknow for their valuable support. I want to say ‘thank you” to Mr. B.P. Chaurasia, Mr. R. K. Singh, Mr. Sameer Srivastava and Prof. (Dr.) N. Badal.

.

I cannot forget the valuable conversation and suggestion of Dr. Sohel Ahemad Khan assistant professor at Indira Gandhi Tribal University MP , India. I have often looked towards his valuable suggestion, and he always helped me whenever needed support in my research. I would like to thank Dr. Manish Madhav Tripathi for their valuable helps in MATLAB software.

I am truly grateful to my parents for their immeasurable love and care. They have always encouraged me to explore my potential and pursue my dreams. They helped me a lot to reach this stage in my life. I would like to thank my wife Raheba Parween for supporting and keeping me motivated throughout this work. Sweet thank is also for my Daughter Bushra, Alisha, Mariyam their puerile support. I would also like to thank my brother and sisters for their everlasting encouragement and supports.

At last I wish to thank many other people whose names are not mentioned here but this does not mean that I have forgotten their help.

(Muhammad Kalamuddin Ahamad)

ABSTRACT

Data mining method is generally used for determining more important information in an enormous dataset. The mining of data is a procedure of being acquainted with consistent patterns in a huge dimension of data applied to the techniques of unsupervised clustering, statistics, genetics, and radial basis function. Data mining concepts extract good information obtained from the dataset where those particular datasets are created well in clusters shape with convergence. The clustering techniques can be categorized namely as partitioning clustering, hierarchical clustering, density-based clustering, and grid-based clustering. It has more utilities to carry the data mining as an essential part of the business.

We have proposed an efficient framework of clustering approach, and its method that improves clustering metrics, analyze the clustering of k-means with other approaches using the software tools, propose and analysis the fitness objective function using Genetic Algorithm (GA), and analysis the clustering metrics SSE using Radial Basis Function of Neural Network of ANN.

An efficient framework is being presented for producing the good quality of clusters. Evaluate metrics related to performance with the component of clusters. Here, we have discussed the every component of the framework. The component of the framework consists are first components proposed methodology, and proposed algorithm is hybridized via PCA and PSO; second component is a statistical analysis with software tools; Third component utilizes the Genetic Algorithm(GA), and fourth component RBFN of ANN theory, using datasets.

There is the first component discussed to the proposed algorithm of clustering and it is implemented on various sizes of the datasets. We have implemented an experiment on MATLAB R2013a to measure the metrics of the cluster and also measure the fitness of fitness function values using particle swarm optimization has been accomplished through the critical literature survey, collects the ideas of experts, require to improvement, and validates the results. Measure the cluster performance using proposed

techniques applies on large datasets. The results have found for a cluster at K by using proposed method, and its proposed method hybridized via PCA is good as compared than proposed. Furthermore, PSO hybridized via proposed method is found that the higher fitness value as compared than earlier method. The results are evaluated which are based on internal and external metrics being used of four datasets. It is observed that the proposed method has better performance for clustering on four datasets in the terms of metrics.

The second component of study is to implement the proposed framework to create good quality of clusters by the statistical analysis by software tools SPSS 17.0 and NCSS2021 on large datasets. PCA techniques utilize the numerical attributes on the database the noisy feature reduces the n^{th}dimensional of the problem considering as the dataset but improves the cluster quality on the basis of the distance between initial centroids. In thesis, shown that the results with reduced no. of components, and also illustrated the comparative analysis of initial centroids value at k cluster levels of proposed concept is more significant than the existing methodology. Further, we have analyzed the F-RATIO or F-SCORE of the original k-means, and it is hybridized via PCA. In the fuzzy approaches to apply and find the silhouette, the compactness of the created clusters all results have revealed the significant and good quality of clusters.

The third component of study is to implement the proposed framework being applied to the genetic algorithm for simulation results of the KFDAFF algorithm. This performance is more favorable in the ordering of datasets. It is mentioned that fitness value of an objective function like in terms of best fit and means, stopping criteria, and the average distance between individuals of the simulation process. The comparative analysis criteria of the objective function proposed concept is lesser than an objective function of K-Means, and the exit criteria are the selections when the no. of generation produced touches the maximum (of population) value.

The fourth component of study is to implement the proposed framework being used by the RBFNN kernel concept and achieve a better sum of square training and testing results

based on the literature review, collect the ideas of experts, requires to improve, and validates the results. RBFNN is touching the appropriate metrics of sum of squared errors outcomes for extra assistance with the good quality of generated clusters.

The proposed guidelines are more helpful to handle the produced good quality of clusters. The validated results of comparative analysis of different components are to reflect the helpfulness and more appropriateness of projected framework model or components and its guidelines. Hence, suggested concepts and it's significant at k clusters are well recognized for development of good quality of clusters.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS & ABBREVIATION

K	Number of clusters
N	Total number of instances in datasets
n	Number of dimension
ω	value of Inertia
μ	Mean of two dimensional dataset
ʎ	Eigenvalues
V	Eigenvectors
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
C_{ij}	Covariance matrix of dataset
SSE	Sum of Squared Error
X_i	Data point at i^{th} position
A^T	Transpose Matrix
GA	Genetic Algorithm
ϕ	Function of feature space
σ	Variance
A	Define the finite set
$\varphi(r)$	Gaussian Radial Function
$K(x_i, x_j)$	Kernel
RBFNN	Radial basis function of Neural Network

KFDAFF	Kernel fisher’s discriminant analysis fitness Function

CHAPTER 1

INTRODUCTION

The mining method is a generally used for determining more important information in an enormous dataset. The mining of data is a procedure of being acquainted with consistent patterns in a huge dimensions of data applied to the techniques of unsupervised clustering, statistics, genetics, radial basis function. Data mining concepts extract good information obtained from the dataset where those particular datasets are created well clusters shape with convergence.

1.1 Introduction

Clustering data and extracting patterns within the clusters is the stepladder used in the well-defined methods of data analysis in the Information Technology era. Clustering is a method that groups an assortment of physical objects (or conceptual objects) placed into groups of similar objects. The cluster of data objects can be treated together as an assembly for supplementary processing, alleviating the management of the huge quantity of diverse data. Another ways like pattern removal, especially class narrative, and generating metaphors for characterizing the information are some applications of clustering.

The characteristic of cluster concept is information of a group into a set that the resemblances of similarity of intra cluster are minimized, while the intra-cluster is maximized. This method is used in different areas like mining of information, pattern detection, client dissection, a study for similarities, examine of tendency.

The clustering of data point has immense applications in every field of life, either consciously or unconsciously. The usage of data clustering has broad applications in field of the information technology, especially retrieval of information. Some of other

applications include: extending the production of volume information through technological developments, and building the cluster on the demanding job for data.

In the consequences to handle the difficulties of clustering, there are a few researchers who can be an effort to develop the efficient concept of clustering. But still, some pitfalls are there in the present clustering algorithms which prerequisite to be solved to attain a better result.

1.2 Data Mining

This is the one of best exciting and dynamic research fields where it is extracted or skill of mining from the huge quantities of the data. The mining of data is a natural result of the advancement of information technology. The progression of the internet becomes more gorgeous where huge capacities of data and information are flooded. The data usage becomes common in the various fields related in engineering, nuclear science, biology, medicine, radar scanning, segmentation of image, study, and the progress of preparation, and mining of data. It is impossible for the human race to investigate for every data to recognize and the unexploited values of patterns. In mining, we are starving for skill, but are covered in the data. Novel intelligent techniques and automatic tools are requisite for converting this massive data into useful information. Data mining helps in discovering of treasured material and shapes. This technique applies to decision-making in corporate, business administration, marketing analysis, production management, science exploration, and engineering design. It is more utilities to carry the data mining as an essential part of the business.

The scholars from the various areas like database, artificial intelligence, visualization, and statistics are inspired by all these trends. Thus, data mining becomes most important advancement of interdisciplinary for development of information technology. The mining of data can be realized as dynamic steps of KDD (Knowledge Discovery in Databases) progress. It is defined as an important procedure of recognizing the validity, usefulness of theoretical, novel, and finally reasonable patterns from the huge volume of data. Data mining is sometimes coined as a synonym for KDD as it is an important and crucial

example of KDD. In figure 1.1 depicts the process of KDD and involvement of an iterative order of the steps as the following:

1. Selection of data is used to retrieve, and analysis from the considering database.

2. Data are integrated and /or cleaned in the pre-processing step.

3. Data transformation: data can be transformed or combined from the source format to destination format.

4. Data mining: A vital procedure was applied to the intelligent technique, in consequence to mine the pattern and also knowledge.

5. Interpretation: The interpretation or evaluation recognizes genuine, fascinating patterns, and illustrating skill based on the several remarkable measures.

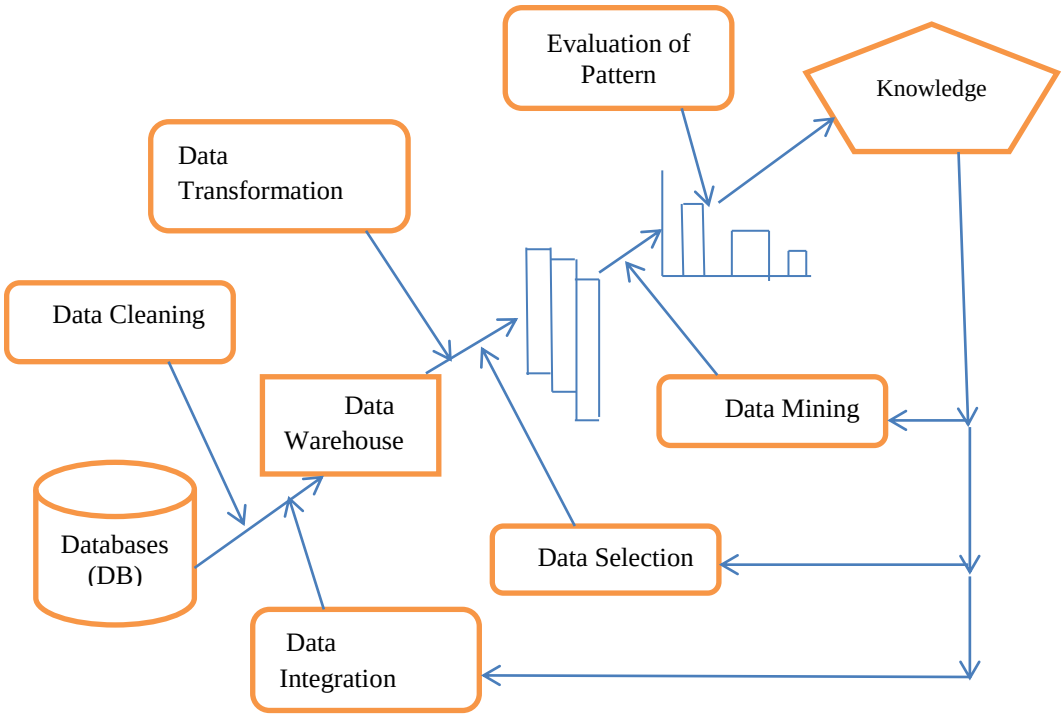


Figure1.1 Illustrate the procedure discovery of knowledge from databases

Data mining techniques are categorized by calculating huge quantities of data seriously. Performance and scalability are still the challenging problems in the mining field.

An algorithm is scalable if it taking the execution time increases the linear proportion of dataset size with the availability of system resources such as hard disk and primary memory. The innovation of method, relationship, forecast, and a study of the development, analysis of clustering similarity, deviation analysis, and classification is some of the data minimum functionalities.

1.3 Clustering

In the grouping process, similar types of objects are placed in a set of objects called clustering, similar objects in one cluster and dissimilar object in another cluster. This concept is used in the application for segmentation because the huge dataset is divided into the groups, allowing resemble and detecting outliers. The application of clustering is in different areas namely medical science especially in medicine, statistics, psychology, manufacturing (production), and engineering etc. For representation one example, in a particular organization, grouping and also categorizing the manufactured products but these products are not more demand in the market and may lead to support in dipping their manufacture to cut the losses. Further, in an academic organization, grouping based on the academic presentation of the student may help in classifying the students with lower grades. These students can be motivated to attend remedial classes to overcome their difficulties. In the data mining, health sector and clustering process may help in finding the link between disease's symptoms.

The clustering concept can be used in banking to lead to a set of the cluster of overdue payments by credit card, and research in marketing recognized the customers purchasing similar patterns. A lot of work is being done to apply these techniques in different areas. Clustering techniques are used in types of data like data matrix and dissimilarity of the matrix. The data is used in clustering numerical, categorical, ordinal, and the ratio of the scaled variable. The not-resemble (resemble) among the objects described by measured the based on Euclidian distance between pairs of the objects. Measuring distance between the two objects is used in more popular methods like Euclidean distance, Manhattan distance, and Weight Euclidean distance. Determined the distance between the clusters is like average, medoid, centroid, single, and also in between complete connection. The

centroid is the middle of the cluster and it needs not to be an actual point in the cluster but one centrally located object in the cluster is called medoid.

K-Means clustering comes close to is an extensively used partition based clustering algorithm which organizes input dataset predefined into the clusters. In classification, the speed and easiness of huge data reflect two features accepted in k-means concept. This procedure is applicable on numerical data; however, its extensions like k-modes, k-medians, and k-medoid types work on categorical and mixed data sets respectively. K-Means has a major limitation. The most repeated value related to k-mode for each attributes but K-Median middle value is taken for each ordered attribute.

It is very simplest theory identified for its speed. There is no expense in terms of cost and works well with high dimensional and huge datasets, but still there are certain limits in this technique. There is one major limitation that is creating the clusters are more reliant on the initially chosen objects as a centroid. The prototype centroids randomly pick this algorithm, which does not provide the same consequences of the same dataset for more iteration. More effort has been carried out to overcome its limitation. The requirements of this algorithm are scalability, the ability to handle the different kinds of attributes, create clusters with the attribute, skill to handle noisy data, high dimensionality, and interpretability.

A significant component of this concept measures the closeness between the data objects. If the component is the example of vectors in all its identical physical units, then it is probably that the metric of Euclidean distance is enough to effective the group related to the example of data. Therefore, in this case, measure the distance by Euclidian can be occasionally confused. Measurements are required for the same units; a knowledgeable conclusion has to be complete to the comparative scaling.

K-means method is applied in various fields optimized by the solution with a hybridized method (Anusuya and Lattha, 2011) [19], neurological disease (Alashwal, et al., 2019) [20], and in e-banking (Aryuni, M., 2018)[21], working in image processing (Pena, J .M. et. al., 1999) [29], clustering approaches (Celebi M. E. et. al,, 2013)[30], outlier detection problem (Han, X., et. al., 2013)[31], optimizing with numerical problem

(Scholkopf, B. et. al., 1997) [32], and hybridizing the concept (Mustafa, A. et.al. , 2015 ; Xinfeng, W. et. al., 2007) [33][34].

1.3.1 Measure of Similarity

Definition

It is a kind of a function known as the function of similarity. The similarity function is quantified on the real value between the two objects. This is the inverse of a metric distance, and it can take a high value of a similar object, but the value for a dissimilar object is either zero or negative. In Table-1, measurement of the Distance and Similarities of quantitative features are shown.

(Haq, E. U., 2017; Xu and Wunsch, 2005) A distance or dissimilarity functions on a data set $X = \{x_1, x_2, x_3, \dots\dots\dots, x_{n\}}$ is defined to satisfy the conditions as given below:

- Symmetry Rule:$D = (x_{l,}x_m) = D(x_{l\,,}x_m)$;
- Positivity Rule :$D(x_{l,}x_m) > 0, for\ all\ \ x_l\ and\ x_m$;
- Triangle Inequality Rule:

  $$D_n(x_{l\,,}x_m) \le D(x_{l\,,}x_n) + D(x_{n\,,}x_m), for\ all\ x_l, x_m\ \ and\ x_n\ ;$$
- Reflectivity Rule:$D = (x_{i\,,}x_j) = 0, if\ x_i = x_j,$ also hold, it is called metric.

Similarly, defined a similarity function and also to satisfy the conditions in the following manner:

- Symmetry Rule :$S = (x_{l\,,}x_m) = s(x_{l\,,}x_m)$;
- Positivity Rule:$0 \le S(x_{l\,,}x_m) \ge 1\ for\ all x_l\ and\ x_m$;
- Triangle Inequality Rule:

$$(x_{l\,,}x_m)S(x_{m\,,}x_n) \le |S(x_{l\,,}x_m) + S(x_{m\,,}x_n)|S(x_{l\,,}x_n), for\ all\ x_l, x_m\ \ and\ x_n\ ;$$

- Reflectivity Rule:$S = (x_{l\,,}x_m) = 1,\ \ if\ x_l = x_m$, it is called Similarity metric [1][16].

It measures from the table given below is required in different clustering techniques. We should apply it to determine the closeness between pairs of objects. The object of data is explained by a set of features, usually represented as a multidimensional vector. The

features can be continuous or binary, nominal or ordinal quantitative or qualitative, which find out the corresponding evaluating mechanisms (Alsayat and El-Sayed, 2016; Xu and Wunsch, 2005) [6][16].

Table 1 Measure of Distance and Similarities for Quantitative Feature

Measures	Forms
Minkowski distance	$D_{ij} = (\sum_{l}^{d} \mid x_{il} - x_{jd} \mid^{\frac{1}{n}})^{n}$
Euclidian Distance	$D_{ij} = (\sum_{l}^{d} \mid x_{il} - x_{jd} \mid^{\frac{1}{n}})^{n}$, n=2
City block distance	$D_{ij} = \sum_{l}^{d} \mid x_{il} - x_{jd} \mid$
Sup distance	$D_{ij} = MAX_{i<j<k} \sum_{l}^{d} \mid x_{il} - x_{jd} \mid$
Mahalanobis distance	$D_{ij} = (x_i - \mathrm{x_j})^r S^{-1} (x_i - \mathrm{x_j})$
Pearson correlation	D_{ij} =(1- r_{ij})/2 , where $r_{ij} = \frac{\sum_{l=1}^{d}(x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})}{\sqrt{\sum_{l}^{d}(x_{il} - \overline{x_i})^2 \sum_{l}^{d}(x_{jl} - \overline{x_j})^2}}$
Point Symmetry distance	$D_{ij} = Min_{j=1,2,3,\ldots n, and\ ,i\neq j} \frac{\lvert\lvert(x_i - x_r) + (x_j - x_r)\rvert\rvert}{\lvert\lvert(x_i - x_r)\rvert\rvert\lvert\lvert(x_j - x_r)\rvert\rvert}$
Cosine Similarities	$D_{ij} = Cos\alpha = \frac{x_i^r\ x_j}{\lvert\lvert x_i\rvert\rvert\lvert\lvert x_j\rvert\rvert}$

1.4 Types of clustering

In general, clustering techniques can be categorized as the following:

- Partitioning clustering
- Hierarchical clustering

- Density-based clustering
- Grid-based clustering

These clustering concepts are based on measuring a distance between pair of the given objects. Therefore, the aim is to measure the distance minimized of all objects from the cluster center which the object has its place in details (Han, J., et. al., 2011) [14].

1.4.1 Partitioned Clustering

This technique creates the clusters by dataset which partitioned into different kinds. Therefore, partition gives for each data object to the clustered index. There are k numbers of the cluster provides by the user and few criteria of function is used to find the solution of proposed partitioned. Find the cluster quality depending on the SSE (Sum of Squared Error) which can be determined by the distance between the clusters. Therefore, the partitioned method can be classified into the known algorithms as K-Means, partitioning around medoid (PAM), and CLARA.

1.4.1.1 K-Means Clustering

It is a very popular clustering technique suggested by Mac Queen in the year 1967, and created by the idea of the partitioning of the dataset. It can be applied in the various fields of research because it's simple perceptible mechanism. This takes the k input parameter and the dataset distributes a set of n number objects placed in k clusters. Therefore, a similarity of inter-cluster is very low, but the intra-cluster is high. The similarity of a cluster has determined the mean of data object in a cluster. This is viewed as the centroid of clusters.

Algorithm: This algorithm is applied for partitioning, where every cluster is a centroid Illustrated by the data object in the cluster

Steps- (1)**:** Choose the arbitrary no. of k object from the dataset as prototype centroids.

Steps- (2): Repeat

Steps- (3): reassign every object to the very closest centroid

Steps- (4): Update: centroid of every cluster, i.e., recomputed mean of the objects for every cluster.

Steps- (5): Until no change of centroids

This procedure has added an effective and scalable during the handling of huge datasets. The complexity is $O(n * k * t)$ for execution time.

Where,

n= total no. of objects

k= no. of the clusters

t= time of the iterations, normally $k < n$ and $t < n$.

The convergence criteria is widely used in the k-means algorithm and determined as the Sum of Squared Error (SSE) defined

$$Sum\ of\ Squared\ Error\ (\text{SSE}) = \sum_{j=1, x_i \in c_j}^{K} | x_i - c_j |^2$$

Where,

c_j= the mean of cluster

x_i= no. of instances in a dataset

Merits of k-means

Merits of the k-means as follows:

(1): This is more efficient for deals of huge datasets.

(2): It frequently ends at the local optimum.

(3): It is performed only on the numeric numbers.

(4): The shape of the created cluster is like coil.

Demerits of k-means:

This algorithm is very simple and has better convergence, but still has some demerits that are as follows:

(1): It is not fit for determining the clusters with the shape of the non-convex or very different size of clusters.

(2): When using the small class of the dataset, beginning of the grouping will find a significant cluster.

(3): Sensitive to the initial conditions, for taking the various initial conditions may create the various outcomes of the clusters and it may be stuck in local optima.

(4) Inputting the various sequences of the same data may create various clusters if lessen the number of data.

(5): This algorithm is never recognized as which data attribute backs into more alignment of the process, assuming that every attribute has taken an equal weight

(6): Measuring the erroneous arithmetic means it is not forcefully to outliers in this method.

(7): The experimental results have illustrated that the outliers can be problematic and its algorithm can robust to recognize the bad clusters.

(8): The experimental results have illustrated, that higher-dimensional data degrade the cluster performance.

1.4.1.2 K-Medoid

In this technique, it is represented that the partitioning is sampling data from the origin data. This method is useful in some special cases such as a cluster of the data is random objects. This is not important for discussing related functions of these given objects. Here represented one example, consider the discrete sequence of the object or set of the network, it doesn't say the significance concerning their median or means.

Similarly, these cases are the representative of partitioning which are taken from the given data and iterative procedures are applied in the consequence to improve the quality of its representatives. Therefore, in every iteration, one representative changes from the present data and also checks to quality of the cluster. Particularly, this technique requires a number of iterations compared to k-means and k-median algorithms. This method is used in structural data of the objects only. The Partitioning Around Medoids (PAM) and clustering LARge Applications (CLARA) are beneficial in partitioned clustering.

1.4.2. Hierarchical Clustering

In this method, creating a hierarchy of the cluster or constructing a tree of the clusters is known as dendrogram. This is a tree structure frequently used to analyze the consequences of a cluster. It can be categorized into the bottom-up approach (agglomerative) and top-down approach (divisive). In the first, bottom–up approach of the clustering begins with one point cluster and applied to the recursive procedure combined into two or more suitable clusters. In the second, the top-down approach of the clustering begins with a single cluster of all data points which is given and further splitting into separate clusters by the recursive procedure. This process is continued to attain the exit criteria.

The hierarchical technique has some difficulty from the fact at one stage where merger or splitting is complete, it can never be undone. This method cannot rectify the wrong decisions. Furthermore, there are two methods used in hierarchical clustering to improve the quality of clusters, (1) Chameleon: performs very well object analysis “linkages” at every level of hierarchical partitioning, (2) BIRCH Method (Balanced Iterative Reducing Clustering using Hierarchies):

1.4.3 Density Based Clustering

This clustering method is designed to recognize the arbitrary shape of the clusters. In density-based clustering, it is assume as a denser region of the objects in data space which can be alienated by lesser density regions. DBSCAN (Density-Based SCAN) is a very typical method of density-based clustering. In DBSCAN, the core point is like a fixed radius and in every neighborhood object; a cluster is consisted of at least the smallest number of other objects. The density of all the objects is calculated by including $\in -neighbourhood$ without earlier discretization. There is a cluster created by density-based model, defined as a collection of dense objects in $\in -neighbourhood$ including more than the smaller point.

1.4.4 Grid Based Clustering

In this clustering object space is quantized in the prearranged no. of cells created in Grid structure. Every clustering operation is implemented in the grid configuration. The advantage of this concept is that it has fast execution time, which is independent of data objects, but dependent on the cells in every dimension of the quantized space. Here, the two examples of the grid-based method are mentioned and they are: (1) STING and (2) CLIQUE.

1.5 Challenges in Cluster Analysis

The clustering concept is a more exciting and challenging research field. Its applications are required in various areas. The research efforts are focused on extraction methods and analysis of enormous datasets for more effective and also efficient in the mining field.

The following challenges as:

(1). Scalability: Handles the huge datasets by clustering algorithm for more scalable.

(2). Capability to handle various types of attributes: The methods should be more able to use for data like numerical data, binary data, and also categorical data.

(3). Detection of clusters with arbitrary shape: The algorithm of clustering should be able to find the arbitrary shape and not be confined to measure the distance to find out the small size of a spherical cluster.

(4). High dimensionality: The clustering algorithm to deal with low dimensionality as well as high dimensionality space.

(5). Ability to handle the noisy data: Consists of the database like noisy and misplaced or erroneous, therefore, few algorithms are more sensitive. Such type of the data may lead the cluster not well.

(6). Interpretability: The results of clustering should be more interpretable, usable, and more understandable.

1.6 Data Types

Here, a particular data type that shows a remarkable influence on the choice based clustering. In the past, clustering method was developed under the assumption of an attribute of numerical data. However, the data could be drawn from any types of possibilities like categorical (discrete), structural or temporal. Further, it is discussed about the various data types and its effect on procedure of clustering. The different types of data used in the analysis of cluster are:

- **Interval scaled variable**

  It is approximately measured continuously by the linear scale. There are several examples, including height and weight, longitude and latitude when the clustering of the houses, and also climate temperature is there.

- **Binary variables**

  It has two states Zero (0) and one (1): where 0(zero) means no variable and 1(one) means that it exists. If the binary variable is symmetric, then both states have equally valuable and transfer the same weight. There is no preference given on which the result may code like 0 or 1. Typical example like attribute of the gender has the male and female states.

  The dissimilarity based on the binary variables is known as symmetric binary dissimilarity.

- **Categorical variables**

  It has simplified the binary variable in a way that it can be tacked on greater than 02(two) states. Typical example, map colorist, a variable of categorical that may include the five states: blue, green, pink, yellow, and red. This variable is very simplified in a way that it can be considered as it is taken as greater than or equal to two states.

1.7 Applications of the Clustering

Clustering algorithms can be applied in various fields. They are as follows:

- Marketing: Finding a particular set of similar behavior of a customer from a given huge database of customer data, including their properties and also previous records of the purchasing.
- Insurance Company: categorizing the groups of the policyholders who claim their cost and detecting the frauds.
- Planning of city: Recognizing the groups of constructed houses based on their type, number, landmark, or geographical site.
- Studies of the earthquake: Detecting the very dangerous zone which observes the clustering epicenters of the earthquake.
- In biology: making the classification of animals, and plants according to their features.
- In the library: Maintain the orders of books.
- Image processing
- Spatial data analysis: create the maps in the GIS by feature spaces of clustering.
- WWW document classification: Clustering of weblog data is used to find out a group access of the similar patterns.

1.8 Identification of Research Gap and Problem

The clustering concept has played a more important role in the research area of data mining. There are different clustering methods available in the literature. There are few research challenges as follows:

- While analysis of the clustering, arises many issues in the method of K-Means like outliers, changing of the accuracy, and empty cluster.
- They are reflecting the contrary effects when initializing the incorrect and include the lesser convergence, unfilled clusters, and dropping in wrong local minima.
- The traditional K-means method losses its effectiveness due to increasing dimensionality.
- K-means traditional clustering method is quite sensitive for a set to the initial location of cluster centers.

- Centroids of the cluster may not be optimum, as the traditional method can be converging to find the local optimum results.
- The classical cluster is not denser, therefore convergence is slowed down.
- One of the big issues of traditional K-mean was that the cluster number at the start of the process of clustering used by the user.
- The traditional clustering has not good implementation in ambiguous and huge data sets.
- Still there is the deficiency of sensitivity to original centers in the traditional k-means method.
- Some essential modifications to consider necessity of the standard of K-Mean (traditional K-Mean), can be used to several sizes of data sets.
- Sum of Squared Error (SSE) of the original k-means is very high.

As a consequence of literature review, we proposed a framework for improve cluster performance by method with a hybridize technique, to overcome some limitations. Therefore, to eradicate the shortfalls of the existing methods, a new methodology is proposed.

Limitations:

There are few limitations as follows:

- **Handling Empty Clusters**

  There is one issue by earlier K-means that if no points are assigned to a cluster while step of assignment, then find the empty clusters. If this occurs, a technique is required to pick a centroid substitute, since otherwise the squared error would be greater than sufficient.

- **Outliers**

  Where outliers are present a representative cluster may not be followed on the centroids of a cluster (prototypes), and hence the SSE may also be higher.

- Minimizing the SSE during the process.

Drawbacks:

- For the better clustering, SSE must be minimized, which is a more complex task.
- It does not produce a similar outcome for all iteration since the final clusters are based on the initial random assignments. Minimize the intra-cluster variance but does not guarantee that the outcome has a minimum global variance.
- Value of K (cluster no.) measure is more challenging.
- Existing method did not work properly with universal clusters.
- Different kinds of beginning partition can lead the outcome in numerous last (i.e. final) clusters.
- Different sizes of clusters of dataset and its associated densities do not work properly.
- Outcomes are more dependent on selection based set of initial centroid.
- Using trial and error in selecting the number, implementation of the SPSS method is limited to calculate the distances between samples using Euclidean distance.

1.9 Motivation for Research

Computational methods for handling datasets have become an essential part of science, engineering, and business. Another ways, the extensive usage of computers increases the amount of data warehouse in different forms of dimensions, instances, and data types. These datasets become more problematic with huge dimensions or huge instances. Historical data plays a crucial role in most of the decision-making and strategic planning. Traditional methods do not perform well to analyze these stored data when the data instance or dimension is huge.

The KDD is an emerging research field, and discovers information from the huge volume of data. The process of KDD starts from data collection to data interpretation. To a certain extent, KDD is also called data mining which plays a vital step in KDD processes. Data mining mentions taking out valuable information from huge quantities of raw data. The mining procedures like association analysis, classification analysis, and clustering have been

widely used by different users for purposes. Classification is working under the supervised technique, which classifies the data with predefined labels. To find out correlation relationships or interesting associations among data, association analysis is used.

Apart from classification and association of analysis, they are used as a standalone tool in the most cases, clustering, an unsupervised MLT (Machine Learning Technique), is used to identify underlying data structure and can also act as a pre-processing tool. It is worthwhile to do research in the cluster to enhance it.

1.10 Scope of Research

This research allows running an evaluation study of four widely used clustering methods (K-Means, PSO, PCA, and Genetic algorithm) with the scope of clustering group's assessment and the cardinality of the methods used to assess their preferences. The capacity of the proposed algorithm is characterized for coding the system with a prototype and it's worth pointing towards cluster quantity to which the particular outline fits in K-Means. The survey of the k-means concept in the various aspects is studied and a measuring function is proposed with this algorithm for clustering the various datasets. The proposed algorithm accurately maximizes the cluster precision by diminishing outliers.

1.11 Objective of the Research

To achieve the main objectives are enlisted following as.

- To proposed an efficient framework of clustering approach.
- To proposed an effective proposed method that improves clustering metrics.
- To analyze the clustering of K-Means with other approaches using the software tools.
- To proposed and analysis the fitness objective function using Genetic Algorithm (GA).
- To analysis the clustering metrics SSE using RBNN of ANN.

1.12 Organization of the Thesis

The organization of this research thesis work is divided into 9 chapters as follows:-

Chapter 1

In chapter 1, the thesis introduced the main concept of data mining, details of the clustering concepts and the proposed objective of the overall thesis work is given. This chapter deals with the concept of clustering algorithms and identifies the appropriate problem, and contribution to this thesis work.

Chapter 2

In chapter 2, the critical review of literature survey of the earlier related work in the clustering field k-means and further also provides a literature review of Principle Component of Analysis (PCA), Particle Swarm Optimization (PSO), Data clustering, and concepts of genetic algorithm, fuzzy approach and RBFNN concept.

Chapter 3

In chapter 3, design an efficient framework of research descriptions of proposed clustering algorithm, its utilities and limitations are mentioned. In this chapter we have proposed the research methodology. The proposed clustering method to hybridize with Principal Component Analysis (PCA), and Particle Swarm Optimization (PSO) used to improve quality of clusters. Furthermore, we have included statistical analysis with software tool SPSS 17.0, artificial neural networks concept of radial basis function neural network, genetic concept is also helpful to detect the significant result compared with k-means to enhance the clustering.

Chapter 4

In chapter 4, proposed the methodology of research related to the clustering algorithm. Details of how to work principal component analysis utilized with the proposed AEIKM Clustering methodology to overcome the dimensionality problem of huge datasets. And also we have discussed the particle swarm optimization for the fitness of clustering in chapter.

Chapter 5

In chapter 5, the analysis of k-means clustering via PCA with a statistical tool is discussed to analyze the reduction of the component based on the eigenvalues. Analysis the min distance between initial centroid of creates clusters for considered datasets, F-ratio. And also using the fuzzy approach is measures of the significance of created clusters and its analysis in this chapter.

Chapter 6

In chapter 6, the analysis of fitness function of k-means with a kernel fisher's fitness function is explained and applied the optimized approach of GA (Genetic Algorithm). Further, we have compared the analysis between them the significant of the best fit along with means.

Chapter 7

In chapter 7, discussed the methodology of research by the RBFN (Radial Basis Function of neural net) concept of the sub field of Artificial Neural Network (ANN), and this result is compared with outcome of the proposed method of clustering. In this chapter we have discussed the analysis of Sum of Squared Error (SSE) value of testing and trained for used the four datasets and also to analyze the results.

Chapter 8

In Chapter 8, experimental results and their analysis are provided. It clearly describes the implementation and performance by used ideas of PCA, PSO, RBFNN, and GA. There are four data sets are taken from the Machine Learning UCI repository for empirical analysis of the proposed clustering algorithms.

Chapter 9

In Chapter 9, discussed the thesis concludes remarks and findings of the proposed research work. Some of the approvals for future work are also discussed.

1.12 Summary

To identify and raise the data is for better understanding as well as for better categorization of techniques. This leads to the involvement of researchers in clustering to

handle the results. There are different approaches suggested in data clustering. The data applied in every concepts of clustering has its own benefits and drawbacks.

K-Means concept is a simple and efficient algorithm that is widely used for data classification, which helps to analyze data in turn. But still, it has many demerits when handling big data. K-means method is assessed in detail and three of its major demerits are identified. We proposed new methodologies hybridized used concept namely PCA, PSO, which automatically boost the efficiency of the K-Means algorithm, applied the optimized concept of genetic to find the significant best fit, and used RBFN Concept of ANN. Further, in this thesis, the identified demerits are tried to resolve.

CHAPTER 2

LITERATURE REVIEW

In this chapter,we have to go through the various reviews of papers as a concern about my thesis title. The pattern of the clustering can see the furnished set of input data, and further, it may be used for analysis. For example, the theme of data mining's steps might identify and recognize the several groups of data, which can be used to get more accurate predicting results. Here, we have focused on clustering, dimensional reduction, optimizing concept, GA, and RBFNN related papers.

2.1 Introduction

Critical review going through the literature facilitates helps the researchers to attain updated information in a selected field of study. Therefore, earlier related to this research and concern about the theory can be supportive to formulate idea of the problem, help to understand of finding the outcomes, and also guiding the study in the right way of the research. In literature, surveying is a more significant progress of the study in certain field for the researchers who want in any kind of research.

The rigorous review of the past research efforts on the problem is the required step to find the conclusion. This surveying of the literature is assisted to investigate the problem in various dimensions, explores every possible technique, and also applies it to be approached as a result.

The aims of the literature's surveying are as follows:

1. Explore the ideas, theory, and hypothesis apply in framing the problem.
2. To propose a method for proper research of the problem.
3. To find the necessities of data assists in the explanation of the consequences.
4. From the literature, surveying has exposed whether the proof already exists to handle the theory of problem is sufficient without more examines and also avoids the possibilities of repetitions.

5. Attain the clarity of the concepts, requisite hypotheses of study the no. of published research papers like a research journal (ACM, Springer, IEEE explorer, and IEEE transaction et.), periodicals, printed and e-books, encyclopedia, and magazines. The review of research studies has examined the consequent in our country and abroad.

The literature reviews were found on the basis of a related article published in various journals and magazines. These surveys are the different research articles, including that which is accessible like the clustering concepts, k-means, principal component analysis, particle swarm optimization, fuzzy, and RBFN concepts.

2.2 Review

The research paper reviews with related issues as following manner:

Capo M. et.al (2020), the author worked out the analysis of huge datasets is continuously in the mining fields. So, design and development of efficient algorithm used in unsupervised learning. The k-means concept of clustering is the most important technique to implement the easily a large dataset with lower cost. Further, some drawback is in this algorithm depending on a large-scaled dataset on the initial condition. The author developed an algorithm on the instances and reduced the problem without affecting the quality of cluster. There are applied some small weights as the representative to distribute for creates for originally given instances. The proposed concept of experimental work is well performed and found a good quality of cluster [109].

Aslam A. et.al (2020), authors have done the effort to develop the k-means in the terms of accuracy based on "availability" of cluster-level k. The authors study and expressed their work improved k-means by selecting the initial center and sorted the datasets by Euclidian distance. They compared the results along the iterative steps and distance find within the cluster. This modified result is better as compared to the old k-means and also finds the good accuracy of such an algorithm [110].

Ahmed M. et.al (2020), the authors discussed the synoptic overview of the research idea on k-means and reduced the shortcomings. They have focused on the recent development

of k-means algorithms. The different datasets apply to their algorithms, analyze the experimental results, and also compared the experiment results between them algorithms. Furthermore, progress of various k-means algorithms to help and suggested a new direction of investigation in mining [112]. In telecom business Alkhavrat M. et.al, (2020), appearances of a large volume dataset for main motive to the determine such as lessen the n^{th}dimension of factual data, lessen the clustering, and analysis. Authors Aydin and Yurdakul, (2020) are study covid-19 analysis against 142 countries. In this paper measure, the clustering analysis and discovers, infections factor and death cases are also examined [111].

Baruri R. et.al, (2019), authors studied the concepts of k-means and also discussed its limitations. The author suggested improving the first version by the greedy concept and further enhanced the old k-means such concepts implemented in python and also validates its work with the optimal method of clustering [113]. The authors emphasized threshold values as a center of the k-means technique based on creating the clusters. Measure the distance between two data points is less than equal$(i.e.\leq)$ to the threshold value than the two points in the similar group otherwise in a dissimilar group. The authors developed a new modified technique to eliminate the existing problems in its original algorithm. This proposed concept dynamically creates the clusters for applies dataset. The author compared the results between the proposed and existing algorithm. The experimental result of the proposed idea is improved as co-related to the original algorithm (Hossain M. Z., et. al., 2019) [26].

Author Al-Zubai I. M., et.al. (2019) they focused on statistics of dataset dimensionality has set of attribute and such variety of data used in the research study and mining concepts are employed in this field like as telecommunication industries for helping the administrative strategy [54]. The authors are used PCA concept in reduced dimensions for changing original data for mining and classifies it's via k-means. Find out the consequences to illustrate the more accuracy of reduced dimensionality of large dataset for analysis. World-wide cause deaths are in among the woman by "breast cancer, and prediction" of its disease to opportunity of it is cure otherwise more risky for health author discussed in detail (Zhang, N., et. al., 2018; Jamal, A., et. al., 2018) [51-52].

Haq E. U., et.al. (2017), addressed the data-mining comprises several concepts like learning by machine, learn of the statistical, database, knowledge-based portfolios, artificial intelligence, neural network, and provides a few well- known clustering algorithms [1]. Advantages of k-means applied in recognition of pattern and its mining for better results (Ali and Kadhum, 2017) [24], prediction and analysis of datasets (Kumar and Kaur, 2017) [25]. In K-Means, modify the centroids, create dissimilar result, and produced the local optimal problem, working outlier detection problem details in (Xu X., et. al, 2018) [18]. Patel and Gondaliya (2017), they have explored in his research work. The authors focused analysis of the effect on the classification theory. In his work express the presentations of the students applied the appropriate concept and also find its accuracy and error rate. They are considering the different types of concepts like rule-based, decision tree, Bayesian [2].

Bae Y., et.al. (2017), authors have addressed his special issue on fuzzy inference system and also explore the information on the area of data mining and its applications. The purpose of this special issue is to survey the most recent modeling related to the concepts like mining of information and fuzzy. Rigorous review criteria are on the basis of procedural contributions, improvement, and completeness [3]. Alsayat and El-Sayed (2016), authors proposed a framework in his research work, the title of the research paper "Task of Detecting Communities by Clustering Message from Large Stream of Social Data". This framework uses the technique of k-means with GA and also applied the optimized cluster distance method for clustering. The aim of this proposed framework is to remove the problem of basic k-means for choosing the first good centroids by GA and got it more accuracy of the cluster by the OCD technique [6].

Jalil A. M., et.al. (2016), they are discussed and addressed the problem in his research work on the procedure of extracting the knowledge from the given databases. They have proposed a comparative study between several clustering algorithms. There are identified two problems; one is correlated to the quality which more converged, and the second is execution in requisite time to reduce [5].

Wang and Bai (2016), they are discussed in his research paper, proposed modified the min- max K-means technique with particle swarm optimization to find out the suitable values of their limits which can raise the issue of the concept to attain the minimum error of the cluster. This process is verified on some preferred datasets in various conditions and also compared to K-means and the technique of min-max K-means. They were proposed the concept can get the minimum error of the cluster [4]. Author Wang D., et.al (2016), proposed the significant concept of network is symbolized in a lower dimension to protect the structure of network node [36]. Boobord F., et.al. (2015), they are discussed in his research paper (1) PCA is used to reduce the redundant dimensionality of the dataset. (2) WK-Means is a hybridized of Invasive Weed Optimization (IWO). (3) Algorithm of k-means utilizes a reduced dataset to get the best possible clusters. It method is tested on five real-world examples and outcome are compared with the proposed algorithm to produce a better performance on datasets [9].

Rathore and Shukla (2015), the authors discussed in his research work in the course of cluster analysis arises the more issues in conventional k-means concept like that changing of accuracy, unfilled cluster, and also outliers. In his proposed work there is some modification in the conventional k-means technique and also implemented it. At the last, the performances of the suggested idea are comparing and find the cluster quality as well as some data point removing from the outlier. Measure the validity of cluster based on the classification such that criteria precision, recall, and F-measure, and rand index (accuracy) [7]. Haraty R. A., et.al. (2015), they have proposed the *G*-Means algorithm, and it is applied on the large data set. G-Means algorithm outperforms K-Means in the word of like as F-score, coefficient of variance, entropy, and execution time also [8]. An improve the k-mean cluster high-quality result, reduced the no. of iterations and better constancy and rapidity of the data processing based on reducing and grid (Ma L., et. al., 2015) [27]. Kamel N., et.al. (2014), they are proposed comprises the several concepts on K-Means, PSO, and Sampling algorithms. Evaluated on data sets and the outcomes are compared to each other [10]. Eslamnezhad and Varjani (2014), they are addressed using the min-max K-Means cluster technique. It is resolved the deficiency of K-Means related to the centers and improved the cluster quality. Also, the technique has a high rate of

detection as well as a lesser rate of false-positive detection [11]. The Min-Max K-Means concept allows K-Means produced to a high quality of clustering during pick initial center is randomly and it more systematically but reduced the variance of intra-cluster (Tzortzi and Likas, 2014) [23].

Anusha and Sathiaseelan (2014), address in a research paper is to maximize the clusters firmness with larger separation between at least two clusters. In his work, compare the two algorithm superiority of EKMG with GGA based on the 'dataset' which is required in daily life, and show the optimum conclusion with accuracy. The Genetic Algorithms are used here because it is a randomized searching concept that provides an improved most favorable result for the fitness function of the optimized problem [13]. Ganganath N., et.al. (2014), propose his research paper some essential modifications in the K-Mean. The modified algorithm of k-means is used and got the outcome a cluster in the anticipated sizes. A possibility of advantage can be acquired the equal sizes of clusters. Therefore, the modified technique creates the cluster being used the past information. It can help to evade local minima; and a modified algorithm leads to accurate results [12]. Suganya and Shanthi (2012), they have discussed in this research paper the soft as well a hard clustering method. In hard clustering, data are divided into separate clusters means every object belonging the same one group. In the soft clustering technique, every object belongs to more than a single cluster, but every object having in a set of membership. This idea is applied to the real dataset to make a better result and is also used in mining [15].

Author Yong and Xincheng (2012), presented his research work in the 12th international conference ICCSE, IEE, identify difficulties and how can enhanced the minority class performance. They are assessed the criteria of the mixed dataset of two classes there is one minority and secondly majority basis on the right and false classification. It is reflecting the performance of its classification and validates the outcome by KNN with supporting vector machine (SVM) [64]. We have study the analysis of three algorithms discussed by author among them first Genetic Algorithm (GA), second Differential Evolution (DE), and third Particle Swarm Optimization (PSO). Further, it is more

beneficial for separate optimization over two algorithms details in (Kachitvichyanukul V. , 2012)[65].

Han J., et.al. (2011), proposed the idea of data mining and clustering approaches for huge quantities of data. They are described by what method to be calculating the dissimilarities of object signified by their different attribute. There are different kinds of clustering like a partitioning concept (k-means, k-medoid, k-median, k-mode, CLARA, CLARANS), hierarchical, density-based, and grid-based clustering method. In this clustering, ideas can also be prerequisite for detection of the outlier [14]. The K-Means algorithm is modifying the centroids, creates a dissimilar result, local optimal problem, and tries to reduce the sum of squared distance and more sensitive for initial centroids (Yelda M. et. al., 2010) [22]. Min and Siqing (2010), author discussed the traditional k-means technique is generally required for clustering of huge data set because it's a very simple concept, and more convergence of the data. It is more responding to the first centroid of the cluster. The cluster of huge-data set is generally affected by the data point. In this algorithm has few drawbacks, but Genetic algorithm used to overcome the responsive of first centroid of the cluster as well as reduced the some data points impact and get more accuracy and high-quality of cluster [63].

Hongxia P., et.al. (2010), they are discussing the idea of kernel trick KPCA and optimized theory of PSO. Kernel's idea is applied to the dataset for feature extraction and measures its fault performance [17]. The author Allam M. N., (2016) proposed the particle swarm optimization concept and discussed it in details [28]. Rui and Donald (2005), they are proposed in his survey paper of clustering algorithms. The data sets are appearing in surveys on different areas like computer science, machine learning, statistics, biological science, TSP (Travelling Salesman Problem). The author is also mentioned in his research paper various topics like a validation of the cluster, proximity calculation [16].

K-Means algorithm is more recognized clustering method and due to the desirability of its good computing efficiency. Therefore, this algorithm is more sensitive of selecting the centroids (Vesterstrom, J., et. al., 2004) [36] and several limitations of its algorithm. In k-means, modify the centroids, create dissimilar result, explore in (Mirkin B., 2012) [35],

but nowadays the researches are adopting the optimization concept found the good cluster and determine the same basic problem (Shi-Wei, L., et. al., 2010; Sethi, C., et. al., 2013) [44-45].

The quality of the cluster is enhanced of k-means integrated with optimization technique of Ant Lion optimization (ALO) and get better quality of cluster for different metrics on the basis of performance (Ratnaweera, A., et. al., 2004) [37]. An improve the k-mean cluster high quality result, reduced the no. of iterations and better constancy and rapidity of the data processing based on reducing and grid (Liu, X., et. al., 2010) [38], cluster performance environment for big data (Poli R., et. al., 2007)[39]. Improve the quality of K-Means performance from ant colony optimization via two phase procedure (Majhi and Biswal, 2018) [40], gene expression analysis of data based on the minimal spanning tree [MST] (Mary and Raja, 2009) [41]. Consequences of Principal component Analysis (PCA) are valid as input data on PSO to decrease the co linearity between the variance (Ren and Zhuo, 2011) [42], and improved k-means clustering (Ma, L., et. al., 2015) [43].K-Means technique can attain the more rapidly convergence, and local optimization (Shi-Wei and Xiao-Dong, 2010) [44].

Author Peres-Neto, P. R., et.al, (2005), they used PCA analysis tool the summarized of giving a regular set of the patterns; determine the deviation for various kind of variables, covariance and also find the performance of the dataset [55]. The PCA is can only extract a linear projection of the data. Consider the consisting of data like as $X = x_1, x_2, \ldots., x_M$ are M vectors authors explained in (Ding and He, 2004)[50]. Authors study the cluster techniques in detail, as well as the elimination of the $(k-1)^{th}$ term from the matrix covariance, PCA projection of the upper to the lower- dimensional space, data placement in the lower space graph, and also the utility of k-means (Ding and He, 2004) [57].

Using k-means with Genetic algorithms, determine the most favorable consequences of the existing challenge faced by every family in terms of how to tackle the commercial plan for accessing the cluster of fiscal and communal conduct. (Babaie S. S. et. al., 2015) [66] and producing the several safe and sound clusters for a large data assemblage have been described in detail (Bhatia, S., 2014)[67].These methods combine to explain the

manifold traveling salesman problem and also proposed genetic algorithms obtained the utmost cluster quality (Rahman M. A., et.al., 2014; Lu Z., et.al., 2016) [68-69]. The route optimization problem and congregate the global result in term of the accuracy, computing time and speed of convergence for online real application (Aibinu A. M., et. al., 2016) [70], and more applications are discussed genetic algorithm with K-Means (Zeebaree D. Q., et.al. 2017) [71]. Identify the problem like consumption of energy, how to strengthen life time of network and they proposed method to support the more in clustering (Barekatain B., et.al., 2015) [72]. A quantum-inspired genetic algorithm with k-means proposed by authors is overcome the convergence of the local optima by GA (Xiao J., et. al., 2010) [73].

2.3 Particle Swarm Optimization (PSO)

A paper published in 1995 at the international conference on the evolution of computation. Introducing the article in conference, the scenario of using its PSO conceptual theory paper to solve various types of issues and problem optimizers is then changed. It is simple and interesting perspective that aids in the worldwide exploring process. (Krishnasamy G. et. al., 2014)) [72], and PSO theory detailed in ([Kennedy and Eberhart (1995); Tharwat A., et. al., (2017); Van der Merwe and Engelbrecht, 2003) [75, 76, 77].

The comparing between genetic algorithm and PSO to found the result PSO is better because it's local search and global searching at the same time. The reflection of PSO is poor and undignified for smaller size of the population, but due to the time bound PSO is good (Prashanth N. A., 2018) [74]. (Kennedy, J., et.al., 1995; Tharwat A., et.al., 2017), author published his research paper in 1995 at the international conference on the evolution of computation. Introducing this research paper in conference, then after change is the scenario of using its paper of PSO conceptual theory to handle the various kinds of complex optimizer problem. It is more simple theory to support the worldwide incisive process [75-76].

In PSO method, result of population is known as a swarm of the particles and as well as find this result shown as the particle. Further all particles have velocity and position. The Particle is being in the move to other position with velocity. When occur the next position is the best position, then which required to update both its location and velocity and this process is repetitive until found the criteria in presented (Van der M. D., et.al., 2003; Zhao W., et.al, 2009; Xinchao Z., 2010) [77, 79 and 81].

This algorithm is an optimization concept of the population based on stochastic global optimization, and its applications. The Particle Swarm Organization (PSO) algorithm useful to dynamically accelerate the constraint is CPSO abbreviate form improve the convergence rate and trend its rate with optimized period (Min and Siqing, 2010; Van der Merwe and Engelbrecht, 2003) [63, 77]. Position and velocity set to initialize according set to the boundary value is upper and lower limit as

$$X_0 = \text{rand}/2((\text{upper limit of particle} - \text{lower limit of particle}) + (\text{upper limit of particle} + \text{lower limit of particle})) \quad (2.8)$$

$$V_0 = \text{rand}/2((\text{upper limit of velocity} - \text{lower limit of velocity}) + (\text{upper limit of velocity} + \text{lower limit of velocity})) \quad (2.9)$$

Computing the fitness function is the motive of contender to apply the objective function evaluation.

The standard PSO algorithm according to the equation (2.8) and (2.9) as follows

$$v_i\ (t+1) = \omega * v_i(t)\ +\ L_1 * random_1(p_i - x_i(t)) + L_2 * random_2(g_i - x_i(t)) \quad (2.10)$$

$$x_i(t+1) = (x_i(t) - v_i(t+1)) \quad (2.11)$$

Where ω = weight of inertia, L_1 and L_2 learning parameter , random1 and $random_2$ are random number generated between 0 and 1, p symbolize the best solution, and g symbolize the globally best solution.

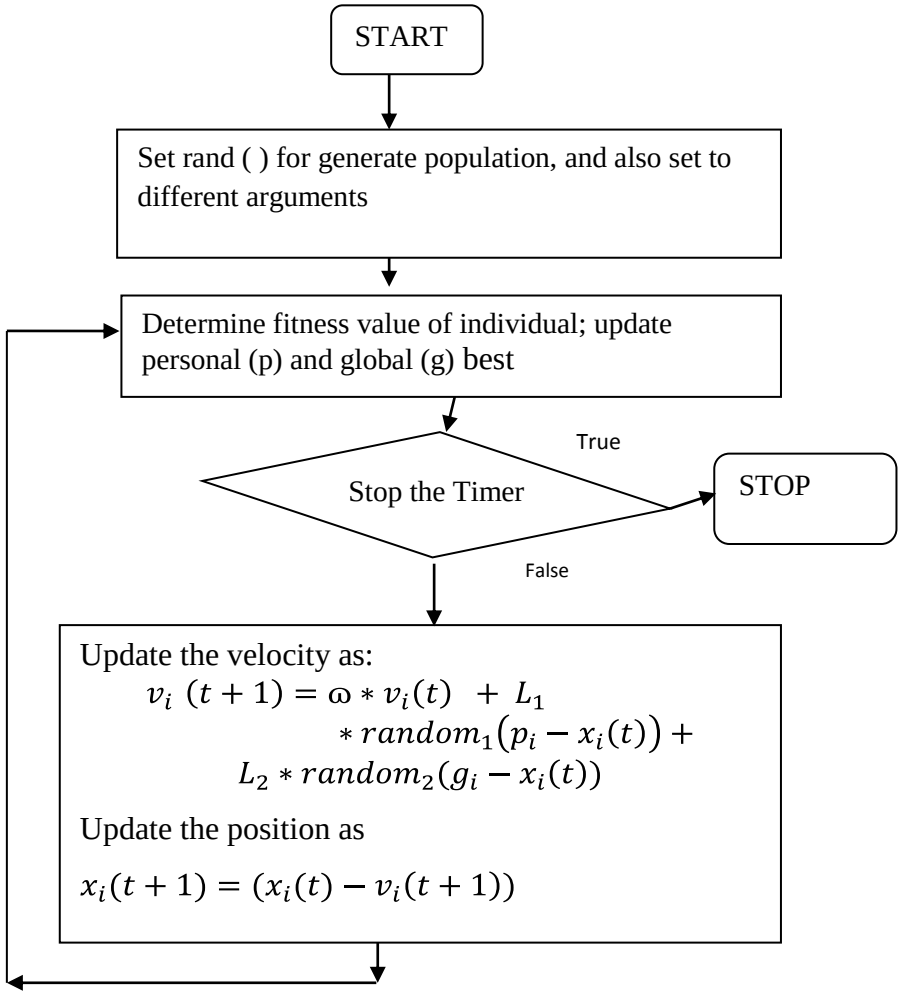


Figure 2.1 Flow chart the procedure of particle swarm optimization

There is several research paper uses the concept of particle swarm optimization included in (Ho, S. L. et.al. 2007; D. Vander Merwe, et.al, 2003; Xinchao, Z., 2010; and Zhao W. et.al, 2009). The procedure of this method is represented in Figure-6.1.

2.3 Kernel Trick

There are given a sample set of data A ={ a_1, a_2, a_3, …..,a_n } and every member of data point belonging in domain field R^N , where R indicate the set of domain , N represent the value of 1to n. Nonlinear mapping mathematically represented as the following manner

$$\emptyset : R^N \to F, a = \emptyset(a) \quad (2.1)$$

Where, $\emptyset$ = represent the nonlinear function of mapping, and the symbol F indicate the range of function.

Note that function of Kernel $K(a_i, a_j)$ is a function of input space and its advantage to avoid the mapping $\emptyset(a)$ at all.

$K(a_i, a_j) = \emptyset(a)_i^{Transpose} \emptyset(a_j)$ With Mercer's conditions (2.2)

The model recognized of the PSO kernel parameter based on FDA instruction. There two type of characteristic samples in feature space F. The first dataset

$A_1 = \{a_{11}, a_{12}, a_{13}, \ldots\ldots\ldots\ldots, a_{1i}\}$,

And second dataset $A_2 = \{a_{21}, a_{22}, a_{23}, \ldots\ldots\ldots\ldots, a_{2j}\}$,

Where

$i = 1, 2, 3, \ldots\ldots, n_1, \quad j = 1, 2, 3, \ldots\ldots, n_2$

The mean vector of two type feature space in F is, Mean vector of one feature Space (F1) is

$$\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \emptyset(a_{1i}) \quad (2.3)$$

Mean vector of second feature space (F2) is

$$\mu_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \emptyset(a_{2j}) \quad (2.4)$$

Determine the square distance between the mean vector of two vector spaces F1 and F2

Square Distance = $|\mu_1 - \mu_2|^2$

$$= |\mu_1 - \mu_2|^{\text{transpose}} |\mu_1 - \mu_2|$$

$$= |\frac{1}{n_1}\sum_{i=1}^{n_1} \emptyset(a_{1i}) - \frac{1}{n_2}\sum_{j=1}^{n_2} \emptyset(a_{2j})|^{tranpose} \left|\frac{1}{n_1}\sum_{i=1}^{n_1} \emptyset(a_{1i})\right) -$$

$$\frac{1}{n_2}\sum_{j=1}^{n_2} \emptyset(a_{2j})|$$

$$= \frac{1}{n_1} * \frac{1}{n_2} (\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \emptyset(a_{1i}) \emptyset(a_{2j})) -$$

$$2*\frac{1}{n_1} * \frac{1}{n_2} (\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \emptyset(a_{1i}) \emptyset(a_{2j}))$$

$$+\frac{1}{n_2} * \frac{1}{n_2} (\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \emptyset(a_{1i}) \emptyset(a_{2j}))$$

$$= \frac{1}{n_1} * \frac{1}{n_2} (\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(a_{1i}, a_{1j})) -$$

$$2*\frac{1}{n_1} * \frac{1}{n_2} (\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(a_{1i}, a_{2j}))$$

$$+\frac{1}{n_2} * (\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(a_{2i}, a_{2j})) \quad (2.5)$$

Determine the dispersion within the sample of feature space F1

$$\text{Disp\_w\_f1} = \sum_{i=1}^{n_1} |\emptyset(a_{1i}) - \mu_1|^2$$

$$= \sum_{i=1}^{n_1} \emptyset(a_{1i})^{transpose} \emptyset(a_{1i}) - n_1 \mu_1{}^{transpose} \mu_1$$

$$= \sum_{i=1}^{n_1} K(a_{1i}, a_{1i}) - \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(a_{1i,} a_{1j}) \quad (2.6)$$

$$\text{Disp\_w\_f2} = \sum_{j=1}^{n_2} |\emptyset(a_{2j}) - \mu_2|^{transpose}$$

$$= \sum_{j=1}^{n_2} \emptyset(a_{2j})^{transpose} \emptyset(a_{2j}) - n_2 \mu_2{}^{transpose} \mu_2$$

$$= \sum_{j=1}^{n_2} K(a_{2j}, a_{2j}) - \frac{1}{n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(a_{2i,} a_{2j}) \quad (2.7)$$

Arrange the particle swarm optimization fitness function according to the fisher's discriminant criteria from the equation (2.5), (2.6) and (2.7) respectively.

2.5 Genetic Algorithm (GA)

To develop the idea of a genetic algorithm by the Goldberg who was inspired the concept theory of evolution proposed by C. Darwin's. In this theory C. Darwin quotes survival of an organism can be maintained the procedure of reproduction, crossover and also mutation. The evolution concept applied to the computational algorithm known usually

trend as like objective function. A solution produced by genetic algorithm is called chromosome, the population is another way an assortment of its chromosomes workings. It's are matched to the Genes to determine whether they have a numerical value, a binary stream value (0's and 1's), a symbolic value, or a character, reliant on the complexity.

The procedure of this algorithm is represented in Figure-2.2.

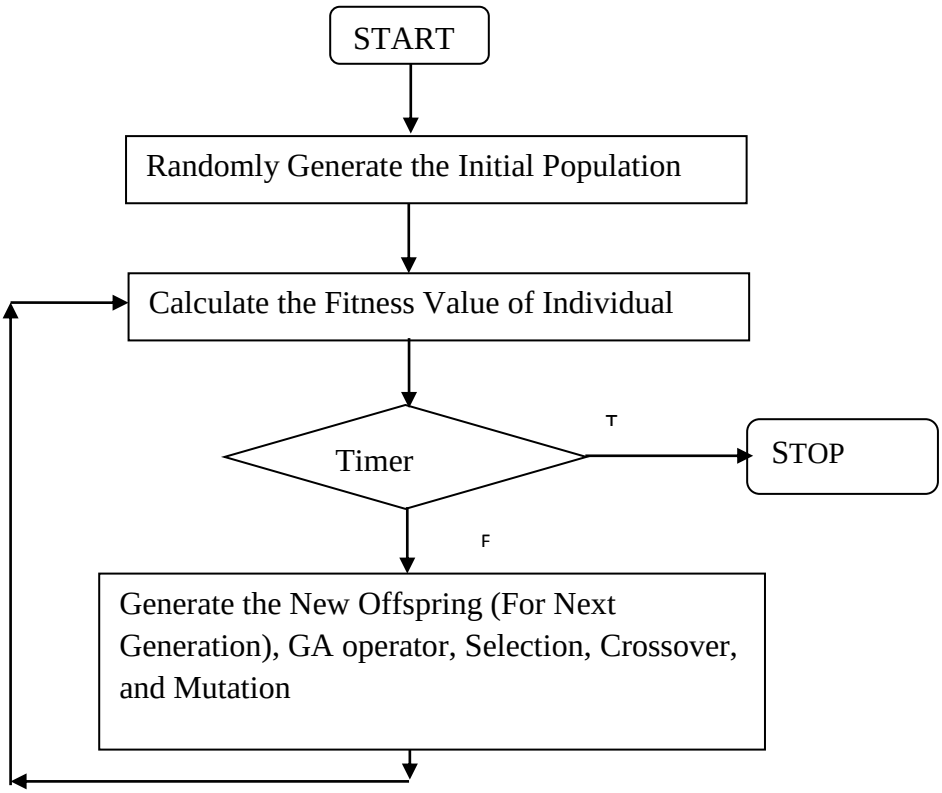


Figure 2.2 Flow Chart procedure of the Genetic Algorithm

These chromosomes are going through the procedure called fitness function, and find appropriateness of problems generated by the genetic algorithm. The higher fitness values of chromosomes have more possibility to prefer in the next generation (Hermawanto et.al. 1997; Chang et.al, 2009) [58, 82, and 85]. The details about the genetic algorithm are available can found to proposed the techniques by researcher Holland in 1975, and by Goldberg in 1989.

Kumar A., et.al. (2016), authors discussed the benefit of genetic algorithm provides more efficient, effective algorithms for optimization, and follows the heuristic searching. This is more beneficial for the searching area extremely high, in Table 2.1, show the comparative analysis and between k-means and GA [107].

Table 2.1 Comparison of k-means and GA algorithm

K-Means Algorithm	GA Algorithm
1. Partitioned based 2. Input: k, dataset, randomly chosen centroids k 3. Objective: Min SSE 4. Final cluster converging to local 5. Termination criteria: no changes in creates new cluster centroids 6. Time complexity:$O(n * k * I * d)$	1. Evolutionary based 2. Input: k, p ,randomly chosen, p chromosomes, t-max 3. Objective: Min sum of distances from every object to cluster centroids 4. Global searching method 5. Termination criteria: MAX no. iterations reached 6. Time complexity:$O(t_{max} * p * k * n * d)$

Where

n= no. of object,

k= no. of clusters,

d= dimension of dataset,

I= no. of iterations,

$t_{max} = \max no. of\ iterartions$, and

P= size of population

2.6 Fuzzy Concepts

Clustering of data is an unsupervised learning concept to extraction of data and ideas of this method the same type data put in the one cluster and other data place in various clusters. Dunn's index is the second component to the measure of cluster quality, it depend on the minimum separation and max-intra-cluster distance (compactness). The Dunn's index is minimum separation divided by max intra-cluster distance are discussed in (Bezdek and Pal, 1995) [92]. Bora and Gupta (2014) authors explained the comparatively studied between hard and c-means of fuzzy clustering and discussed mentioned the conclusion also in [94].

Authors discussed the measure internal quality of cluster: Cohesion of the cluster is sum of assign the weight every links in a cluster its measure by sum of squared errors (SSE) or WSS (Within the cluster Sum of Square), and sum of weights between the node object in cluster and outside the cluster known as the separation (between the cluster sum of square) of the cluster in a graph based cluster. $SSE = WSS = \sum_i \sum_{a \in C_i} (a - m_i)$, $BSS = \sum |C_i|(m - m_i)^2$, $|C_i|$ = Size of the i^{th} cluster. Average silhouette is the measure the cluster validly; pick exact number of cluster, relative quality of the cluster, compactness and the separation Silhouette value is consists both concepts of cohesion and separation. And also discussed the optimum cluster is represented by both F_c (U) and D_c (U) and choose F_c (U) > D_c (U), and FCM is also discussed in detail (Steinbach, M., et. al., 2005; Rousseau and Silhouettes, 1987) [95][96] , and online availability Silhouette is in (Clustering) [97].

$$S(m_1, m_2, \ldots\ldots\ldots, m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I\,(x_i \in C_k) \mid Dist_{i,k} \mid^2,$$

$$I(X) = \begin{cases} 1, & \forall\, x\ belong\ in\ Cluster \\ 0, & Otherwise\ \forall x\ not\ belong\ in\ Cluster \end{cases}$$

$$SSE_{SUM} = S(m_1, m_2, \ldots\ldots\ldots, m_M),$$

K-Means find locally optimal result w.r.to clustering error (variance) main drawback to sensitive initial position of cluster center.

Deal to initialization problem, the proposed algorithm iterative deterministic algorithm employ to as a searching based solution find closed to clustering variance.

For Min Max

$$S_{weight} = \sum_{i=1}^{N} (w_k)^p \ \sum_{k=1}^{M} I\ (x_i\ \in C_k)\ |\ Dist_{i,k}\ |^2\ ,$$

$$\sum_{k=1}^{M} w_k = 1\ 0r\ 0, 0 \leq p \leq 1\ ,$$

P is small to less weight value so difference between the variance

$$w_k = \frac{(v_k)^{\frac{1}{(1-p)}}}{\sum_{k}^{M}(v_k)^{\frac{1}{(1-p)}}}$$

$$\text{Where } v_k = \sum_{i=1}^{N} \ \sum_{k=1}^{M} I\ (x_i\ \in C_k)\ |\ Dist_{i,k}\ |^2$$

Cluster validity with fuzzy clustering technique is supported by similarities and dissimilarities define the pairwise. The silhouette index is generalized and applied on both crisp and fuzzy approaches (Rewashes and Ralescu, 2012) [98]. The detecting of separated clusters concept is reported in research paper proposed by Dunn's in (Dunn J. C., 1973) [99]. The objective function value is minimized, and show facilitates the new relationship same function, stopping algorithm and prove this relationship (Selim and Kamel, 1992) [100].

Ilc N. (2012), author proposed novel cluster validity, modification of Dunn's score and measure the shortest path. The quality of cluster is determined by silhouette; therefore silhouette $s_{(k)}$ is clearly defined as

$$s_{(k)} = [b_{(k)} - \ a_{(k))}]/\min\{a_{(k)}, b_{(k)}\}$$

Differences of the k^{th} object to other rest object in the own cluster, neighboring cluster

And range of silhouette is $-1 \leq sillhouette \geq 1,$

Case: 1- if silhouette close to one (1), cluster clearly distinguish (better quality of cluster) or good cluster,

Case: 2- if silhouette close to zero (0), then sample represent the overlying in cluster or not significant, and

Case: 3- if silhouette close to minus one (-1), then sample misclassified or create the cluster in wrong direction, and simplified form in [93].

2.7 Radial Basis Neural Network (RBFNN)

Radial Basis Neural Network (RBFNN) has consisted of three layers namely the input layer, output layer and hidden layer. Its neural net is strictly for limitation has a single hidden layer. RBFNN has two phases as: first phase has hidden layer trained using concept of the back- propagation, required the method of curve fitting, find out the variance (σ) and receptors. Second phase has updated the weight vectors between a single hidden and output layer. The first stage of training is done by a clustering algorithm and consists k cluster samples or observations into K clusters which is satisfy the condition k > K. Every receptor (i.e. output cluster) is determined variance using Euclidian distance for nearby for all cluster samples.

The author has illustrated that the modification of the error functions without affecting training speed (Bishop C., 1991) [116], and broadly application is used in universal approximation (Park and Sandlberg, 1993) [117].

The authors have discussed the representation are some learn from the training pictures and further then applied to classify the new picture in group the texture and also assessed their merits based on performance (Verma, and Zisserman, (2008) [114]. In research paper, it is mentioned that the some no. of factors of unseen nodes in this system more important to attain the high performance as compared to learning concepts (Coates and Lee, 2011) [115].

2.8 Relevant Findings of Literature Review

After a careful and centric study of available approaches for clustering unsupervised trend technology is found in mining areas. Here, some of the relevant finding enlists as follows:

- Lack of cluster quality due to early centroids of the created clusters.
- During analysis of clustering, arise many issues in the method of K-Means. like outliers, changing of the correctness, and empty cluster.
- They are reflecting the contrary effects when initializing the incorrect and include the lesser convergence, unfilled clusters, and dropping in wrong local minima.
- The early k-means method fatalities its success due to increasing dimensionality.
- The traditional clustering concepts are quite sensitive for an established to initial of cluster centers.
- It has not well implementation in ambiguous datasets, and large dimensional datasets.
- The classical cluster is not denser, therefore convergence is slowed down.
- Centroids of cluster may not be optimum, as the traditional technique can be converging to get local optimum results.
- The internal metrics of cluster quality for original k-means is high.
- In traditional clustering has some deficiency of sensitivity to first centers in cluster.

The available literature on clustering divided into few thoughts: in first categories of approaches which emphasized on framework during design, how to find the good clusters. In second categories of approaches are tries to refine the cluster metrics. The limitation of the second categories is their analysis of metrics is a more tedious task for estimating the relevant outcome. This approach is fundamentally used to design of cluster framework. This chapter discussed the concepts which are applied during design of framework. In addition, it is estimates the more refinement regarding development effort, metrics analysis. The theoretical aspects review of chronological order which specific categories like namely PCA, PSO, fuzzy, and ANN approach of RBFN play vital role in research fields of mining. The regress review process is to strengthen the conceptualization for refining the presentation of clusters.

2.9 Summary

This chapter of the literature review deals with extensive study and more popular clustering k-means algorithm. There are discussed the several changes and related concepts of k-means. The core objective of it process is to assemble a similar type of data point, which is more convenient in different services and utilities in many fields. Meanwhile, it has some own limitations on the data size, it is further redefined and to meet the future challenges nowadays applicable to large size of data.

In this chapter is existing the related literature surveys on k-means partitioning clustering algorithm, principal component analysis, Genetic Algorithm(GA), and Particle Swarm Optimization(PSO) is analyzing in the detail. The study related to literature suggested more solutions to its limitations and use several methodologies to apply on the k-means. The proposed work is the strategy of a well-organized framework to develop its performance of group or cluster in mining using PCA, PSO, GA, and ANN to deal with the dataset facts, and k-means performance.

CHAPTER-3

DESIGN A FRAMEWORK

Design an efficient framework for produced good quality of clusters. Evaluate metrics related to performance with the component of clusters. In chapter- 3, we discussed the every component of the framework. The component of the proposed framework is consisting as traditional k-means, proposed method, proposed concept hybridized via PCA and via PSO, statistical analysis, using datasets, GA, and Radial Basis Function of Neural Network (RBFNN) of ANN theory and mentioned guideline of research methodology.

3.1 Introduction

K-Means clustering methodology is an “unsupervised clustering” algorithm that splits input data into many classes based on their similarity of criteria. By minimizing the number of square distances between the data points and then the classification is achieved. Distance measures and similarity measures are two main types of measures used. To evaluate the similarity of relationship or difference of a pair of objects are used in distance measurements. The data is only clustered as a crisp array and has its drawbacks when handling high-dimensional data and minimal data since it is very simplest clustering tool. In the real world, with the advance of information technology, the bulks of data processed by many applications exceed the Peta-scale threshold, so the clustering of very large-scale data is now a daunting challenge. There are mainly five useful approaches to increase the efficacy of the basic k-means algorithm of clustering. It has proposed to work in this research to deal with high-dimensional data, constraints, and data processing using genetic and neural networks.

3.2 Proposed a Framework

We have design an efficient framework for created good quality of clusters. Measures metrics related to the performance with the component of clusters. In this chapter every related component of the framework is discussed.

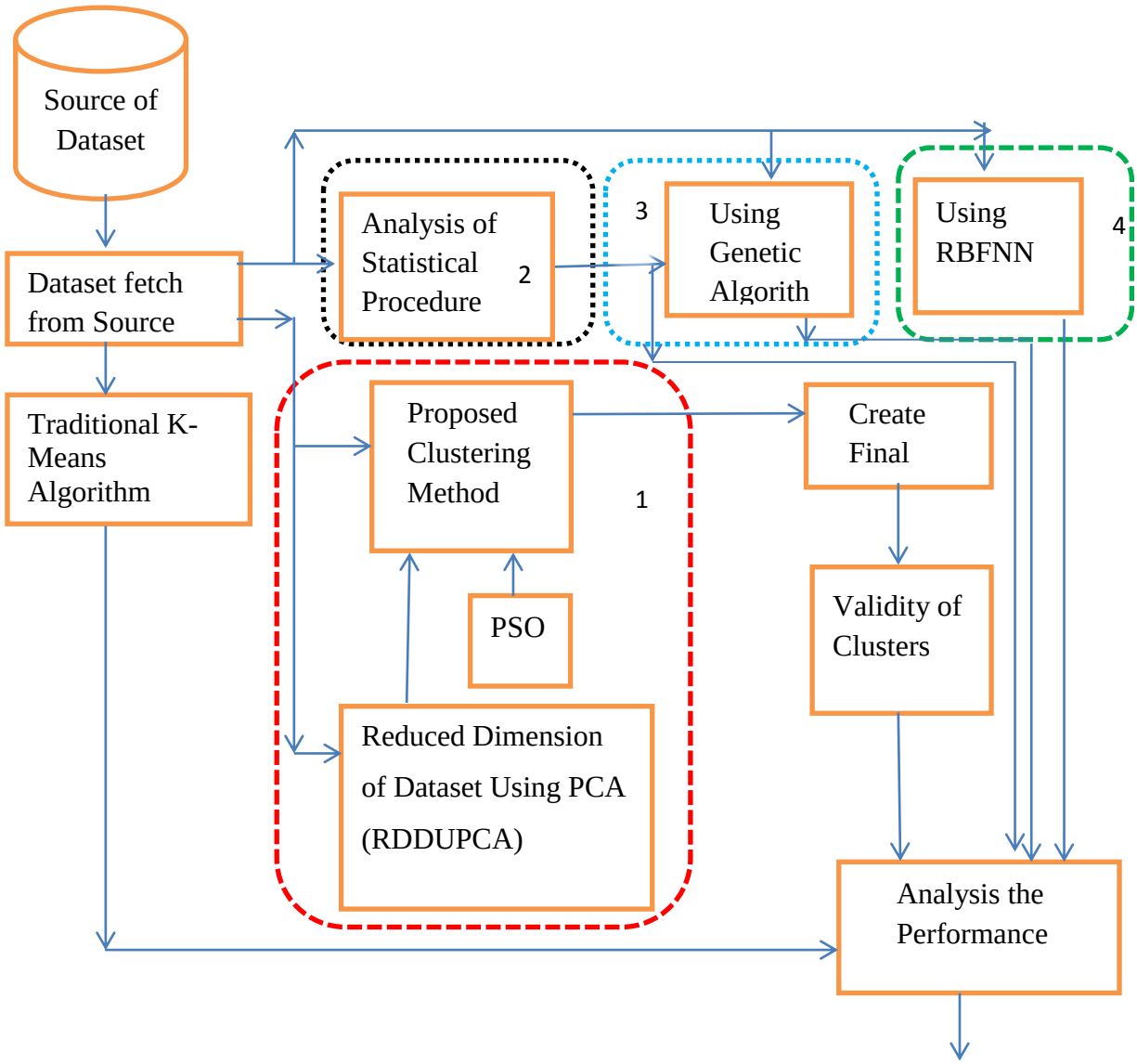


Figure: 3.1 Design an efficient framework for clustering using on large dataset

There are key components of framework as follows:

- First component: Proposed clustering method, Reduced dimension of dataset using PCA, and PSO
- Second component: Analysis of statistical procedure by software tools
- Third component : Applied the genetic algorithm of optimizing tool
- Fourth component: RBFNN of ANN

In figure-3.1, the framework comprises like Particle Swarm Optimization (PSO) and Radial Basis Neural Network (RBFNN) are discussed details in chapter-2.

3.3 Source of Dataset

It is performs a more important role for extracting the features of data for the given dataset, and also finds the information. The factual datasets like iris, wine, and heart disease are taken from the path address http:// https://archive.ics.uci.edu/ml/datasets of UCIML learning repository. The fetch the dataset as like:

Heart disease dataset path address: http:// https://archive.ics.uci.edu/ml/datasets/heart

User Knowledge dataset path address: http:// https://archive.ics.uci.edu/ml/datasets/user

Iris dataset path address: http:// https://archive.ics.uci.edu/ml/datasets/iris

Wine dataset path address: http:// https://archive.ics.uci.edu/ml/datasets/wine

3.4 Traditional K-Means Algorithm

It is a partitioning clustering method that relocates objects by moving from one to the next cluster and beginning from a first partitioning. In clustering, dataset is divide N^{th} observations into k^{th} clusters, with each observation in cluster with nearby the mean. It is unsupervised method for dealing with the recognized difficulty of clustering. This is evolutionary algorithm that gets its name from way of the works.

It is divided N^{th}observations into k^{th} groups, with k being the input constraint. Furthermore, it assigns every observation of clusters depending on how close it is to the cluster's mean. Its significance is reevaluated, and this process is restarts again.

Algorithm steps:

Step 1: Select the K data point at random as the centers of early cluster.

Step 2: Each point of dataset is assigned in nearby cluster, according to the space between points and each cluster center.

Step 3: Every cluster center is recalculated as mean of the cluster's points.

Step 4: Steps 2–3 are repeated until all clusters have converged.

There is liable on the implementation, convergence can be meaning in a different ways. When steps 2 and 3 are there repeated, however, it usually signifies that no opinions change the clusters or that the changes do not create a major effect on the cluster meaning.

K-Means is a well-known technique in "unsupervised learning" and "vector quantization". This is formulated by minimizing a formal objective function,

$$\text{Mean} - \text{squared} - \text{error distortion\_ minimum (P)} = \sum ||x_i - C_j||^2$$

Where

N= Illustrate the dataset sample,

K= No. of clusters, and

d=dimension of data vector,

$X = \{x_1, x_2, \ldots\ldots\ldots\ldots, x_n\}$, symbolized, N data sample set,

$P = \{x_i \mid \ where, i\ is\ 1\ to\ N\}$, symbolized label of X, and

$C = \{\ c_j | where, j\ is\ 1\ to\ k\}$, is k centroids of cluster

Because of its simplicity for implementation, the conventional k-means can be useful to a given clustering algorithm as a post processing stage to improve the final solution. However, the main challenge of the original k-means is that its categorization of highly performance be sure of on the selected initial partition. It is converges to a close by the best result with the randomized of first partitions.

3.5 Reduced Dimension of Dataset Using PCA (RDDUPCA)

RDDUPCA is the most influential concept of dimension reduction technique from a prearranged dataset and then obtain change variable as well as change the dimension also. Extraction of some component from a considered dataset is to effect the separation of clusters. Consequently, obtain the reduced dimension of data sets is an implementation on proposed method to achieve the best clustering. The advantages of this technique namely compress the dataset, minimized the storage capacity, minimized the computation time, eliminate the redundancy, and also improve the performance. PCA technique is working on concept without loss the data.

3.6 Analysis of Statistical Procedure

Statistical analysis procedure of different datasets and creates the clusters quality is initially not well. Here, we have discussed the PCA concept for being used to reduction of n^{th}dimensional datasets. Analyzes the dataset it is constructing clusters and then reduced dimensions. Therefore, creates good quality of the clusters are followed ideas represented by analysis of reduction component, comparative analysis of F-Ratio, comparative analysis of average Silhouette , and comparative analysis of centroids respectively. Statistical Package for the Social Sciences (SPSS) and NCSS Software package tools are being used to study and analysis of the datasets. The software tool is useful to analyze of the clusters technique. And also it is being used simplified and handles the large dimensional datasets.

3.7 Validity of Clusters

There is assessment of the "performance" of a clustering process, then, has many aspects. Here, calculation of the domain area of the dataset is itself procedure for the clustering. Validation is done used a statistical process and checking hypotheses, when we applied the approaches of the statistics in clustering. Validation studies are separated into 3(three) groups as: (1) Improved of the structure (data) is matched to a prioritized structure in an external validity evaluation, (2) Internal validity testing is suitable for data, (3) When two(02) datasets are compared w.r.t. define their merit.

There are 02 (two) basic quality metrics to evaluate whether the clustering is good or bad. These are the internal and external metrics. This evaluation is very much important for efficient **Quality of** clustering.

3.7.1 Internal Metrics

Internal cluster quality (metrics) measures use similarity to determine how compact the clusters are there. It frequently assesses intra-cluster homogeneity, inter-cluster separability, or a mix of the two. It is not use some external data. There are many internal cluster quality metrics like, Sum of Squared Error (SSE), least criteria of the variance, Scatter Criteria. In this research work we are used internal metrics such as Sum of Squared Error (SSE), execution time, and Silhouette is used to evaluate the quality of cluster.

- **Sum of Squared Error (SSE)**

SSE is the very simplest and widely used criterion measure for clustering. It is calculated follow as:

$$SSE = min \sum_{j=1}^{K} \sum_{i=1}^{n} | x_i^j - c_j |^2 \qquad (3.1)$$

Where,

c_j= centroids for j^{th} cluster,

K= number of cluster,

n= no. of objects, and

x_i^j= i^{th} object in k^{th} cluster

- **Intra-Cluster Distance (ICD)**

ICD defines the distance between two the objects belonging to same cluster. Here mentioned the mathematical expression for centroids diameter distance as:

$$ICD(C) = 2\frac{\sum_{x \in C} d(x, \mu)}{|C|}$$

Where

$x =$ represent the object in same cluster

C= represent the cluster

$x \in C$= every object x belong in the cluster C

- **Execution Time**

Measure the execution time (taken by CPU) during processing of datasets is defined as follow:

Processing Time= P*CPI/ CR

Where, CR=Clock rate

CPI= Count clock cycle * no. of used instruction/ Total number of instruction

CPI= Cycle per Instruction

P= count instruction

3.7.2 External Metrics

The external metrics are very beneficial for evaluating whether the building of the clusters tie to few predefined grouping of the instances. The following external metrics as rand index (or Accuracy) of the cluster, adjusted random index, Normalized Mutual Information (NMI), Precision (P),Sensitivity or Recall, Entropy (E), F-Measure or F-Score (F).

But, in thesis we are used external metrics such as Rand Index or Accuracy of the cluster Precision (P), Sensitivity or Recall, F-measure.

In this paper, mentioned the evaluation metrics of cluster Validity based on four criteria authors define in [12], as following:

Precision: Measure how many object classified in correctly

$$\text{Precision} = \frac{a}{a+b}, \quad (3.2)$$

Recall: Measure the fraction of exactly classified

$$\text{Recall} = \frac{a}{a+d}, \quad (3.3)$$

F-measure: Measure the score

$$\text{F-measure} = \frac{2\ \text{precision\_recall}}{(\text{precision}+\text{recall})}, \text{ and,} \quad (3.4)$$

Rand Index: Accuracy of the cluster

$$\text{Percentage of Rand Index} = \frac{a+c}{(a+b+c+d)} * 100, \quad (3.5)$$

Where,

a=TP=True Positive,

b=FP= False Positive,

c=TN=True Negative, and

d=FN=False Negative.

3.8 Problem Statement

There are extract the hidden data patterns from the datasets using k-means algorithm is an unsupervised mining technique. The limitations of current k-means clustering are considered some limitations for study. The expletive of dimensionality is the first restriction of current K-Means clusters. Multiple dimensions are more challenging to reflect about, hard to imagine, and this is impossible to enumerate because of the exponential increase amount of potential values with each dimension. The k-means clustering method does not handle the large dimensional datasets.

By proposed method to handle it limitation and increase the efficiency and accuracy of high-dimensional data clustering, in this thesis, data processed and reduction of dimension by PCA method and also used PSO for fitness of clustering.

The second weakness of current K-Means clustering is its approach to crisp clustering. However, certain poor types of side data about the domain or data sets can also be accessible or derivable in actual application domains. Difficulty by k-means clustering techniques is that even if this is usable, they do not benefit from side data. To enhance the current clustering of k-means with side data and GA creates the automatic clustering for generated of Gene.

The traditional clustering idea working on considered that the original centers are given. It is starting from its initial centers till search for final clusters. The algorithm will yield a weak group, final centroid without appropriate initialization, and this issue can become severe if the data is clustered by traditional K-Means. The Artificial Neural Network (ANN) method is used in pattern training and gets mean square error to overall optimum in current k-means technique. In this thesis, we are using sampling the input dataset for it

algorithm. This is an effort made to enhance the learning skills of a multilayered ANN, reduce the quantity of time interval, and also resources preferred by the learning procedure.

3.9 Proposed Method

There are three main shortcomings of the existing K-Means are considered and overcome by introducing new methodologies. There are three disadvantages considered in this work the lack of large-dimensional data handling, the deficiency of the restriction control, and the effort of capable data clustering.

Objective function

$$f(x) = \sum_{j=1}^{K} \sum_{i=1}^{N} || \ Dist_{i,j} \ ||^2$$

$$|| \ Dist_{i,j} \ ||^2 = \left|| \ x_{i,j} \ - \ C_j \right|^2$$

Data object reassigned nearby cluster $minimum(\{ \ x_{i,j} \ - \ C_j \})$

$$C = \frac{1}{N} \sum_{x_{i \in C_i}} x_i \ , where \ i = 1 \ to \ K$$

We have some modified for the objective function like that with apply the weight$W_{[d,k]}$

- Applied the weight in two dimensional dataset

$$W_{[s,k]} = \begin{cases} 1 & x_i \in cluster \ j \\ 0 & x_i not \in cluster \ j \end{cases}$$

- Modified objective function

$$F_C = \sum_{j=1}^{K} \sum_{i=1}^{N} (W_{[s,k]}) \sum_{N=1}^{KXI} ||Dist_{i,j} \ ||^2$$

$$|| \ Dist_{i,j} \ ||^2 = \left|| \ (x_{i,N} \ - \ C_{j,N} \right|^2$$

- Measure the centroid

$$C_{j,N} = \frac{\sum_{i=1}^{S} w_{i,j} x_{i,N}}{\sum_{i=1}^{S} w_{i,j}},$$

Where

i= 1 to N

j= 1 to K

K= No. of cluster

N = No. of data sample

I= No. of attributes

N= K x I where d is belong in data sample 2D matrix K x I

Algorithm steps as follows:

Step 1: Set initial prototype cluster k\\ select the cluster k initialized initial Centroid

Step 2: All points of dataset is assigned in nearby cluster, according to the Objective function and each cluster center$C_{j,N}$.

Step 3**:** All cluster center is recalculated as mean of the center mean of clusters

$$C_{j,N} = \frac{\sum_{i=1}^{S} w_{i,j} x_{i,N}}{\sum_{i=1}^{S} w_{i,j}},$$

Step 4: Reassigned every object into the cluster based on centroids

Step 5: Evaluate the mean squared error by using this function

\\ function $J_c\,(p)$ using Euclidean Distance

$F_C = \sum_{J=1} \sum_{I=1} (W_{[s,k]})\ \sum_N ||Dist_{i,j}\ ||^2$

Step 6: Steps 2–5 are repeated until Until $|J_C(p) - J_c(p-1)| < \delta$ (thresold value) all clusters have converged

Step 7: Create clusters

The element of the two dimensional matrix expressed as [d, k], where s is belong in data sample N, and k is belong in cluster K. Measure the centroid from 2D matric [k, d] where d is belong in data sample K x I, and k is belong in cluster K. Finally measure the Cluster $minimum_{k \in K}(\{||\ x_j\ -\ \ C_k||\})$.

3.10 Datasets Used

Datasets used in this work are taken from the UCIMLR (Machine_ Learning_ Repository) The learning recourse has database, philosophy of its field, and makes data applied in ML (Machine learning) community for experimental study of ML algorithms.

In this effort, iris dataset, wine dataset, heart disease dataset, user knowledge modeling dataset are used for the experiment.

3.10.1 Iris Dataset

This is the multivariate dataset, which is widely used in the area of pattern recognition. The dataset resides information about varieties of iris flowers. This dataset is created by R. A. Fisher is a well-known dataset for the classification of various data clustering techniques.

It includes 3(three) classes of 50(fifty) instances each and each class refers to a category of Iris flower. The instances no. is 150 (one hundred and fifty), and the no. of attributes 4(four) respectively. There is in a list the attribute all in cm scale like as

1. Sepal_ length,
2. Sepal_ width,
3. Petal_ length,
4. Petal_ width,
5. Class: Iris virginica, Iris versiclor, and Iris setosa

3.10.2 Wine Dataset

There are 30(thirty) variables but in a list the 13(thirteen) continuous attribute using for clustering. It contains distribution of 59 (fifty nine) in class 1(one), 71(seventy one) in class 2(two) and 48(forty eight) in class 3(three). The numbers of instances are 178(one hundred and seventy-eight), and the attributes number are13 (thirteen).There is in list the attribute like as:

1. Alcohal,
2. Malic_ acid,
3. Ash,

4. Alkalinity_ of ash,
5. Magnesium,
6. Total_ phenols,
7. Flavanoids,
8. Nonflavanoid phenols,
9. Proanthocyanins,
10. Color_ intensity,
11. Hue,
12. OD280/od315 of diluted _wines,
13. Proline

No. of instances 178, disappeared the attribute values none.

3.10.3 User knowledge modeling Dataset

It has 403(four hundred three) num. of instances but available 258(two hundred fifty eight) on achieve, number of attribute 5(five), attribute characteristics like integer and characteristics of dataset is multivariate

There are in list the attribute like as: study time of the goal and the repetitions of user for goal object material, study time related for use associated the object and exam performance of user associated object using the target object, performance for exam of user with target object and last attribute knowledge level of the user is categorized in very low, middle, high, very low, and very high. The abbreviation of attributes are namely as STG, SCG, STR, PEG, and UNS.

3.10.4 Heart Disease Dataset

There are 30 (thirty) variables but in list the 14 (fourteen) continuous attribute using for clustering. The dataset has a distribution of 59(fifty nine) in class 1(one), 71(seventy one) in class 2(two), and 48(forty eight) in class 3(three). Numbers of instances are 297(two hundred and ninety seven) and the no. of attributes are13 (thirteen). There are in list the attribute like as:

1. Age in years,
2. Set 1(male) 0 (female),
3. Chest_pain_type
4. Blood_pressure__in_mm_Hg_on_admission_to_the_hospital,
5. Serum _cholesterol in (mg dl),
6. Fbs_fasting_blood_sugar__gt__120_mg_dl, set 1(true), and set 0(false),
7. Resting_ electrocardiographic_ results,
8. Maximum_heart_rate_achieved
9. Exercise_induced_angina__1___yes__0___no,
10. ST depression_induced_by_exercise_relative_to_rest,The_slope_of_the_peak_ exercise _ST Segment
11. Number_of_major_vessels__0_3__colored_by_flourosopy,
12. X3___normal__6___fixed_defect__7___reversable_defect,
13. Target X 1_or_0

3.11 Guideline of Research Methodology

We expressed the guideline of the proposed efficient framework given below.

The pseudo code of research methodology discussed follows as:

```
Function_ metrics (arg_1, arg_2,…. )  // measured the performance
{
    \\ Proposed AEIKM (An Efficient Improved K-Means) Algorithm
    Step 1: Call Function_M1 (arg_1, arg_2…)
\\ PCAH AEIKM (AEIKM using hybridized via PCA)
    Step 2: Call function_M2 (arg_1, arg_2…)
\\ PSOHAEIKM (PSO Hybridized via AEIKM)
    Step 3: Call function_M3 (arg_1, arg_2…)
\\ Optimizing Technique of GA
    Step 4:  Call Finness_function (arg_1, arg_2…)
```

\\ Create the training with target using RBFN Kernel ( )

Step 4: Call RBFN Kernel ( )

}

Function_M1 ( )

{

Step 1: Set initial prototype cluster k\\ select the cluster k initialized initial Centroid

Step 2: Evaluate the distance by Euclidean distance\\ $Dist(X_i, C_j) = |X_i, C_j|$

Step 3: Select the min distance

Step 4: Reassigned every object into the cluster based on centroids

Step 5: Evaluate the mean squared error by using this function

\\ function $J_c\,(p)$

Step 6: Determine thresold value \\ Until $|J_C(p) - J_c(p-1)| < \delta$

Step 7: Create clusters

}

\\ AEIKM using hybridized via PCA

Function_M2 ( )

{

Step 1: Mean of considered dataset

Step 2: Determined value, subtract means from dataset

Step 3: Set Z score

Step 4: Find covariance matrix, and also find the eigenvalues

Step 5: Find new dataset

Step 6: Function call of proposed method

}

\\ PSO Hybridized via AEIKM

Function_M3 ( )

{

Step 1: set all parameters ω, L1, and L2, location X_0 and particle velocityV_0

` Step 2: Find and best index for all particles by fitness function

Step 3: Update position and velocity

Step 4: Update position, both best and global

Step 5: return best

}

\\ Optimizing Technique of GA

\\Function kfdaff= kernel for vectors(x1, x2)

Finness_function (arg_1, arg_2…)

{ $b = Dist(X_i, C_j)$\\ Defined in k-means concept

Function y1 = distance fitness (b)

y = abs (b);

End

Function y2 = kfda_fitness (x)

\\ Define the function of kernel fisher's Discriminant analysis

Function_KFDA ( )

End

}

Where,

C_j = centroids for j cluster,

K= number of cluster= number of object, and

$x_i = x_i^j, i^{th}\ object\ in\ the\ k^{th}\ cluster$

Define the kernel,

$$K(x_i, x_j) = \emptyset(x_{1i})^T \emptyset(x_{2j}) = (x_{i1}.\text{^}2 x_{2j}.\text{^}2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}.\text{^}2\ x_{j2}.\text{^}2)$$

Guideline for ANN:

There are describe the idea of RBFN

\\ MIN-MAX SCALING INTO [0, 1]

Set Function of Normalization ( )

\\ z-score make the conversion of data center is zero scale into range [-1, 1]

F (T) = function of normalization (Dataset X, D)

\\ Create the kernel function

Kernel (arg_1, arg_2)

{

Create the kernel function f (k) = Kernel (data matrix, center)

\\ Exponential function of Kernel

K (I, J) = EXP (-NORM (Dataset X, Center of dataset X))

}

\\ Create the training with target using RBFN Kernel ( )

RBFN Kernel ( )

{

Step 1: Set the all attributes \\ Associated with classes and subject CREATE Target, lambda, id class

Step 2: Call K-Means (Dataset X, Means for K, 1)

Step 3: Call Kernel (Dataset X, Consists the center)

Step 4: Return the result

Step 5: Create network structure // Measure performance

}

3.12 Software / TOOL

MATLAB is a tool used in the mathematical calculation and also applicable in create the graph in research. The data element in matrix form necessitates in use of array-based data

manipulation software. It's a HLL (High Level Language) and more interactive platform for numerical calculation, picturing, and program design. MATLAB is applied to examine the set of data, develop algorithms, and create simulations and its uses. In MATLAB tools built-in the methodical functions are enable to explore for the multiple approaches, and it also useful this functions working using in creation of matrices.
Both SPSS and NCSS Software tools are enable to study of datasets. The software tool is useful to analyze of the clusters technique. It is simplified and handles the large datasets.

3.13 Summary

The k-mean method is most popular clustering concept to make a data cluster. But the original k-means method has some limitations which degrades performance of it algorithm. The some limitations of original k-means are identified and new methodologies are proposed to rectify the restrictions.

The first limitation is the difficulty in behavior and metrics of large dimensional data, and it is handled by leading with principal component analysis concept in proposed approach, and also evaluates the fitness of cluster by hybridized via PSO. The second proposed methodology is to include the analysis by statistical tool of the centroids, F-score, silhouette for significant clusters. The analysis criteria of applied the objective function is lesser than an objective function of K-Means. The exit criteria of selection is produced the maximum population value when used the no. of generation.

Finally, work on proposed an idea which uses the RBFNN recognizing for neural network to find trained dataset sum squared errors. Four datasets namely Iris, Wine, Heart disease, and User knowledge modeling taken from UCI machine repository are considered for experiment by proposed research methodologies. In thesis, outcome is validated by MATLAB (Matrix Laboratory) Ra 2013a software tool.

CHAPTER-4

IMPLEMENT OF CLUSTER PERFORMANCE

Chapter-4, the limitation of clustering k-means technique is more difficulty in behavior of large dimensional data, such kind of problem handled by leading concept using Principal Component Analysis (PCA) concept, and also used particle swarm optimization concept to measure the fitness value for objective function , it is also used in proposed approach.

4.1 Introduction

Clustering concept is a mechanism of grouping a dataset object on the basis of measure the similarity. The partitioning is created by an algorithm of cluster. Therefore clustering is more beneficial and more effective for discovering the group structure in giving data sets. The principal component analysis method is mainly preferred to minimize the dimension of the dataset. This algorithm works like linear algorithm. This method is required to lessen the dimensions of the large dimensional data sets. This technique successfully eliminates a feature of characteristic attribute to help the minimization of dimension, but ensures that main attribute is not to going misplaced.

Cluster technique performs more important role for extracting the features of data for given the dataset, and also finds the information. The factual dataset like iris, wine and heart disease are taken from the UCI Irving Machine learning repository. The k-means technique is more popular for partitioning of datasets into clusters but it has some limitations. This method does not performed well on large dataset therefore, proposed a concept of k-means with some modification to create a better cluster. And also, k-means concept used to hybridize with popular algorithm principle component analysis to overcome its limitations. We have proposed research methodology to enrich the cluster performance and implemented on various sizes of the datasets. In this chapter,

we have implemented an experiment on MATLAB R2013a to measure the metrics of the cluster and also measure the fitness of fitness function values use particle swarm optimization with proposed method.

In this chapter we discuss and try the proposed concept to combine with PCA algorithm to resolve the drawback of K-Means. The PCA techniques utilized on numerical attributes on the database the noisy feature reduces the dimension of problem in vector but improve the fitness accuracy, and also PSO concept to apply for fitness.

4.2 Back Ground

4.2.1 K-Means Algorithm

This algorithm is to apply on a given d-Dimensional dataset $X = x_{1,}x_2, x_3, \dots\dots\dots, x_M$, $\forall\, i, and\ i = 1\ to\ M$, and set $\{x_i | i = 1,2,3, \dots\dots\dots, M\}$, $x_i \in F^{dimension}$, x_i denoted the i^{th} point of data from the dataset as well as $F^{dimension}$, illustrate in d-dimensional feature space. K-means algorithm, authors explained details in (Anusuya and Lattha, 2011; Pena J.M., et. al.1999; Celebi et. al., 2013) [19, 29, and 30].

4.2.2 Principal Components Analysis (PCA)

This is the most influential concept of dimension reduction technique from a prearranged dataset and then obtain new variable and change the new dimension also. Extraction of some component from a considered dataset is to effect the separation of clusters. Consequently, obtain the reduced dimension of data sets is an implementation on k-means algorithm to achieve the best clustering. Consider the n vectors data points like such as$x_1, x_2, \dots\dots\dots, x_n$. This technique is can only extract a linear projection of such $x_i \forall i$data points, and described concepts (Anusuya and Lattha, 2011; Sethi, and Mishra, 2013) [19, 45].

4.2.3 Particle Swarm Optimization (PSO)

This concept is encouraged by public and supportive performance to show by different aspect to fulfill their required search space move the particle and find best search value. This algorithm is an optimization concept of the population based on stochastic global

optimization, and its applications. The PSO algorithm useful to dynamically accelerate the constraint is CPSO abbreviate form improve the convergence rate and trend its rate with optimized period. To initialized the position (X_0) and velocity (V_0) according using to set the boundary condition is upper and lower limit ([Ratnaweera, A., et.al. 2004); Poli, R., et.al. 2007)) [37, 39]. There is the boundary condition define as follows

$$X_0 = \frac{rand()}{2}[(a - b) + (a + b)] \quad (4.1)$$

$$V_0 = \frac{rand()}{2}[(v_a - v_b) + (v_a - v_b)] \quad (4.2)$$

Where

a =upper limit of particle, b= lower limit of particle, V_a = upper limit of velocity, & V_b = lower limit of velocity

The standard Particle Swarm optimization method according to the equation (4.1) and (4.2) update the velocity $v_i\ (t+1)$ and position $x_i(t+1)$ respectively. The parameter used as namely ω = weight of inertia, L_1 and L_2 learning parameter , $rand_1()$ and $rand_2()$ are generated the random number between 0 (zero) and 1(one), p symbolized the finest solution, and g symbolized the globally best solution. Authors explained the PSO concept in detail (Sethi and Mishra, 2013; Shi-Wei and Xiao-Dong, 2010; Poli, R., et. al., 2007)) [39, 44, and 45].

4.3 Research Methodology

4.3.1 Real World Datasets
4.3.2 Proposed Methodology

4.3.1 Real World datasets

Heart disease dataset, User knowledge modeling dataset, Iris dataset, and wine dataset are generally available on the UC Irving Machine Learning Repository (UCIMLR) archive and its datasets are used in research works. These factual world datasets retrieve from the UCI learning repository the path address is mentioned here as https://archive.ics.uci.edu.ml/datasets. Heart disease dataset comprises the no. of instances 297, no. of attributes 14, and multivariate characteristics of data set, User knowledge modeling dataset has 403 no. of instances but available 258 on achieve, and

no. of attribute 5, attribute characteristics integer and characteristics of dataset is multivariate. Iris dataset comprises the no. of instances 150, no. of attributes 4, and multivariate type of data characteristics, and wine dataset comprises the no. of instances 178, no. of attributes 13, and multivariate type of data characteristics.

4.3.2 Proposed Methodology

4.3.2.1 AEIKM Algorithm \\ Proposed An Efficient Improved K-Means Algorithm

AEIKM algorithm follows as several steps

Algorithm steps as follows:

Step 1: Set initial prototype cluster k\\ select the cluster k initialized initial Centroid

Step 2: All points of dataset is assigned in nearby cluster, according to the Objective function and each cluster center$C_{j,N}$.

Step 3**:** All cluster center is recalculated as mean of the center mean of clusters

$$C_{j,N} = \frac{\sum_{i=1}^{S} w_{i,j} x_{i,N}}{\sum_{i=1}^{S} w_{i,j}},$$

Step 4: Reassigned every object into the cluster based on centroids

Step 5: Evaluate the mean squared error by using this function

\\ function $J_c\,(p)$ using Euclidean Distance

$$F_C = \sum_{J=1} \sum_{I=1} (W_{[s,k]})\ \sum_N ||Dist_{i,j}||^2$$

Step 6: Steps 2–5 are repeated until Until $|J_C(p) - J_c(p-1)| < \delta$ (thresold value) all clusters have converged

Step 7: Create clusters

Where, C_j= centroids for j^{th} cluster, , $C_{j,N}$= center for j^{th}cluster N observations

K= number of cluster,

N= number of observations, and

$X_i = x_i^j$= i^{th} object in k^{th} cluster

δ= threshold value

4.3.2.2 PCAHAEIKM Algorithm \\ AEIKM using hybridized via PCA

PCAHAEIKM Algorithm follows as several steps

1. Find mean of data set X,

$$\mu = 1/N(\sum_{i=1}^{N} x_i)$$

2. Find subtract of mean from dataset,

$$\tilde{x} = \sum_{i=1}^{N}(x_i - \mu)$$

3. Set z-score, $z = \frac{\check{x}}{\sigma}$

4. Find covariance matrix. $C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T$

5. Find the eigenvalues $|\mathbf{C} - \boldsymbol{\lambda}\mathbf{I}| = \mathbf{0}$, $\lambda_1 > \lambda_2$ (eigenvalue)

6. Set find the eigenvector $CX = \lambda X$, X= eigenvector

7. Reduced and return new dataset

8. call **AEIKM ( )**

9. Return updated

Where,

σ = Standard deviation,

λ = Eigenvalue,

X= Eigenvector,

C= Covariance matrix

4.3.2.3 PSOHPAEIKM\\ PSO Hybridized via AEIKM

PSOHPAEIKM Algorithm follows as several steps

1. to initialize parameter ω, L1, and L2.

2. To initialize population, particle at location X_0 and particle velocityV_0.

3. Set k=1.

4. Find fitness of particle for all, find index of BEST particle.

5. Set the boundary of particles.

6. To update X (position) and V (velocity) of particle.

7. Measure fitness at k++, and find index of best particle.

8. Update, (best + Global) position of particle.
9. Set k++, Goto (6) otherwise Goto exit from loop

4.4 Experimental Results

Chapter-4 is focused on a proposed algorithm AEIKM method, proposed method hybridized via PCA, and it compare by PCAHAEIKM method. The computing results of PSOHPAEIKM method (PSO hybridized via AEIKM method) is better than AEIKM method and creates good clusters.

4.4.1 Sum of Squared Error (SSE)

In Table-4.1, illustrate the comparative analysis of Sum of Squared Error (SSE) of four datasets creates the clusters follow as:

Table 4.1 Analysis of SSE metrics of clusters

Datasets	Methods	2	3	4	5	6	7	8
Dataset D1	AEIKM	26.90	19.89	15.75	12.35	10.87	8.38	7.4
	PCAHAEIKM	25.64	17.79	12.71	10.398	9.25	8.45	6.82
Dataset D2	AEIKM	30.59	28.69	15.37	14.76	10.71	8.2	7.52
	PCAHAEIKM	28.623	25.004	14.72	12.96	9.19	7.157	6.84
Dataset D3	AEIKM	13.343	10.768	10.78	10.304	9.3182	6.592	3.592
	PCAHAEIKM	12.026	9.504	9.45	9.349	8.7653	4.506	2.755
Dataset D4	AEIKM	17.52	15.35	15.02	12.03	11.025	9.76	8.34
	PCAHAEIKM	16.63	14.04	13.95	11.42	11.004	8.78	7.89

Heart Disease Dataset (D1)

In sequence to estimate the performance, there are applied AEIKM method and PCAHAEIKM method on D1 (Heart Disease Dataset). Estimated values are Sum of Squared Error (SSE), In Figuere-4.1(a), shown the SSE value at K= 2, 3, 4, 5, 6, 7, 8 number of clusters, and show the SSE value of PCAHAEIKM method is gradually decrease from k=2 to k=8, but at k=8 approx. closed of AEIKM method**.**

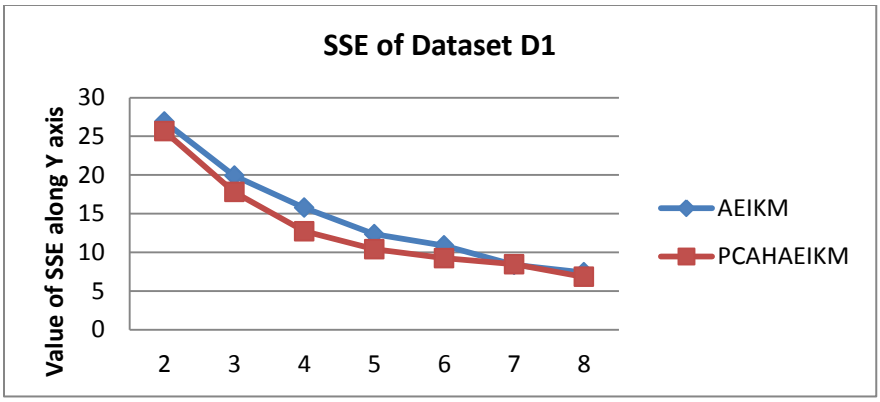


Figure 4.1(a): Analysis cluster vs. value of SSE

User Knowledge Modeling Dataset (D2)

In sequence to estimate the performance, there are applied AEIKM method and PCAHAEIKM method on User Knowledge Modeling Dataset (D2).Estimated values are Sum of Squared Error (SSE), shown in figure at K= 02, 03, 04, 05, 06, 07, and 08 clusters. In Figure-4.1(b) show the sample data of User Knowledge Modeling Dataset (D2), and show the SSE value of PCAHAEIKM method is gradually decrease from k=2 to k=8 than of AEIKM method**.**

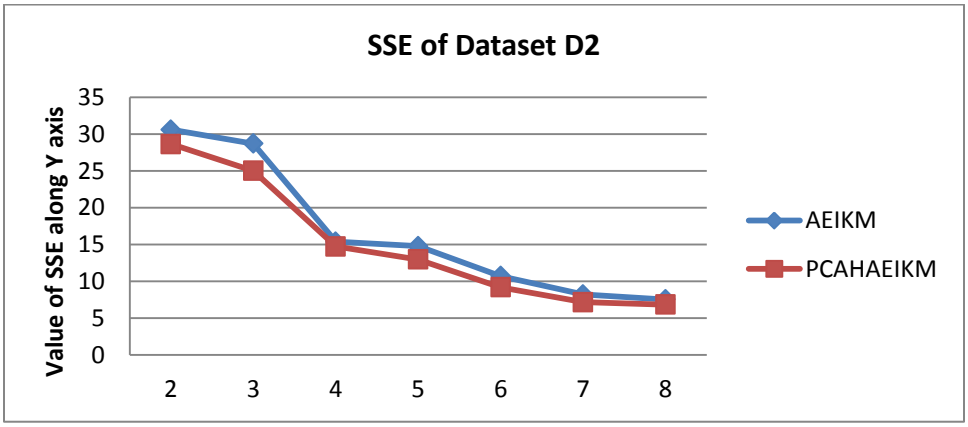


Figure 4.1(b): Analysis of cluster vs. value of SSE

Iris Dataset (D3)

In sequence to estimate the performance, there are applied AEIKM method and PCAHAEIKM method on Iris Dataset (D3).Estimated values are Sum of Squared Error (SSE), shown in figure at K= 02, 03, 04, 05, 06, 07, and 08 clusters. In Figure-4.1(c)

show the sample data of Iris Dataset, and show the SSE value of PCAHAEIKM method is gradually decrease from k=2 to k=8, but at k=6 approx. closed of AEIKM method**.**

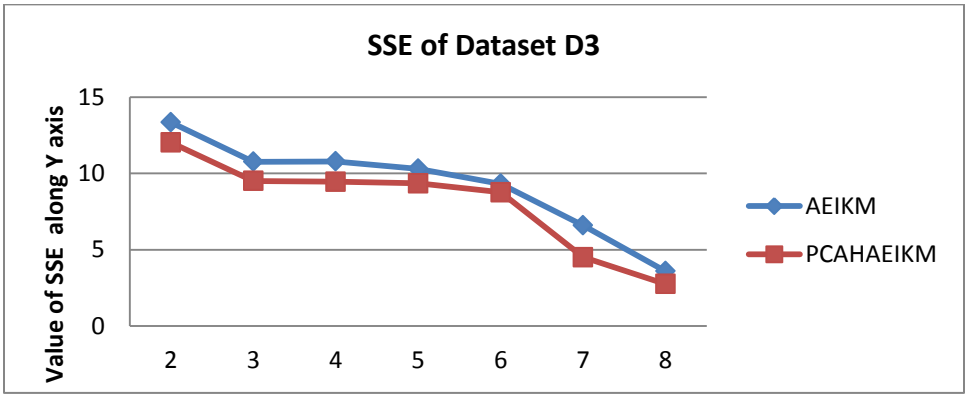


Figure 4.1 (c): Analysis cluster vs. value of SSE

Wine Dataset (D4)

In sequence to estimate the performance, there are applied AEIKM method and PCAHAEIKM method on wine Dataset (D4). Estimated values are Sum of Squared Error (SSE), shown at K= 02, 03, 04, 05, 06, 07, and 08 clusters. In figure-4.1(d), show analysis of cluster vs. SSE of Wine Dataset, and SSE value of PCAHAEIKM method is gradually decrease from k=2 to k=8, but at k=6 approx. closed of AEIKM method**.**

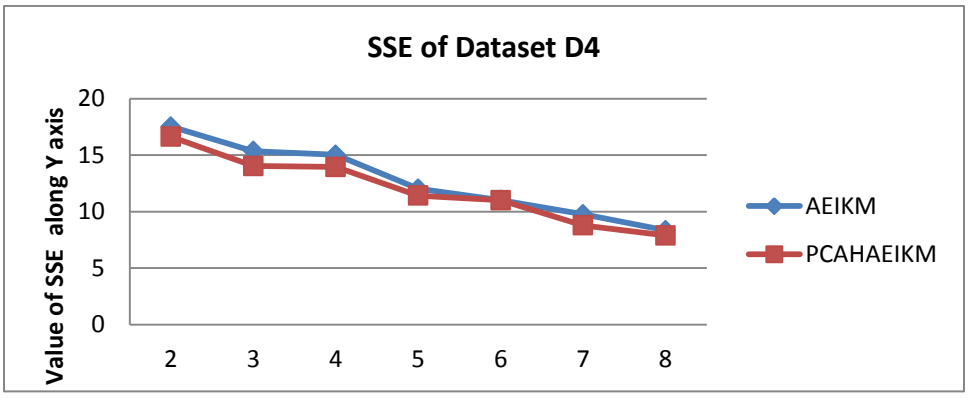


Figure 4.1(d): Analysis cluster vs. value of SSE

4.4.2 Intra Cluster Distance (ICD)

In Table-4.2, illustrate the comparative analysis of Intra Cluster Distance (ICD) of four datasets creates the clusters follow as:

Table 4.2 Analysis of Intra Cluster Distance metrics of clusters

Dataset		Number of Clusters						
		2	**3**	**4**	**5**	**6**	**7**	**8**
Dataset D1	AEIKM	12.5261	6.2281	5.2655	3.8188	5.9623	6.6601	5.0407
	PCAHAEIKM	12.2363	5.7264	4.4729	6.3277	4.3881	4.3067	2.9712
Dataset D2	AEIKM	0.1262	0.0984	0.1240	0.1071	0.1035	0.0906	0.1010
	PCAHAEIKM	0.0789	0.0686	0.0455	0.0360	0.0345	0.0396	0.0358
Dataset D3	AEIKM	0.1155	0.0599	0.0599	0.1038	0.1038	0.1034	0.0965
	PCAHAEIKM	0.0550	0.0407	0.0407	0.0535	0.0535	0.0668	0.0759
Dataset D4	AEIKM	7.2290	8.3187	6.3211	7.0383	4.2669	1.3806	5.0233
	PCAHAEIKM	1.4639	1.4108	1.5033	1.2289	1.1368	1.1791	1.2112

Heart Dataset D1

In figure-4.2(a), at cluster K=2, 3, 4, 5, 6, 7, 8, illustrate analysis of intra cluster distance of heart disease (D1) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure at k=2 approximately same but k= 3, 4, 6, 7, 8, slightly reduced in hybridize concept but at k=5 is more. The performance PCAHAEIKM (proposed method hybridize via PCA) method is better than AEIKM method.

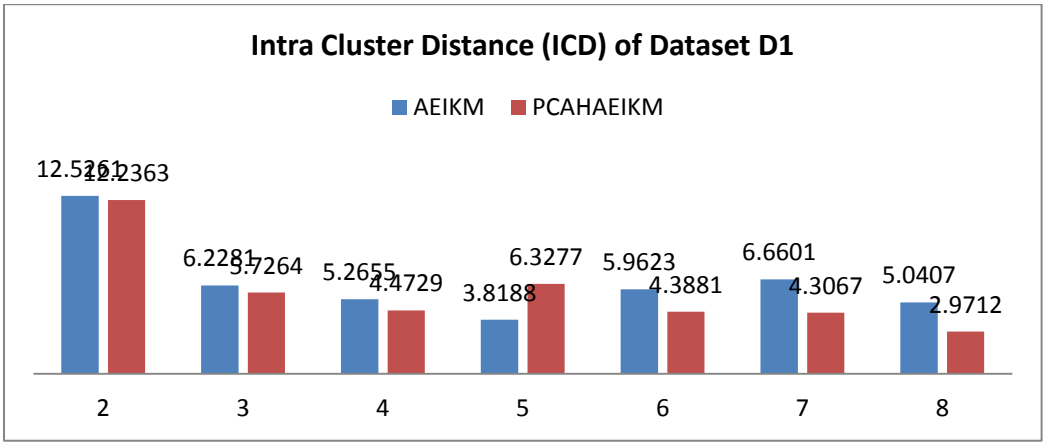


Figure 4.2(a): Analysis cluster vs. intra cluster distance

User Knowledge Modeling Dataset D2

In figure-4.2(b), at cluster K= K=2, 3, 4, 5, 6, 7, 8, illustrate analysis of intra cluster distance of user knowledge modeling (D2) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure show the ICD value of PCAHAEIKM method lower than AEIKM method

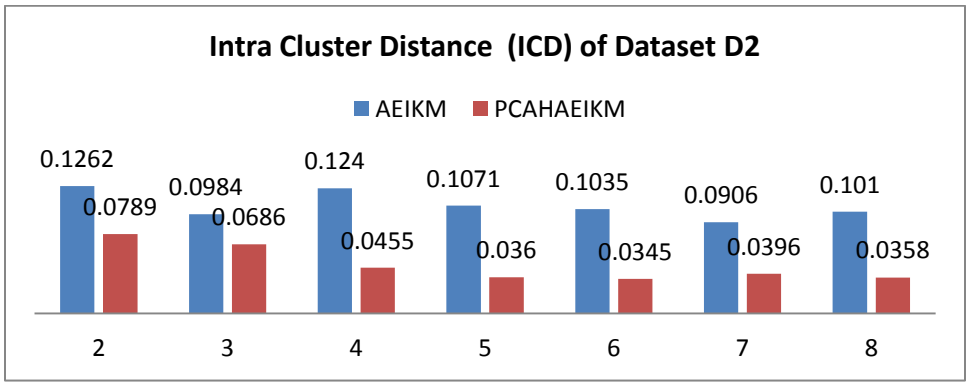


Figure 4.2(b): Analysis cluster vs. Intra cluster distance

Iris Dataset D3

Illustrate in figure-4.2(c), at cluster K= K=2, 3, 4, 5, 6, 7, 8, analysis of intra cluster distance of iris (D3) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure show the ICD value of PCAHAEIKM method lower than AEIKM method**.**

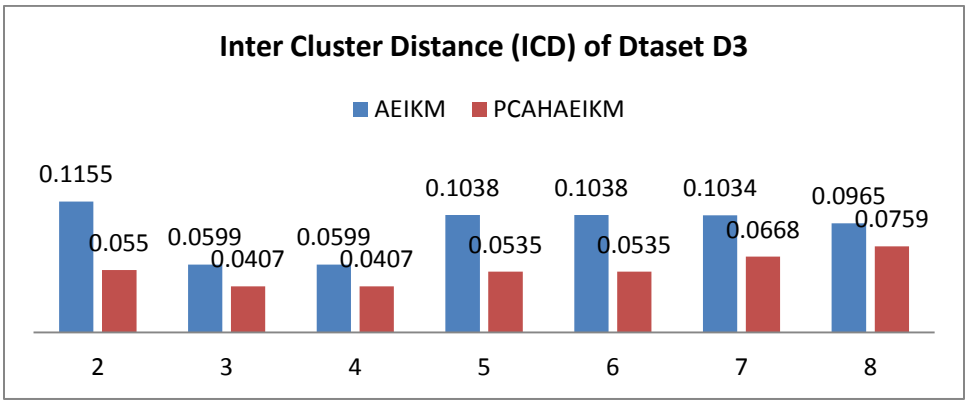


Figure 4.2(c): Analysis cluster vs. Intra cluster distance

Wine Dataset (D4)

In figure-4.2(d), at cluster K=2, 3, 4, 5, 6, 7, 8, illustrate analysis of intra cluster distance of wine (D4) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure show the ICD value of PCAHAEIKM method lower than AEIKM method, but at k=7 approximately closed.

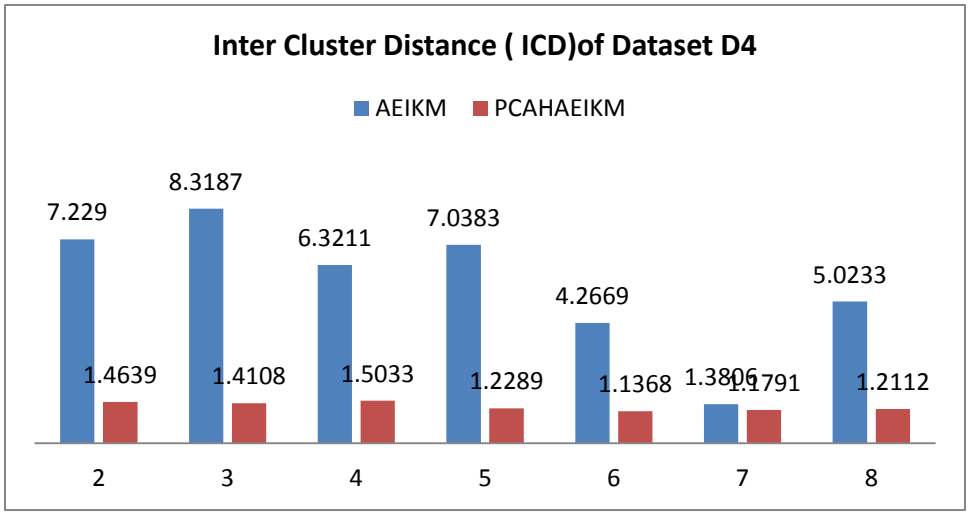


Figure 4.2(d): Analysis cluster vs. intra cluster distance

4.4.3 Execution Time

In Table-4.3, illustrate the comparative analysis of execution time taken by the CPU for four dataset follow as:

Table 4.3 Analysis of Execution Time (in Second)

Datasets	Methods	2	3	4	5	6	7	8
Dataset D1	AEIKM	0.083079	0.141423	0.157602	0.159408	0.193956	0.18536	0.211756
	PCAHAIEKM	0.081699	0.130487	0.142256	0.148058	0.186267	0.184486	0.190353
Dataset D2	AEIKM	0.086072	0.11098	0.163334	0.165998	0.16809	0.178608	0.203497
	PCAHAEIKM	0.081145	0.102346	0.144753	0.158375	0.162628	0.163021	0.199462
Dataset D3	AEIKM	0.042995	0.0446	0.101979	0.11201	0.112891	0.128458	0.13043
	PCAHAEIKM	0.03354	0.0415	0.081975	0.102258	0.103791	0.121398	0.12303
Dataset D4	AEIKM	0.07277	0.097863	0.121516	0.131029	0.151821	0.169077	0.185373
	PCAAEIKM	0.067307	0.085154	0.10774	0.13076	0.145083	0.157862	0.180895

Heart Disease Dataset (D1)

At cluster K=2, 3, 4, 5, 6, 7, 8 In figure-4.3(a), illustrate analysis of execution time (in second) of heart disease (D1) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure show the execution time in second value of PCAHAEIKM method just lower than AEIKM method, but at k=2, 6, and 7 approx. same.

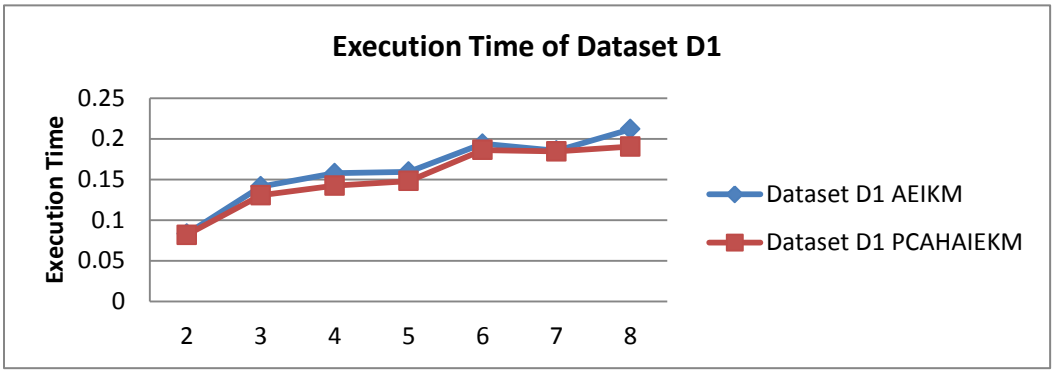


Figure 4.3(a): Analysis cluster vs. execution time

User Knowledge Modeling Dataset (D2)

At cluster K=2, 3, 4, 5, 6, 7, 8, in figure-4.3(b), illustrate analysis of execution time (in second) of user knowledge (D2) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure show the execution time in second value of PCAHAEIKM method just lower than AEIKM method, but at k=2, and 8 approx. same.

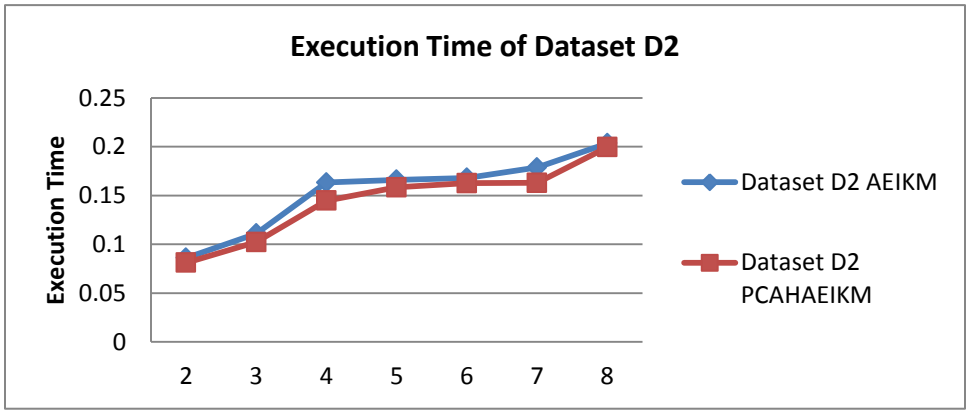


Figure 4.3(b): Analysis cluster vs. execution time

Iris Dataset (D3)

At cluster K=2, 3, 4, 5, 6, 7, 8 In figure-4.3(c), illustrate analysis of execution time (in second) of iris (D3) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure show the execution time in second value of PCAHAEIKM method just lower than AEIKM method.

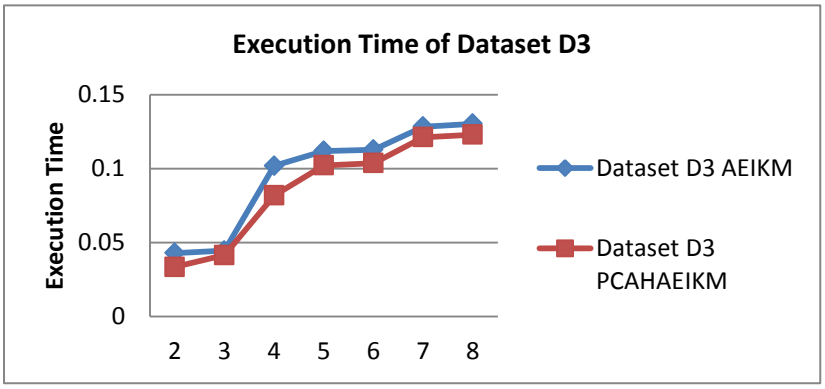


Figure 4.3(c): Analysis cluster vs. execution time

Wine Dataset (D4)

At cluster K=2, 3, 4, 5, 6, 7, 8 illustrate in figure-4.3(d), analysis of execution time (in second) of wine (D4) dataset using AEIKM method, and PCAHAEIKM method respectively. In figure show the execution time in second value of PCAHAEIKM method just lower than AEIKM method, but at k=5 approx. same.

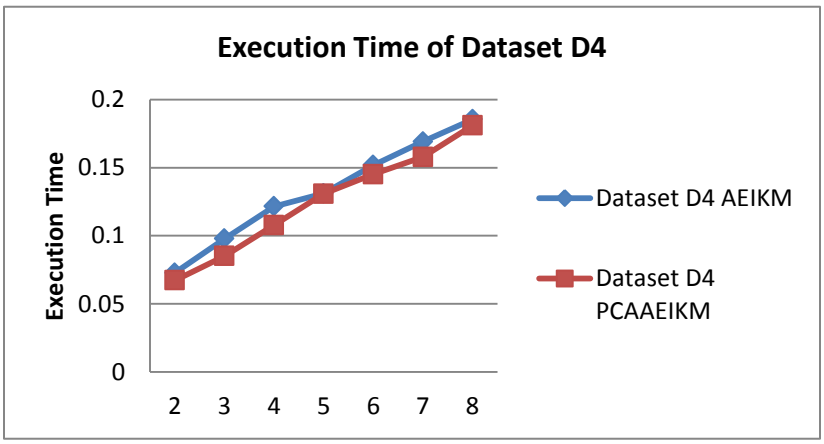


Figure 4.3(d): Analysis cluster vs. execution time

At cluster K=2, 3, 4, 5, 6, 7, 8, in figure-4.1(a), figure-4.2(a), and figure-4.3(a) illustrate analysis of Sum of Squared Error(SSE), intra cluster distance, and execution time (in second) of heart disease (D1) dataset, using method AEIKM, and PCAHAEIKM respectively. Performance of PCAHAEIKM method is better than AEIKM method.

At cluster K=2, 3, 4, 5, 6, 7, 8, in figure-4.1(b), figure-4.2(b), and figure-4.3(b) illustrate analysis of Sum of Squared Error(SSE), intra cluster distance, and execution time (in second) of user knowledge modeling (D2) dataset, using method AEIKM, and PCAHAEIKM respectively. The evaluate results of PCAHAEIKM method compared than AEIKM method is well.

At cluster K=2, 3, 4, 5, 6, 7, 8, in figure-4.1(c), figure-4.2(c), and figure-4.3(c) illustrate analysis of Sum of Squared Error(SSE), intra-cluster distance, and execution time (in second) of Iris (D3) dataset, using method AEIKM, and PCAHAEIKM respectively. Method PCAHAEIKM is better than AEIKM method.

At cluster K=2, 3, 4, 5, 6, 7, 8, in figure-4.1(d), figure-4.2(d), and figure-4.3(d) illustrate analysis of Sum of Squared Error(SSE), intra cluster distance, and execution time (in second) of Wine (D4) dataset, using method AEIKM, and PCAHAEIKM respectively. The measured performance (metrics) of PCAHAEIKM method is better than AEIKM method.

4.4.4 External Metric Performance

The external metric performance is illustrated in Table-4.4 at k=4. The value of precision, recall, F-measure, and rand index (or accuracy) for heart disease (D1), user modeling knowledge (D2), iris (D3), and wine (D4) datasets are using methods AEIKM method, and PCAHAEIKM method.

Table 4.4 Comparative Analysis of Cluster Metrics

Dataset	Metrics	k-means	AEIKM	PCAHAEIKM
Heart Dataset at k=4	Precision	0.592	0.678	0.701
	Recall	0.610	0.682	0.718
	F-measure	0.623	0.683	0.710
	(Accuracy) in percentage	65.4	68.6	86.6
User Knowledge Modeling at k=4	Precision	0.601	0.687	0.716
	Recall	0.641	0.664	0.766
	F-measure	0.653	0.675	0.742
	(Accuracy) in percentage	61.5	69.2	75.8
Iris Dataset at k=4	Precision	0.701	0.827	0.927
	Recall	0.734	0.812	0.912
	F-measure	0.785	0.819	0.920
	(Accuracy) in percentage	84.3	92.3	94.89
Wine Dataset at k=4	Precision	0.736	0.860	0.960
	Recall	0.744	0.835	0.955
	F-measure	0.756	0.874	0.957
	(Accuracy) in percentage	81.6	87.9	91.89

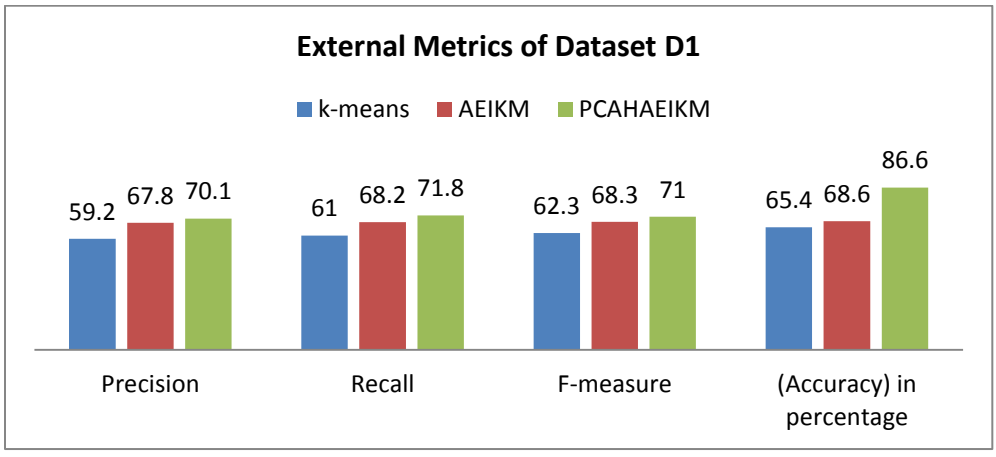


Figure 4.4 (a) Analysis of external metrics of Dataset D1

In figure-4.4(a), analysis of precision, recall, f-measure, and rand index (or accuracy) are mentioned of heart disease (D1) dataset, measured by k-means, proposed methods AEIKM and proposed method hybridize via PCA (PCAHAEIKM) method.

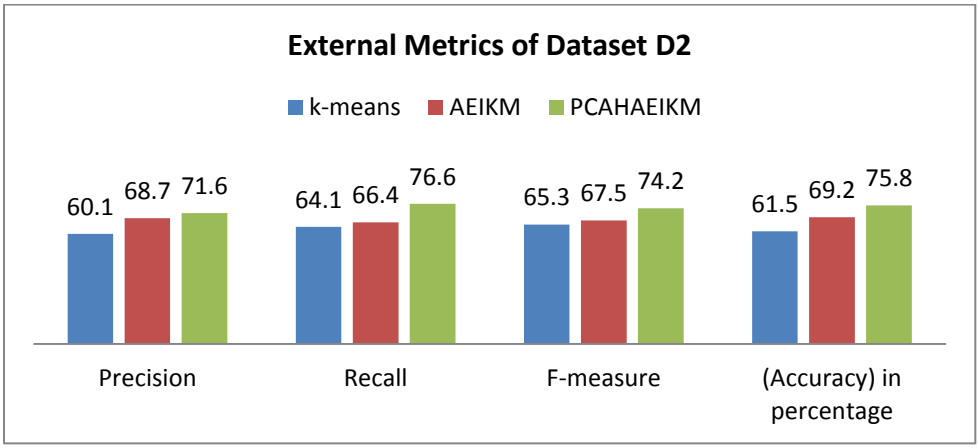


Figure 4.4 (b) Analysis of external metrics of Dataset D2

In figure 4.4(b), analysis of precision, recall, f-measure, and rand index (or accuracy) mentioned of user knowledge modeling (D2) dataset, measured by k-means, proposed methods AEIKM and proposed method hybridize via PCA (PCAHAEIKM) method.

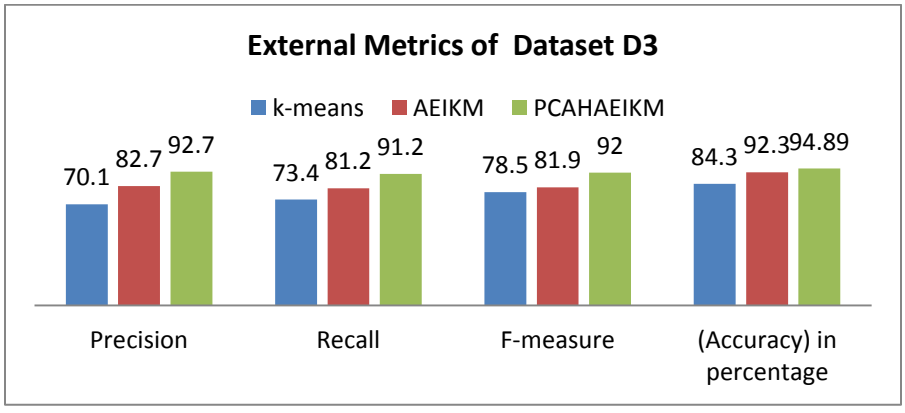


Figure 4.4 (c) Analysis of external metrics of Dataset D3

In figure-4.4(c), analysis of precision, recall, f-measure, and rand index (or accuracy) mentioned of iris (D3) dataset, measured by k-means, proposed methods AEIKM and proposed method hybridize via PCA (PCAHAEIKM) method.

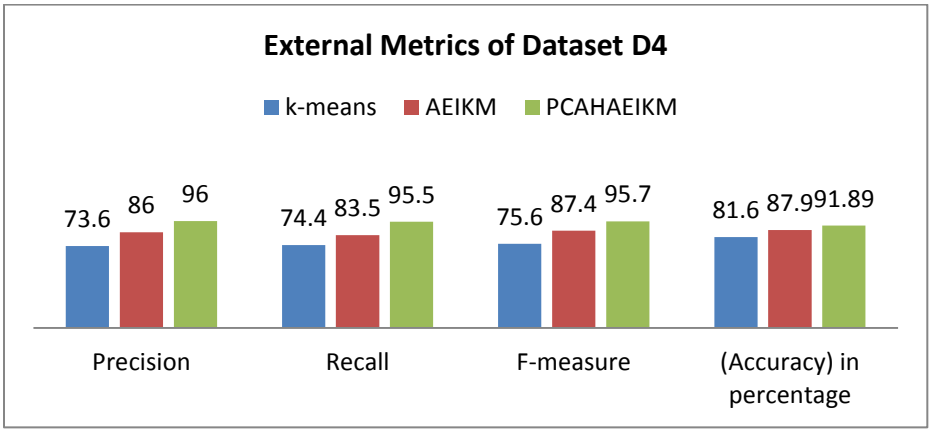


Figure 4.5 (d) Analysis of external metrics of Dataset D4

In figure-4.4(d), analysis of precision, recall, f-measure, and rand index (or accuracy) mentioned of wine (D4) dataset measured by k-means, proposed methods AEIKM , and proposed method hybridize via PCA (PCAHAEIKM) method.

4.4.5 Compare Fitness Function Using AEIKM Methods

Particle swarm optimization to be relate on k-means algorithm for more convergence the different parameters to set L1=1.5, L2=1.5, max ω =0.8, at no. of cluster k=3 for iris dataset (D3) and K= 4 for rest three datasets D1, D2, and D4. The obtain the Best _fitness of the objective function for heart disease dataset (D1) at the population size 297, user knowledge modeling dataset (D2) size 403, iris dataset (D3) at the population size 150, and wine dataset (D4) at the population size 178,. In Table-4.5 represent the fitness of same four datasets are as k-means method as compared by proposed method An Efficient Improved K-Means (**AEIKM**), and also compared proposed method hybridized via Principal Component Analysis (PCA) and Particle Swarm Optimization (PSO).

In Table- 4.5 shown measures the fitness of fitness function $F_i^K(X)$by applied the methods namely as AEIKM method, PCAHAEIKM method, and PSOHAEIKM method respectively. In Figure-4.6, illustrate that method PSOHAEIKM is better than the rest methods.

Table 4.5 Performance of Fitness Analysis of Cluster

Datasets	AEIKM	PCAHAEIKM	PSOHAEIKM
Dataset D1	76	84	135
Dataset D2	67	73	140
Dataset D3	89	89	164
Dataset D4	83	85	156

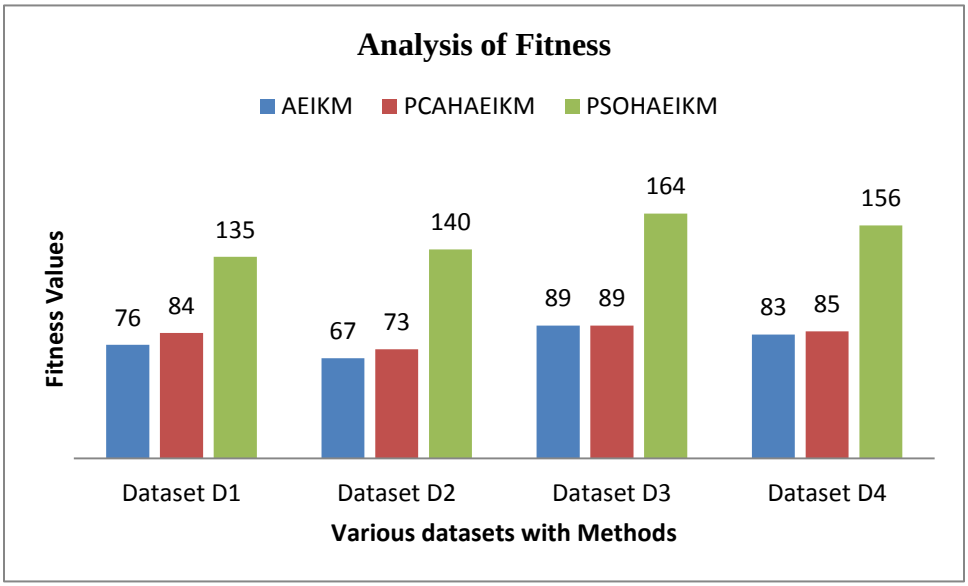


Figure 4.6 Analyses of Fitness of Datasets D1, D2, D3, and D4

4.5 Summary

Chapter-4, Measure the cluster performance using proposed techniques applies on heart disease (D1), user knowledge modeling (D2), iris (D3), and wine (D4) datasets respectively. The results have a lesser amount of execution time with reduced the intra cluster distance, minimized SSE at various cluster level and also analysis the sound effects of precision, recall, F-measure, and increase the accuracy for a cluster at K=4 by using AEIKM method, PCAHAEIKM method, but the result find by PCAHAEIKM method is good as compared than AEIKM method. At the present time, advancement of research works in the capacity of mining for large dimensional datasets to explore by metrics of clustering. Furthermore, PSO hybridized via AEIKM method is found the higher fitness value as compared than AEIKM method, PCAHAEIKM method.

CHAPTER-5

STATISTICAL ANALYSIS AND IMPLEMENTATION

Measure the performance of cluster by proposed techniques applied on heart disease, user knowledge modeling, iris, and wine datasets respectively. The results have of cluster quality are better discussed in chapter 4. But in the chapter-5, statistical analyses of a cluster are using software tools SPSS 17.0 and NCSS2021 on large datasets.

5.1 Introduction

Partitioning of the problems handle by the idea of a cluster and this technique is playing the essential work mining of data from the given dataset. The K-Means method of clustering is the most suitable recognized idea; it is too applied on large data but has the several shortcomings. Therefore, getting the outcome of this method is required to overcome the drawbacks and improve the quality of cluster using the Principal Component Analysis (PCA) on the dataset given. In this chapter illustrate the experimental results for four data sets with different sizes. We have carried out to validate the experimental effects on clusters metrics and also component's size of the different datasets to reduce while the processing on SPSS tool on the basis of eigenvalues. In this chapter, we have also discussed, it is the relative study of the distance between the original centroid of iris, wine and heart disease dataset at the different clusters.

The extraction and analyzed of data from huge dataset to apply the statistical software tool with used the idea of AI (Artificial Intelligence). The data mining research field is very supportive in e-business and its applications to require a demo of functionality for industries (Olson, D. L., 2007) [48], applications in business like as marketing, advertising, and the promoting reported (Kudyba and Hoptroff, 2001) [49]. Tsai, C. F.,

et.al. (2004) they study the approaches of k-means clustering and also developed the method to enhanced the performance [53].

In this chapter PCA techniques utilize on numerical attributes on the database the noisy feature reduces the dimensions of problem consider as the dataset but improve the cluster quality on the basis of the min distance between initial centroids.

5.1.1 Mathematical Illustration of Coefficient Matrix

Definition-1: Consider that the set of data value X is consist nonempty set members of attribute m , extraction of data sample n and normalize of all members of attribute illustrate in the mathematical form

$$Mean\ of\ X_{at\ j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \tag{5.1}$$

$$Normalization = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij})^2 = 1 \tag{5.2}$$

Where, x_{ij} values normalized at set j=1, 2, 3……, m

Definition-2: Consider the collection of set of nonempty attribute

$X_1, X_2, X_3, \ldots\ldots\ldots\ldots\ldots, X_m$ And associated the coefficient matrix

C at the m attributes, illustrate set as equation (3), as matrix equation (4), and as sparse matrix equation (5)

Set C

$$= \{ Cov_{11,} Cov_{12,} Cov_{13,} \ldots\ldots, Cov_{1m}; Cov_{21,} Cov_{22,} Cov_{23,} \ldots\ldots, Cov_{2m}; \ldots\ldots\ldots\ldots\ldots; Cov_{m1,} Cov_{m2,} Cov_{m3,} \ldots\ldots, Cov_{mm} \} \tag{5.3}$$

Illustrate the Matrix C as

$$C = \begin{bmatrix} \mathrm{Cov}_{11} & \mathrm{Cov}_{12} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \mathrm{Cov}_{1m} \\ \mathrm{Cov}_{21} & \mathrm{Cov}_{22} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \mathrm{Cov}_{2m} \\ \mathrm{Cov}_{31} & \mathrm{Cov}_{32} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \vdots & \vdots & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \mathrm{Cov}_{m1} & \mathrm{Cov}_{m2} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \mathrm{Cov}_{mm} \end{bmatrix} \tag{5.4}$$

Sparse Matrix

$$C = \begin{bmatrix} \mathrm{Cov}_{11} & \cdots & \mathrm{Cov}_{1m} \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}_{m1} & \cdots & \mathrm{Cov}_{mm} \end{bmatrix} \quad (5.5)$$

Where the notation $C_{ij} = C$ is covariance matrix is the covariance coefficient between the X_i and X_j

5.1.2 Evaluate the Eigenvalues and Covariance Matrix

Determine the characteristics value from consider the "characteristic equation" $|\lambda I - C = 0|$, find the "eigenvalues" λ_i and sorted it's like $\lambda_1 \geq, \lambda_2 \geq, \lambda_3, \geq, \ldots\ldots \lambda_m$. and find the orthogonal eigenvector. Here simulated the eigenvalues$\lambda_1 \geq, \lambda_2 \geq, \lambda_3, \geq, \ldots\ldots \lambda_m$, related to$p^{th}$ principal component($p \leq m$). Therefore involvement of the eigenvalues rate is higher than value 1.00 to pick from the given datasets.

5.2 Back Ground

Clustering algorithm applies on d-Dimensional dataset, Choose k-prototype centroid at random from data point, create the early dividing of cluster by assigning the object to the closest Centroid, and finally create the cluster (Zhang, N., et.al. 2018; Jamal, A. et.al, 2018)[51][52].

Principal Component Analysis (PCA) projects the input space to feature space through linear mapping. The PCA is can only extract a linear projection of the data as following manner- Consider the data $X_1, X_2, X_3, \ldots\ldots, X_M$ are N vectors (Ding and He, 2004) [57].

Ilc, N.: (2012) he fined the quality of cluster by silhouette; therefore silhouette $s_{(k)}$ is clearly defined as

$$s(k) = \begin{cases} 1 - [\frac{a(k)}{b(k)}] & if a(k) < b(k) \\ 0 & if a(k) = b(k) \\ [\frac{b(k)}{a(k)}] - 1 & if\ a(k) > b(k) \end{cases} \quad (5.6)$$

Differences of the k^{th} object to other rest object in the own cluster, neighboring cluster and range of silhouette is -1≤ $s_{(k)}$ ≥1,

Case: 1- if $s_{(k)}$ close to 1, cluster clearly distinguish (better quality of cluster) or good cluster,

Case: 2- if $s_{(k)}$ =0, then sample represent the overlying in cluster or not significant,

Case: 3- if $s_{(k)}$ close to -1, then sample misclassified or create the cluster in wrong direction, and simplified form

5.3 Statistical Clustering Approaches

K-Means Cluster Analysis Using NCSS 2021:

This is the classification show separation into form groups. The main objective of it analysis categorize n objects in K Groups if satisfy the condition K >1 and using V variable and set condition V>1 where K is called cluster. It analysis consists some variants and its own procedure of clustering and using Euclidean metric.

Measure the validity index of cluster (K-Means cluster) by using formula

$$\text{F-ratio} = \frac{\frac{\text{Distance between Mean Squares}}{(\text{DF1})}}{\frac{\text{Distance within Mean Squares}}{(\text{DF2}))}} \qquad (5.7)$$

Where, $(C-1)$= DF1, $(n-C)$=DF2

C= no. of cluster, n =no. of observations

Estimates separation between all clusters and should be high, in the examination of variance from table significance level of F-ratio is more considerable, significance value is high so minimum contribution of particular variable to separation of this clusters.

Fuzzy algorithm trying to find reduce the objective function described by author in (Bezdek and Pal , 1995)[92], follow as

$$C = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_{ik}^{2}\,\mu_{jk}^{2}\; d_{ij}}{2\sum_{j=1}^{N}\mu_{jk}^{2}} \qquad (5.8)$$

Where C= Objective function of cluster form the membership m and d distance

Where

μ_{ik} = Unknown membership of i^{th} object into k cluster,

d_{ij} = Indicate the dissimilarities pair of i^{th} and j^{th}object

The membership is non-negative for individuals, but sum to 1.

For **goodness- fit**;

1. $F(U) = {}^{1}/_{N} \sum_{k=1}^{K} \sum_{i=1}^{N} \mu_{ik}^{2}$, Where, $\frac{1}{K} \leq F(U) \leq 1$
2. $F_c(U) = \frac{F(U)-(\frac{1}{K})}{1-(\frac{1}{K})}$
3. $D(U) = {}^{1}/_{N} \sum_{k=1}^{K} \sum_{j=1}^{N} (h_{ik} - \mu_{ik})^2$, Where,$O \leq D(U) \leq (1 - {}^{1}/_{K})$
4. D_c= D (U)/ (1-(1/K))

Dunn's partitioned coefficient (FU), and it's normalized therefore vary from zero (0) to one (1) (F_c (U)); Kaufman's partitioned coefficient $D(U)$ and its normalized therefore vary from 0 to 1 (D_c (U))

Optimum cluster is represented by both F_c (U) and D_c (U) and choose F_c (U) > D_c (U).

5.4 UPCAKM\\ Utility of Principal Component Analysis with K-means

Statistical analyses of large various datasets and make cluster quality is initially not good. PCA is good concept the reduction of dimension of large datasets. Analyzes the dataset, it is construct the clusters and then reduced dimensions. Make good quality of clusters are followed two steps first one POSA (Procedure of Statistical Analysis), and second steps KMWPCA (K-Means with PCA). These ideas are represented as:

5.4.1 Procedure POSA \\ Procedure of Statistical Analysis

The adopted procedure is as follows:

1. Initially set the name of variables or attribute then after filled the data into data view filed
2. Create the structured data set in 2D
3. Go to analyze the 2D structure data set until return the eigenvalues

i. Select the dimension reduction tool factor
ii. Select " coefficient" of component from descriptive field
iii. Choose Removal concept to build the "variance matrix" based on eigenvalues, with eigenvalues is set greater than one and "maximum convergence" repetition is set to 25.

4. return " eigenvalues"
5. Stop

5.4.2 Procedure KMWPCA \\ procedure K-Means with PCA

1. Set organized dataset, and then applies on the reduction tools
2. Removal of particular constituent by step-1
3. Consider the set of variance and initial eigenvalues.
4. To projecting the data in lower dimensional subspace, getting the reduced dimension of real datasets
5. After reducing the component of dataset to evaluate k-means, and obtained the separation between original centroid for the best creates the clusters.
6. Stop the procedure

5.5 Experimental Results

The discussion of experimental results is divided into sub section as follows:.

5.5.1: Analysis of Reduction Component

5.5.2: Comparative Analysis of F-Ratio

5.5.3: Comparative Analysis of Average Silhouette

5.5.4 Comparative analysis of centroids

5.5.1: Analysis of Reduction Component

In this chapter, an idea used to analyze the PCA implementation on SPSS (Statistics 17.0 software tool). The comparative analysis of extraction component via PCA implemented on this tool, and shown are the results in Table-5.1 Figure-5.1, Table-5.2 Figure-5.2, Table-5.3 Figure-5.3, and Table-5.4 Figure-5.4 respectively.

Heart Disease Dataset (D1)

It has a no. of instances 297, no. of attribute 14 and multivariate characteristics of the dataset.

Table- 5.1 Variance Explained of Heart Disease Dataset (D1)

Total Variance Explained						
	Initial Eigenvalues(λ)			Extraction Sums of Squared Loadings		
Component	Values of λ	% of Variance(σ^2)	Cumulative %	Total	% of Variance	Cumulative %
1	3.097	23.823	23.823	3.097	23.823	23.823
2	1.578	12.139	35.962	1.578	12.139	35.962
3	1.261	9.702	45.664	1.261	9.702	45.664
4	1.108	8.524	54.188	1.108	8.524	54.188
5	1.005	7.728	61.916	1.005	7.728	61.916
6	0.877	6.750	68.666	-	-	-
7	0.837	6.438	75.104	-	-	-
8	0.752	5.784	80.888	-	-	-
9	0.683	5.253	86.140	-	-	-
10	0.559	4.303	90.443	-	-	-
11	0.464	3.566	94.010	--	-	-
12	0.417	3.210	97.219	-	-	-
13	0.361	2.781	97.778ss	-	-	-
14	0.311	2.222	100	-	-	-

All the descriptions shows and five component extraction on the basis of eigenvalues indicate in table 5.1. The all data of heart disease (D1) are consists into three clusters, in

cluster (K1) 56, in cluster (K2) 61, in cluster (K3) 103, in cluster (K4) 77 valid 297, and missing object is 0.00. The cluster data have total 297 valid, but no missing data. In Table- 5.1, show the variance of Heart Disease Dataset (D1) and figure- 5.1 extractions of five components of same dataset.

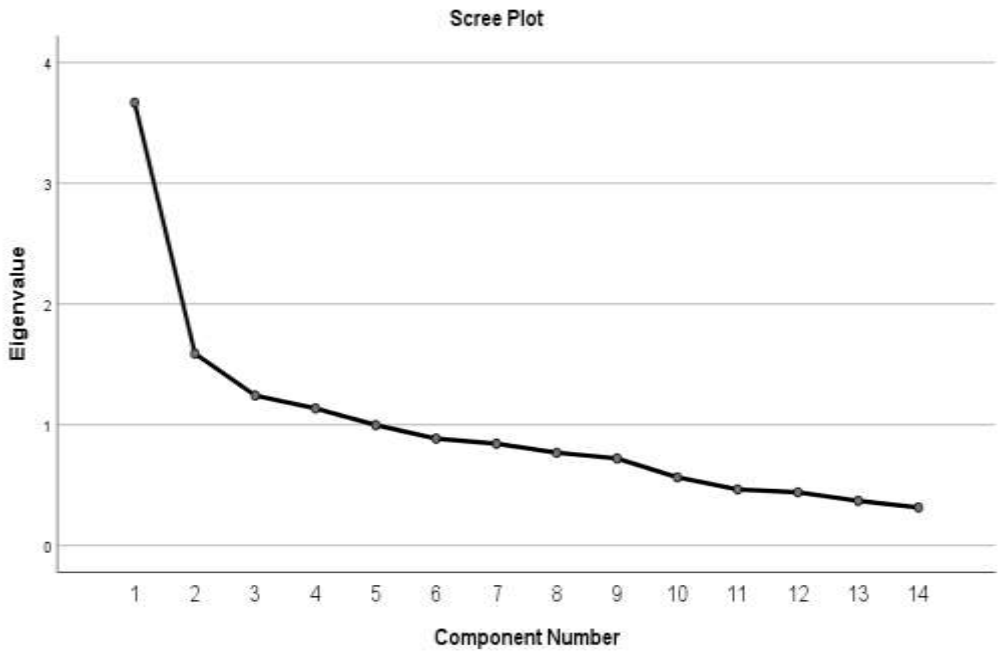


Figure 5.1 Five extraction component of dataset D1

User Knowledge Modeling (D2)

It has a no. of instances 258 (used), no. of attribute 05 and multivariate characteristics of the dataset. All the descriptions shows and 02 component extraction on the basis of eigenvalues indicate in table- 5.2. The all data of heart disease (D1) are consists into three clusters, in cluster (K1) 60, in cluster (K2) 94, in cluster (K3) 93, in cluster (K4) 61 valid 257, and missing object is 0.00. The cluster data have total 257 valid, but no missing data. In Table-5.2, show the variance of User Knowledge Modeling (D2) and figure 5.2- extractions of two components of same dataset.

Table- 5.2 Variance Explained of User Knowledge Modeling (D2)

Total Variance Explained						
	Initial Eigenvalues (λ)			Extraction Sums of Squared Loadings		
Component	Values of λ	% of Variance(σ^2)	Cumulative %	Total	% of Variance	Cumulative %
1	1.382	27.640	27.640	1.382	27.640	27.640
2	1.173	23.451	51.091	1.173	23.451	51.091
3	0.973	19.454	70.545	-	-	-
4	0.915	18.305	88.850	-	-	-
5	0.558	11.150	100.000			

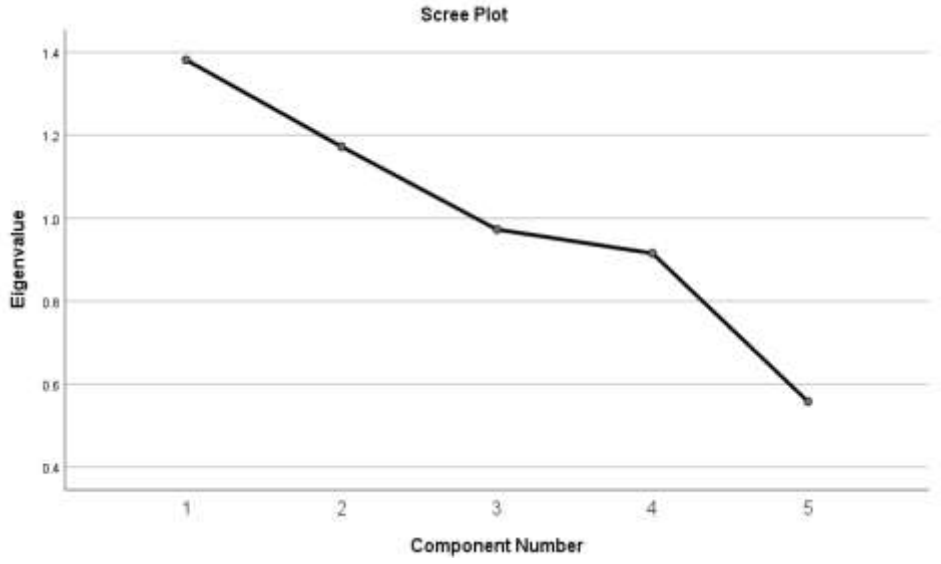


Figure 5.2 Two extraction component of User Knowledge Modeling (D2)

Iris Dataset (D3)

It has a no. of instances 150, no. of attribute 4 and multivariate characteristics of the dataset. All the descriptions shows and one component extraction on the basis of eigenvalues indicate in table 5.3.

Table 5.3 Variance Explained of Iris Dataset (D3)

Total Variance Explained						
	Initial Eigenvalues (λ)			Extraction Sums of Squared Loadings		
Component	Values of λ	% of Variance(σ^2)	Cumulative %	Total	% of Variance	Cumulative %
1	2.921	73.0176	73.016	2.921	73.016	73.016
2	0.918	22.949	95.965	-	-	-
3	0.141	3.518	99.484	-	-	-
4	0.021	0.516	100.000	-	-	-

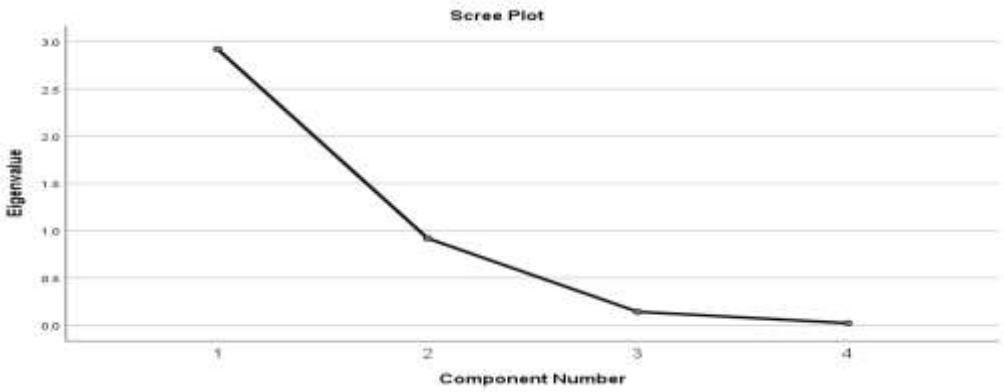


Figure 5.3 One extraction component of Iris dataset (D3)

The all data of iris are consists into three clusters, in cluster (K1) 38, in cluster (K2) 50, in cluster (K3) 62, valid 150, and missing object 02. The cluster data have total 150 valid

no missing data. There is one component extracted from the dataset represent in figure-5.3

Wine Dataset (D4)

Table- 5.4 Variance Explained of Wine Dataset (D4)

Total Variance Explained of Wine Dataset						
Component	**Initial Eigen values(λ)**			**Extraction Sums of Squared Loadings**		
	Values of λ	% of Variance(σ^2)	Cumulative %	Total	% of Variance	Cumulative %
1	4.706	36.199	36.199	4.706	36.199	36.199
2	2.497	19.207	55.406	2.497	19.207	55.406
3	1.446	11.124	66.530	1.446	11.124	66.530
4	0.919	7.069	73.599	-	-	-
5	0.853	6.563	80.162	-	-	-
6	0.642	4.936	85.098	-	-	-
7	0.551	4.239	89.337	-	-	-
8	0.348	2.681	92.018	-	-	-
9	0.289	2.222	94.240	-	-	-
10	0.251	1.930	96.170	-	-	-
11	0.226	1.737	97.907	-	-	-
12	0.169	1.298	99.205	-	-	-
13	0.103	0.795	100.000	-	-	-

It has a no. of instances 178 (one hundred and seventy eight), no. of attribute 13 (thirteen) and multivariate characteristics of the dataset. All the descriptions shows and three component extraction on the basis of eigenvalues indicate in Table-5.3. The all data of wine are consists into three clusters, in cluster (K1) 61, in cluster (K2) 69 (sixty nine), in cluster (K3) 41(forty one), in cluster (K4) 07valid 178 (one hundred seventy eight), and missing object 0.00. The cluster data have total 178 (one hundred seventy eight) valid, but no missing data. There is three component extracted from the dataset represent in figure-5.4.

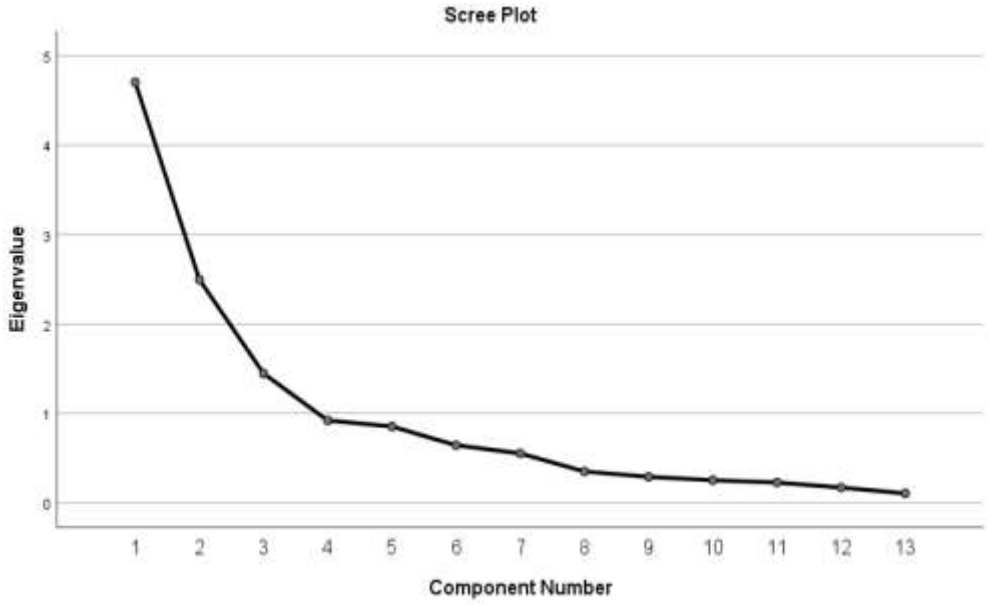


Figure 5.4 Three extraction component of Wine dataset (D4)

5.5.2: Comparative Analysis of F-Ratio

And also discussed is comparative statistical analysis of cluster by K-Means, and its clusters concept via PCA by NCSS 2021 are shown in Table 5.5 to Table 5.10 respectively. The average Silhouette values of four datasets namely heart disease (D1), user knowledge modeling (D2), iris (D3), and wine (D4).

5.5.2.1 Expected F-Ratio by K-Means Clustering

Heart Disease Dataset (D1)

From the Table-5.5 analysis of F-Ratio by k-means at K=4 heart disease dataset (D1) by k-means represented that the estimated value of F-Ratio which is less than tabled value of 2.0838 at 5% with d. f. being used df1= 3, df2= 293/ or INF, and hence could have increase by chance.

Table 5.5 Analysis of F-Ratio by k-means at K=4 Dataset (D1)

Attribute	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
1	3	293	932.3322	73.19019	12.74
2	3	293	15.74588	0.06051897	260.18
3	3	293	16.55704	0.7709596	21.48
4	3	293	1838.158	299.9271	6.13
5	3	293	2435.536	1823.511	1.34
6	3	293	10.62702	0.01670084	636.32
7	3	293	3.565037	0.9634864	3.70
8	3	293	17432.14	353.2181	49.35
9	3	293	5.684576	0.1647308	34.51
10	3	293	32.60195	0.6360666	51.26
11	3	293	7.715275	0.3070717	25.13
12	3	293	20.50617	0.6807207	30.12
13	3	293	133.1428	2.433525	54.71
14	3	293	78.62064	0.7347307	107.01

Hence, estimated value of F-Ratio (1.34) is < F-Rato of tabled value (2.0838).It is support to the Null-Hypothesis no difference attribute-5 data means. We, conclude that statistical insignificant of attribute-5 data but rest attribute-1,2,3,4,6,7,8,9,10,11,12,13, and 14 for support to clusters creations

User Knowledge Modeling Dataset (D2)

From the Table-5.6 analysis of F-Ratio by k-means at K=4 Dataset (D2) by k-means represented that the estimated value of F-Ratio which is greater than tabled value of 2.0838 at 5% with d. f. used df1= 3, df2= 293/ or INF. Hence, estimated values of F-Ratio are > F-Rato of tabled value (2.0838). It is not support to the Null-Hypothesis has differences of data means. We, conclude that statistical significant of attribute- 1, 2, 3, 4 and 5 for support to clusters creations.

Table 5.6 Analysis of F-Ratio by k-means at K=4 Dataset (D2)

Variables	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
1	3	254	2.163348	0.01918484	112.76
2	3	254	2.354623	0.01764816	133.42
3	3	254	0.1380075	0.05957077	2.32
4	3	254	1.212924	0.04795877	25.29
5	3	254	2.822357	0.03256695	86.66

Iris dataset (D3)

From the Table-5.7 analysis of F-Ratio by k-means at K =4 iris dataset (D3) by k-means represented that the estimated value of F-Ratio which is greater than tabled value of 2.0838 at 5% with d. f. used df1= 3, df2= 293/ or INF. Hence, estimated values of F-Ratio are > F-Rato of tabled value (2.0838). It is not support to the Null-Hypothesis has differences of data means. We, conclude that statistical significant of attribute- 1, 2, 3, and 4 for support to clusters creations.

Table 5.7 Analysis of F-Ratio or F-Score by k-means at K=4 Dataset (D3)

Attributes	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
1	3	146	26.36147	0.1581092	166.73
2	3	146	143.7958	0.222441	646.44
3	3	146	5.817593	0.07232754	80.43
4	3	146	26.11542	0.0577636	452.11

Wine dataset (D4)

Table 5.8 Analysis of F-Ratio or F-Score by k-means at K=4 Dataset (D4)

Attributes	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
1	3	174	23.49265	0.2653798	88.52
2	3	174	25.48553	0.8301273	30.70
3	3	174	0.7384082	0.06383113	11.57
4	3	174	213.3451	7.66661	27.83
5	3	174	2756.99	159.9721	17.23
6	3	174	14.61365	0.1464834	99.76
7	3	174	46.14998	0.2192313	210.51
8	3	174	0.3022202	0.01054499	28.66
9	3	174	7.987355	0.1955298	40.85
10	3	174	184.3616	2.288464	80.56
11	3	174	1.680321	0.02417468	69.51
12	3	174	21.12819	0.1484985	142.28
13	3	174	4272219	27217.55	156.97

From the Table-5.8 analysis of F-Ratio by k-means at K=4 wine dataset (D4) by k-means represented that the estimated value of F-Ratio which is greater than tabled value of 2.0838 at 5% with d. f. used df1= 3, df2= 174/ or INF.. Hence, estimated values of F-Ratio are > F-Rato of tabled value (2.0838). It is not support to the Null-Hypothesis has differences of data means. We, conclude that statistical significant of attribute- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 for support to clusters creations.

5.2.2.2 K-Means Algorithm Hybridized via PCA

Heart Disease Dataset (D1)

From the Table-5.6 analysis of F-Ratio by k-means at K=4 heart disease dataset (D1) by k-means using hybridized PCA represented that the estimated value of F-Ratio which is greater than tabled value of 2.0838 at 5% with d. f. used df1= 3, df2= 293/or INF.

Table 5.9 Analysis of F-Ratio or F-Score by k-means via PCA at K=4 Dataset (D1)

Attribute	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
1	3	293	3469.018	47.2173	73.47
2	3	293	3.123318	0.1897602	16.46
3	3	293	18.03772	0.7557991	23.87
4	3	293	4021.135	277.5758	14.49
5	3	293	14395.83	1701.051	8.46
6	3	293	11.60596	0.006677548	1738.06
7	3	293	10.91227	0.8882588	12.29
8	3	293	15887.33	369.0353	43.05
9	3	293	15.50666	0.06416339	241.67

Hence, estimated values of F-Ratio are > F-Rato of tabled value (2.0838).It is not support to the Null-Hypothesis has differences of data means. We, conclude that statistical significant of attribute- 1, 2, 3, 4, 5, 6, 7, 8, and 9 for support to the clusters creations.

Iris Dataset (D3)

From the Table-5.10 analysis of F-Ratio by k-means at K=4 iris dataset (D3) by k-means using hybridized PCA represented that the estimated value of F-Ratio which is greater than tabled value of 2.0838 at 5% with d. f. used df1= 3, df2= 146/or INF. Hence, estimated values of F-Ratio are > F-Rato of tabled value (2.0838). It is not support to the Null-Hypothesis has differences of data means. We, conclude that statistical significant of attribute- 1, 2, and 3 for support to clusters creations.

Table 5.10 Analysis of F-Ratio or F-Score by k-means via PCA at K =4 Dataset (D3)

Attributes	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
1	3	146	6.287773	0.06266631	100.34
2	3	146	142.577	0.2474842	576.11
3	3	146	26.11561	0.0577596	452.14

5.5.3: Comparative Analysis of Average Silhouette

Further, comparative statistical analysis of cluster by Fuzzy approach and its clusters concept via PCA is represented in Table 5.11 to Table 5.16 respectively. The average Silhouette values of four datasets namely namely heart disease (D1), user knowledge modeling (D2), iris (D3), and wine (D4).

5.5.3.1 Fuzzy Approach

The silhouette value close to high then object is matched to itself cluster and very poorly matched to nearby clusters. Further, given maximum object have closed to high, then

formation of cluster is suitable (Table-5.13 and Table-5.14) otherwise have few or more clusters (Table 5.11 and Table 5.12).

Heart Disease Dataset (D1)

Table 5.11 Analysis of metrics by fuzzy at K=4 Dataset D1

Number of Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	1336.969288	0.172883	0.5013	0.0025	0.4602	0.9204
3	887.187572	0.143425	0.3422	0.0133	0.5547	0.8320
4	665.798047	0.074896	0.2558	0.0077	0.6614	0.8819
5	532.288185	0.058233	0.2056	0.0070	0.7160	0.8950

User Knowledge Modeling Dataset (D2)

Table 5.12 Analysis of metrics by fuzzy at K=4 Dataset D2

Number of Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	20.507428	0.116382	0.5047	0.0095	0.4211	0.8422
3	13.596835	0.108502	0.3337	0.0006	0.6459	0.9688
4	10.196604	-0.006689	0.2502	0.0003	0.7376	0.9835
5	8.159669	0.042512	0.2003	0.0004	0.7848	0.9810

Iris Datasets (D3)

Table 5.13 Analysis of metrics by fuzzy at K=4 Dataset D3

Number of Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	36.867300	0.677035	0.7109	0.4219	0.0822	0.1644
3	22.633965	0.522717	0.5841	0.3762	0.1650	0.2475
4	17.053815	0.477877	0.4409	0.2545	0.2938	0.3917
5	13.711244	0.306845	0.3407	0.1758	0.4315	0.5394

Wine Dataset (D4)

Table 5.14 Analysis of metrics by fuzzy at K=4 Dataset D4

Number of Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	227.925300	0.432712	0.5310	0.0619	0.3026	0.6053
3	145.815713	0.356429	0.4170	0.1255	0.3411	0.5117
4	109.453281	0.251874	0.3134	0.0845	0.4711	0.6281
5	87.562927	0.199430	0.2521	0.0651	0.5572	0.6966

5.5.3.2 Fuzzy Approach Hybridized via PCA

The silhouette value close to high then object is matched to itself cluster and very poorly matched to nearby clusters. Further, given maximum object have closed to high, then formation of cluster is suitable (in Table-5.16) otherwise have few or more clusters (in Table-5.15).

Table 5.15 Analysis of metrics by fuzzy via PCA at K=4 Dataset D1

Number of Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	1664.089692	0.168686	0.5013	0.0025	0.4599	0.9198
3	1104.028257	0.150270	0.3425	0.0137	0.5532	0.8298
4	828.554707	0.075607	0.2560	0.0080	0.6602	0.8803
5	662.380091	0.064685	0.2058	0.0072	0.7148	0.8936

Table 5.16 Analysis of metrics by fuzzy via PCA at K=4 Dataset D3

Number of Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	35.548398	0.722930	0.7357	0.4713	0.0686	0.1372
3	20.897879	0.590457	0.6272	0.4408	0.1266	0.1899
4	15.774233	0.504718	0.4828	0.3104	0.2510	0.3347
5	12.761452	0.343379	0.3721	0.2152	0.3791	0.4738

5.5.4 Comparative analysis of centroids

Statistical analysis the separation between the initial centroid results in Table-5.17 and represent in Figure-5.5, and Figure-5.6 of heart disease, iris, and wine dataset respectively.

Table 5.17 Comparative analysis of centroids

Dataset		Measure Min_distance between Initial Centroids						
		Various level of K Clusters						
		2	3	4	5	6	7	8
Dataset D1	Existing Method	193.811	130.603	102.359	91.985	83.294	69.303	70.235
	Proposed Method	131.031	65.023	41.304	31.357	23.087	19.075	17.170
Dataset D2	Existing Method	1.320	1.156	0.992	0.957	0.950	0.768	0.705
	Proposed Method	1.181	0.948	0.750	0.512	0.484	0.433	0.351
Dataset D3	Existing Method	06.759	03.824	1.942	1.338	1.288	1.336	1.122
	Proposed Method	05.954	02.520	2.04	1.118	1.020	0.781	0.762
Dataset D4	Existing Method	1402.192	895.845	370.027	327.55	233.077	165.045	140.306
	Proposed Method	92.143	895.844	370.03	327.45	223.03	165.024	140.291

In the above Figure-5.5, and Figure-5.6 show the minimum distance of original centroids of cluster at K=2, 3, 4, 5, 6, 7, and 8 of four datasets is decrease from K=2 to K=8. The proposed method is enhanced than the existing algorithm based on least separation between the original centroids convergence achieved due to smaller changes in cluster centroids.

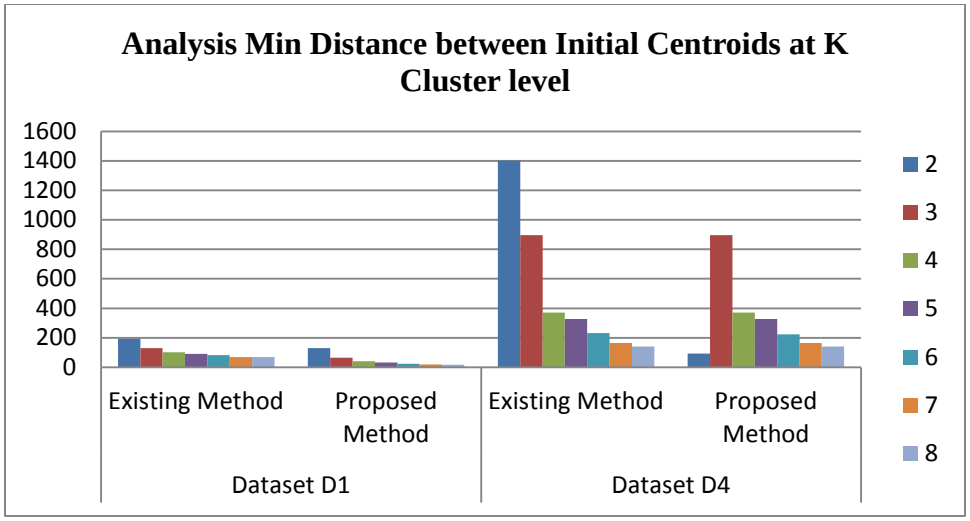


Figure 5.5 Analysis of min distance between initial centroid of dataset D1and D4

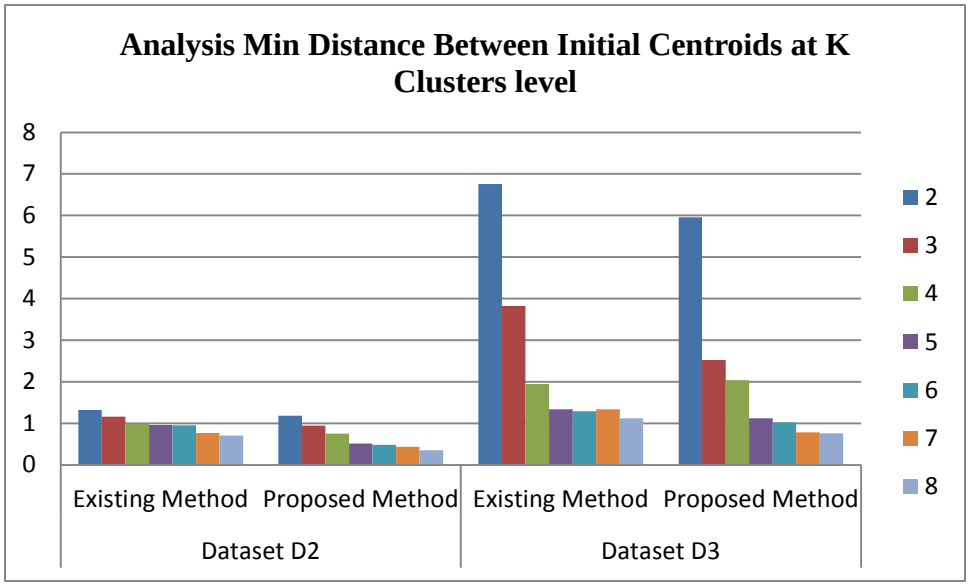


Figure 5.6 Analysis of min distance between initial centroid of dataset D2 and D3

5.6 Summary

In chapter are worked out the lessening reduction component and its statistical analysis of performance with PCA using SPSS statistics 17.0 tool. The dataset of dataset D1 (heart disease), dataset D3 (iris), and dataset D4 (wine) are simulated on the reduction component tool. It is reducing the dimension of datasets by a threshold value on the measured eigenvalues. Statistical method is more significant and used successfully for partitioning the large dimensionality for datasets and helpful for validating of clustering performance. In this chapter, show comparative study for existing algorithm with proposed procedure at cluster level. If the cluster level increases then the small distance between initial centroid deceases and gets a well-defined cluster. And also comparative statistical cluster analysis by K-Means, Fuzzy approach concept applies on four dataset namely heart disease, user knowledge modeling, iris, and wine dataset.

This concept is to categorize the appropriate causes for disease and support treatment of heart ailment, and also relevant information extract of wine dataset.

CHAPTER-6

ANALYSIS OF FTNESS WITH GENETIC ALGORITHM (GA)

Chapter-6, comparatively analysis the fitness function of k-means and kfda, used the optima tool of GA (Genetic Algorithm) approach and to measure the fitness for objective function with means.

6.1 Introduction

Growing the research in field of data mining the k-means technique is well accepted extracting of data from datasets with some limitations. The improve drawback of this technique to employed the kernel concepts and resolve the cluster limitation of separability. We suggested an optimum technique that incorporates a "fisher's discriminant analysis" into the kernel of the PSO concept, as well as a genetic algorithm for evaluating fitness functions. The value of fitness function is required to select offspring for the next generation. In the sequence is to reduce the noise and enhance performance of cluster. The genetic algorithm employed to optimize the objective of fitness function on provided the input parameter. The kernel technique is performs more fault identification feature than principle component analysis. Getting is more benefited by this method like such that fitness value, fitness scaling with score and average distance between individual. In this chapter discuss the comparison analysis with objective function of k-means and kernel fisher's discriminant analysis in the domain of large dataset.

The idea of genetic algorithm explains the computation of numerical easily solve like a mathematical problem reported (Hermawanto, 2013). There are available the several techniques for mining of feature and arrangement of multivariate dataset. The LDA analysis is a statistical method for ordering based on eigenvalues. There are well recognized procedure of classification like as principle component analysis, fisher's

discriminant analysis, particle least square, and discriminant particle least square (Al Malki, et. al., 2016; Kemsley, 1966). The principal component analysis work very crucial role classification of dataset, but fisher's analysis discriminant analysis produced to better outcome as compared than PCA. KPCA is nonlinear result of PCA and similar manner KFDA is same as of the FDA (Sayed et. al., 2009; Oujezsky and Horvath, 2018). In the linearly nature of dataset classification occur not more good, but kernel idea handles the non-linearity problem.

Kernel concept to handles the nonlinear problems this study to overcome the few difficulty using of optimization technique for evolutionary computing. In this chapter discussed the methodology of comparatively analysis of performance of fitness function.

6.2 Back Ground

6.2.1 Kernel Trick

The consider a sample data set $A = \{ a_1, a_2, a_3, \dots\dots, a_n \}$ and every member of data point belonging in domain fieldR^N, where R indicate the set of domain, $\forall\, n \in N, and\ i = 1, 2, 3, \dots\dots. n$. Nonlinear mapping mathematically represented as the following manner

$$\emptyset: R^N \to F, and\ a = \emptyset(a) \qquad (6.1)$$

Where, $\emptyset\ \ is\ non\ linear\ mapping\ function$, and symbol F is indicate the variety of function (feature space).

The model recognized of the PSO kernel parameter based on FDA instruction. There two type of characteristic samples in feature space F.

The first dataset $A_1 = \{ a_{11}, a_{12}, a_{13}, \dots\dots, a_{1i} \}$, and

Second dataset$A_2 = \{ a_{21}, a_{22}, a_{23}, \dots\dots, a_{2j} \}$,

Where $i = 1, 2, 3, \dots\dots. n_1$, $i = 1, 2, 3, \dots\dots. n_2$

The mean vector of two type feature space in F is, Mean vector of one feature Space (F1) is

$$\mu_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \emptyset(\, a_{1i}) \qquad (6.2)$$

Mean vector of second feature space (F2) is

$$\mu_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \emptyset(\, a_{2j}) \tag{6.3}$$

Determine the square distance between mean vector of two vector spaces F1 and F2

$$\text{Square Distance} = |\, \mu_1 - \mu_2 \,|^2 \tag{6.4}$$

Determine the dispersion within the sample of feature space F1

$$\text{Disp\_w\_f1} = \sum_{i=1}^{n_1} |\emptyset(a_{1i}) - \mu_1\,|^2 \tag{6.5}$$

$$\text{Disp\_w\_f2} = \sum_{j=1}^{n_2} |\emptyset\big(a_{2j}\big) - \mu_2|^{transpose} \tag{6.6}$$

Arrange the particle swarm optimization fitness function as according to fisher's discriminant criteria from the equation (6.4), (6.5) and (6.6) respectively

$$Fitness\ function = \frac{(\text{Disp\_w\_f1} + \text{Disp\_w\_f2})}{\text{Square Distance of mean vector}} \tag{6.7}$$

This fitness function is useful for separation of between max and min of classifications. This fitness function is varying on the set value of inertia from low to high. So that, this is more fit for other kernel trick.

6.2.2 Genetic Algorithm

The details about this algorithm is available can found to proposed the techniques by researcher Holland in 1975, and by Goldberg in 1989.

There are discussed the operators of genetic algorithm as the follows:

SELECTION OPERATOR: In this concept offer the preference of higher fitness value of the chromosome allowed to go for next generation.

CROSSOVER OPERATOR: In this concept matching it individuals. The selecting of two individual by the theory of selection operator and apply crossover operator can be rearranged at the site creating the new individual known as the offspring.

MUTATION OPERATOR: In this concept inserts the genes in the offspring getting by crossover operator to maintain the size of a population to keep avoid the early convergence.

The procedures of generic algorithm summarize as follows:

Step 1: Set to initialize the value of population with random values.

Step 2: Fitness function evaluate on the no. of population

Step 3: Until the convergence repeat

I: Select the descendants from the population

II: Crossover and generate the new offspring

III: Apply the mutation of new offspring

IV: Determine the fitness of new offspring

6.3 Methodology

The objective function is getting for associated to "kernel trick" of fisher's discriminant analysis. Further, in this function are required for the parameter of the performance like as means and square of the distance. Consider data size is including the two factor firstly row or instances and secondly columns or attribute. The index of performance defined from equation (8) and compare with objective function of K-Means.

Procedural steps as:

Function kfdaff= kernel for vectors(x1, x2)

1. Begin
2. For i=1 to $n_1$do
3. For j=1 to n_2 do
4. $E1 = \frac{1}{n_1} * \frac{1}{n_2} K(x_{1i}, x_{1j})$
5. $E2 = 2 * \frac{1}{n_1} * \frac{1}{n_2} K(x_{1i}, x_{2j})3$
6. E3== $\frac{1}{n_2} K(x_{2i}, x_{2j})$
7. SDMV= E1-E2+E3
8. For i=1 to n1
9. E4= $K(x_{1i}, x_{1i})$

10. For j=1 to n2 do
11. E5=$\frac{1}{n_1} K(x_{1i}, x_{1j})$
12. Disp_w_f1=E4-E5
13. For i=1 to n1
14. For j=1 to n2 do
15. E6=$\frac{1}{n_2} K(x_{2i}, x_{2j})$
16. E7= $K(x_{2j}, x_{2j})$
17. Disp_w_f2=E7-E6
18. A= Disp_w_f1
19. B= Disp_w_f2
20. Z= A+B
21. KFDAFF=Z/SDMV
22. END

Where

$$K(x_i, x_j) = \emptyset(x_{1i})^T \emptyset(x_{2j}) = (x_{i1}.\text{^}2x_{2j}.\text{^}2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}.\text{^}2\, x_{j2}.\text{^}2)$$

The procedure of simulate it, defined objective function by genetic algorithm for using the OPTIMTOOL for optimizing. There are few steps as follows:

Result of any optimization problem to convert into the binary stream is called a chromosome.

Create the initial random number of chromosomes.

1. To calculate the chromosomes by defining the objective function with Score.
2. To set its current population for producing the child.
3. Evaluate the crossover performance
4. Evaluate the performance of mutation.
5. Repeat the step 3 to 4 until the exit

Here mentioned the exit criteria for pick up that produced the no. of generations to be reached maximum (of a population) value the parameter set of option to measuring performance from Table- 6.1.

Table 6.1: Set the Optional Parameters Measuring the Performance

Parameters	Range/ value
Set Mutation	0.8
Generation of Random Number	[0, 1]
Use default Population	20
Size of Population	1000
Variables	2
Set size of Variable	10
Type of Population	Double Vector

6.4 Experimental Results

In this chapter, compare the fitness objective function of k-means (traditional) and proposed KFDAFF objective function by genetic algorithm. The experimental set up of determine fitness function in problem solver Genetic Algorithm(GA), and fitness function defined @ problem fitness and set variables.

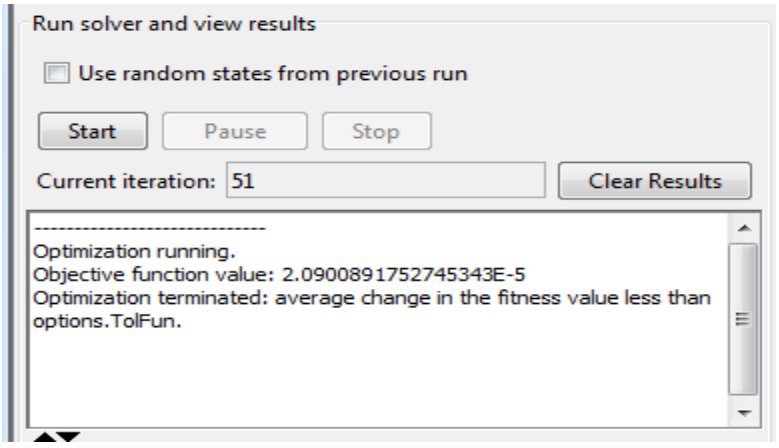


Figure 6.1 For Objective function of KFDA

For objective function of Kernel FDA

Fitness function= @kalam_fitness, number of iteration is 51 at variables on the run solver view the result as in the Figure- 6.1. The optimization of running as:

Value of objective function: 2.0900891752745343E-5

Optimization Terminated: Average change in fitness value less than option.

For Objective function of k-means

Fitness function= @kalam1_fitness, number of iteration is 51 at variables on the run solver view the result as following the Figure-6.2.

The optimization of running,

Value of objective function: 3.38184024594581E-7
Optimization Terminated: Average change in fitness value less than option

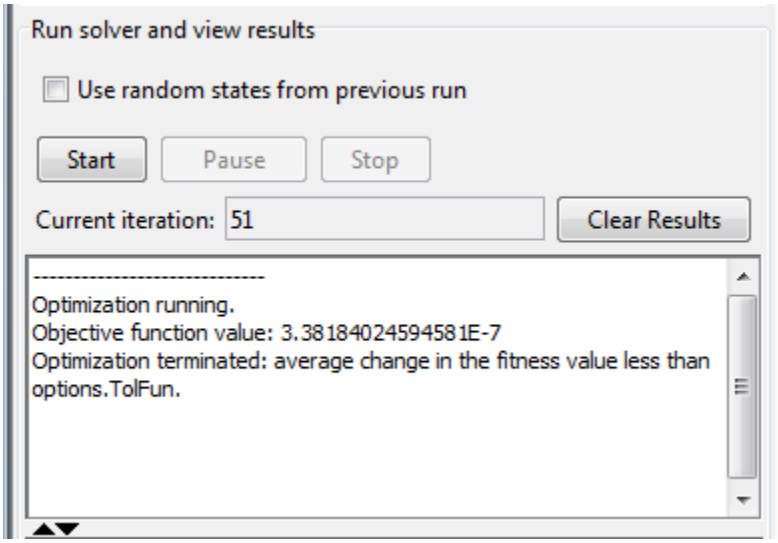


Figure 6.2 For Objective function of k-means

Therefore, we have set the fixed type of population for some attribute of dataset: double size of which represent as uniformly, operator crossover mutation set at 0.8 vector, size (population): default value at 20, set the initial random generation rang [0, 1], and fitness scaling(R). In addition, select the stochastic function, and constraints mutations dependent based on fitness function but where the cross over function is scattered.

Fitness Value and Mean

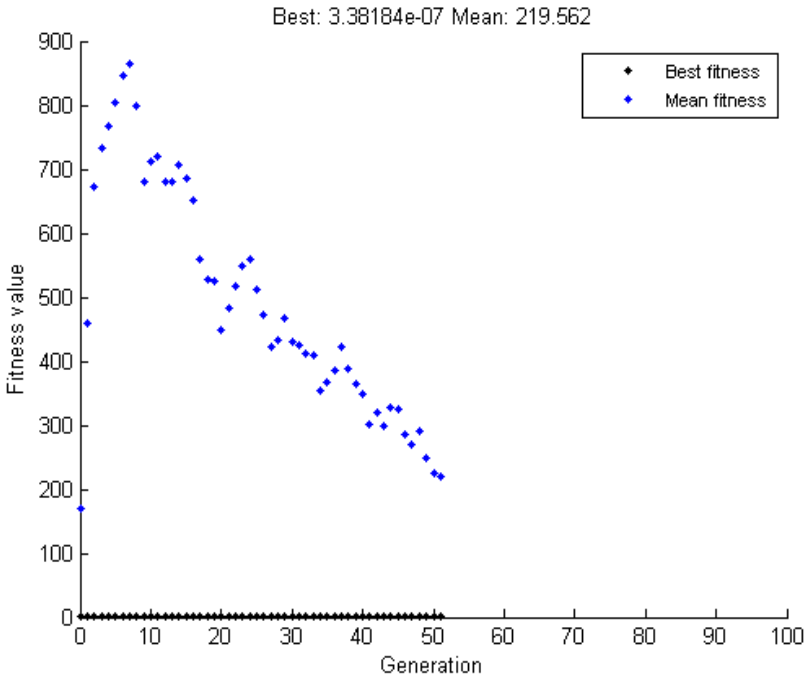


Figure 6.3 Fitness of k-means Objective function

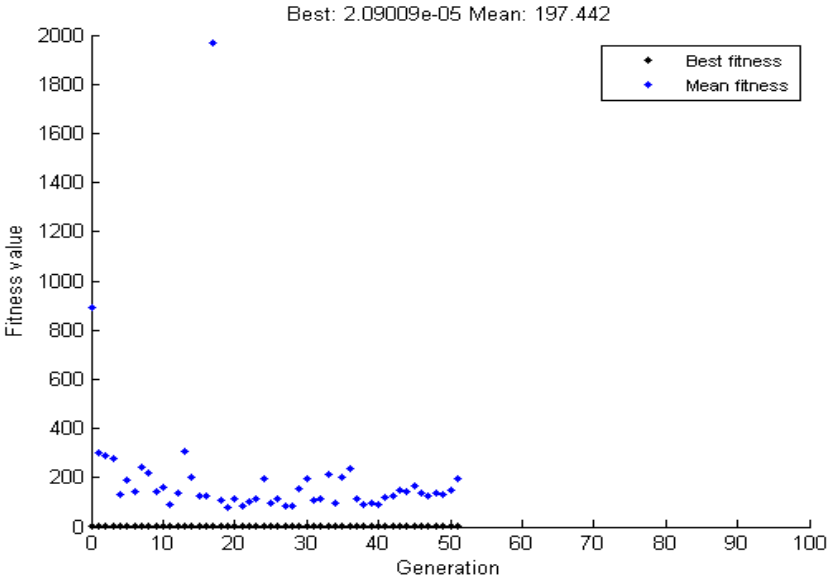


Figure 6.4 Fitness of KFDAs Objective function

Table 6.2 Comparative Analysis of Various Attribute

Criteria of K-Means objective function	**Criteria of KFDAFF objective function**
1.Mean =219.562	**1**.Mean =197.442
2. Fitness scaling(Expectation) =30	**2.** Fitness scaling(Expectation) =35
3. Every individual (fitness) at 800 lies between 36 and 39	**3**.Every individual (fitness) at 800 above 500
4.Stopping % criteria met S(G) is below 80	**4**.Stopping % criteria met S(G) is above 50
5.Stopping % criteria met S(T) is above 50	**5.**Stopping % criteria met S(T) is 20
6.Average distance between individual approximate above 01(one)	**6**.Average distance between individual approximate above 02(two)

There are the parameter n1 and n2 both where set at 10. The result of performance show an objective function of k-means and function of KFDA according in figure- 6.3, and figure- 6.4 mentioned that the fitness of objective function in term of best fit and means. And also determine the rest value fitness scaling, fitness of every individuals, stopping criteria, and average distance between individual respectively. Analyze of comparatively study the KFDA objective function is more prefer than objective function of K-Means in Table-6.2

6.5 Summary

Genetic Algorithm (GA) applies for simulation results of proposed KFDAFF algorithm. The KFDAFF algorithm is superior and more significant as related to other procedures. This performance is more favorable in the ordering of datasets. In this chapter, it is mentioned that fitness value of an objective function like in terms of best fit and means, stopping criteria, and average distance between individual of simulation process. The comparative analysis criteria of objective function proposed KFDAFF algorithm is lesser than an objective function of K-Means. The exit criteria is the selection when the no. of generation produced touches the maximum (of population) value.

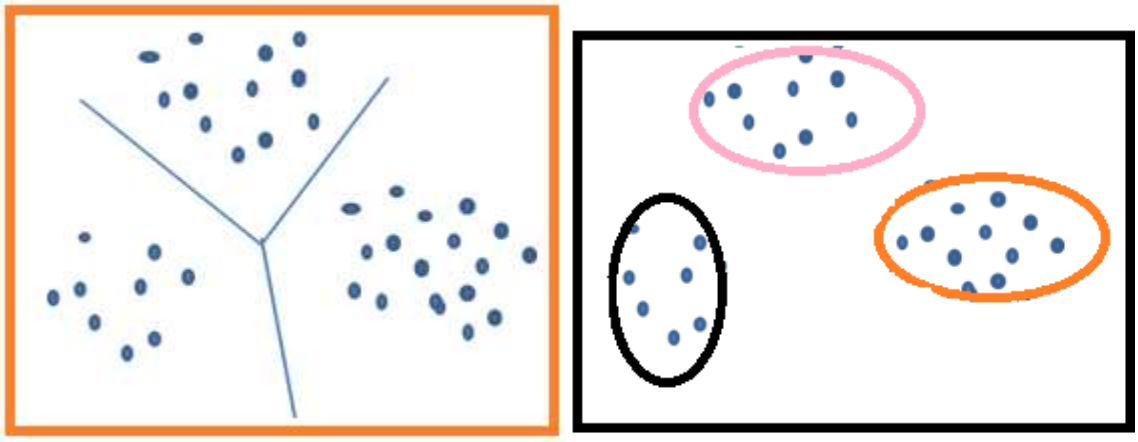
CHAPTER-7

CLUSTERING USING MACHINE LEARNING RBFNN

In this chapter, we applied the concept of radial basis neural network and pattern recognition to measure the sum of squared error and mean squared error for using the real world dataset.

7.1 Introduction

The concept of radial basis neural network has consisted of three layers namely the input layer, output layer and hidden layer. Its neural net is strictly for limitation has a single hidden layer. This hidden layer is also called a feature vector. Training in this neural network is very fast, very easily explained function of all nodes in a single hidden layer, and also take more time as compared to MLP. In figure 7.1, illustrate the clusters difference between MLP vs. RBF.



MLP **RBF**

Figure 7.1 Show the clusters difference between MLP vs. RBF

There is illustrate the Gaussian Radial Function (GRF)

$$\varphi(r) = exp^{\left(-\frac{r^2}{2\sigma^2}\right)}$$

In this neural network idea have two phases as follows:

- In the first phase: The hidden layer trained using concept of the back- propagation, required the method of curve fitting, find out the variance (σ) and receptors.
- In the second phase (training phase): In this phase update the weight vectors between a single hidden and output layer.

For all nodes at a single hidden layer represent every transformation using basis function, and also it function satisfies the nonlinear separability [101].

7.2 Back Ground

Therefore, the first stage of training is done by a clustering algorithm. Define the number of cluster centers for necessities and required. Furthermore, consists N cluster samples or observations into M clusters (N > M). Outputs "clusters" are the "receptors". Every receptor, determined variance mentioned that $\frac{1}{N} * |\ Data\ sample\ X\ - t\ |^n, set\ n = 2$ using Euclidian distance for nearby for all cluster samples. The explanation of the first training phase is that the "feature" vector is anticipated onto the converted space.

Wetterschereck et. al., (1992) they worked in to improve the RBFN by pseudo-inverse and unsupervised learning (k-means) method in detail [102], Chun-Tao et. al., (2009) authors explore the neural network concept by optimize method particle swarm optimization in detail [103], application of the least mean Square technique and fuzzy c-means used in neural net by (Ziyang, et al., 2008)[105], De Lacerda et. al., (1994) Genetic Algorithms[104]. Dash et. al., (2016), authors discussed the progress of radial basis function networks using and emphasized on some kind of learning concept like that the novel Kernels[106].

Applied four data like that Heart Dataset, User knowledge modeling, Iris, and Wine Dataset using RBFN and analysis training SSE.

Defined the K-Means

Set the function of k-means

K-means (Data matrix X, Means for K, D)

7.3 Methodology

There are describe the idea of RBFN

Function of Normalization \\ MIN-MAX SCALING INTO [0, 1]

\\ z-score make the conversion of data center is zero scale into range [-1, 1]

F (T) = function of normalization (Dataset X, D)

Begin

1. Begin

2. IF Margined is greater than 1 then

$$X_{-min} = \min(X)$$

$$X_{-max} = \max(X)$$

$$T = \frac{(X - repmat(X_{-min},\ size\ (X,1),1)}{repmat((X_{-max} - X_{-min}),\ size\ (X,1),1)}$$

3. Find the means of dataset X

4. Determine Variance of dataset X

$$T = \frac{(X - repmat(X_{-means},\ size\ (X,1),1)}{repmat(standard\ deviation\ of\ dataset\ X,\ size\ (X,1),1)}$$

5. End

\\ Create the kernel function

Create the kernel function f(k)= Kernel(data_matrix, centre)

Begin

\\ Set loop

1. For I=1 to N do \\ N= size(dataset X, 1)
2. For J=1 to M do \\ M=size(consists center of the dataset X, 1)
3. K(I, J)= EXP(-NORM(Dataset X, Center of dataset X))\\ Exponential function of Kernel

End

End

Measure the training with target using RBFN Kernel

1. Set no. of classes and subject
2. Set the id of class \\ Associated with classes and subject CREATE Target
3. Call K-Means (Dataset X, Means for K, 1)
4. Call Kernel(Dataset X, Consists the center)
5. Set Lamda=Inv(K'*K)*K
6. W=Lamda* set the id class
7. Start training with set of Target
8. Return the result

7.4 Experimental Results

Chapter-7 is focused on RBFN Kernel concept, and proposed algorithms to compare by without clustering algorithm. The computing results of proposed algorithm are better than without k-means and create good clusters.

Heart Dataset (D1)

In figure-7.1 (a), show the case processing summary of total 297(two hundred ninety seven) samples and, reveal of the report 65.7 %(training) and 34.3 %(testing) of 100 percentage sample. The measure of SSE, the best performance is mentioned in figure-7.1(b) of Heart disease dataset (D1), and also represent in the relative error of training is 0.896, and relativity error of testing is 1.22 determined by RBF neural network.

Case Processing Summary

		N	Percent
Sample	Training	195	65.7%
	Testing	102	34.3%
Valid		297	100.0%
Excluded		0	
Total		297	

Figure 7.2 (a) Show the Case summary of dataset (D1)

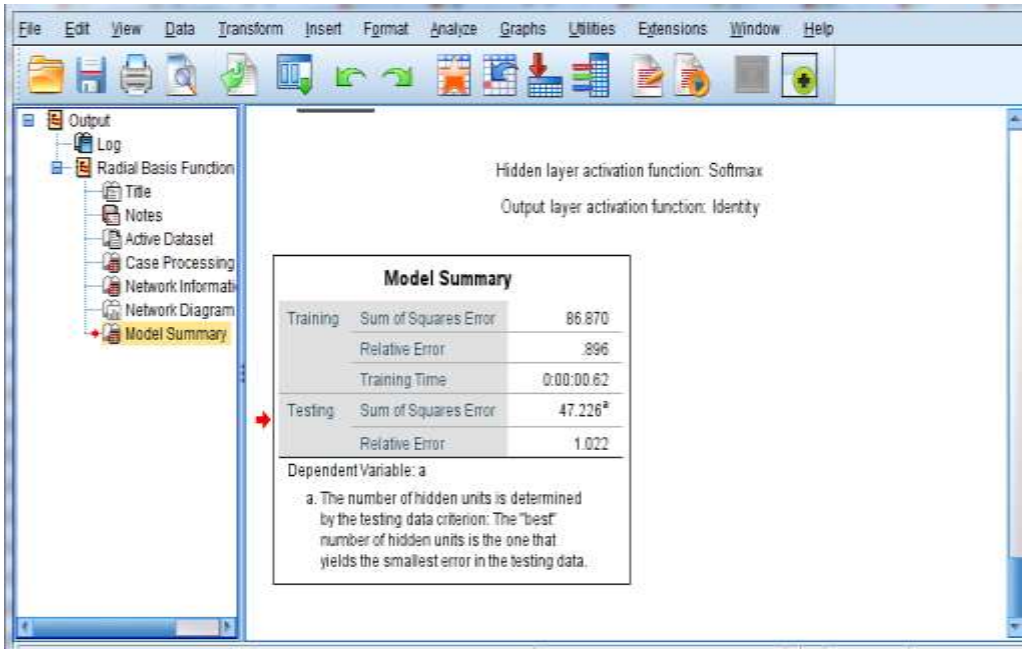


Figure 7.2 (b) Show the SSE summary of dataset (D1)

User Knowledge Dataset (D2)

In figure-7.2 (a), show the case processing summary of total 258(two hundred fifty eight) samples and reveal the report of 84.7% (training) and 15.3% (testing) of 100 percentage sample. The measure of SSE, the best performance is mentioned in figure- 7.2(b) of user knowledge modeling dataset (D2), and also represent in the relative error of training is 0.975, and relativity error of testing is 1.00 determined by RBF neural network.

Case Processing Summary

		N	Percent
Sample	Training	183	84.7%
	Testing	33	15.3%
Valid		216	100.0%
Excluded		42	
Total		258	

Figure 7.3 (a) Show the Case summary of dataset (D2)

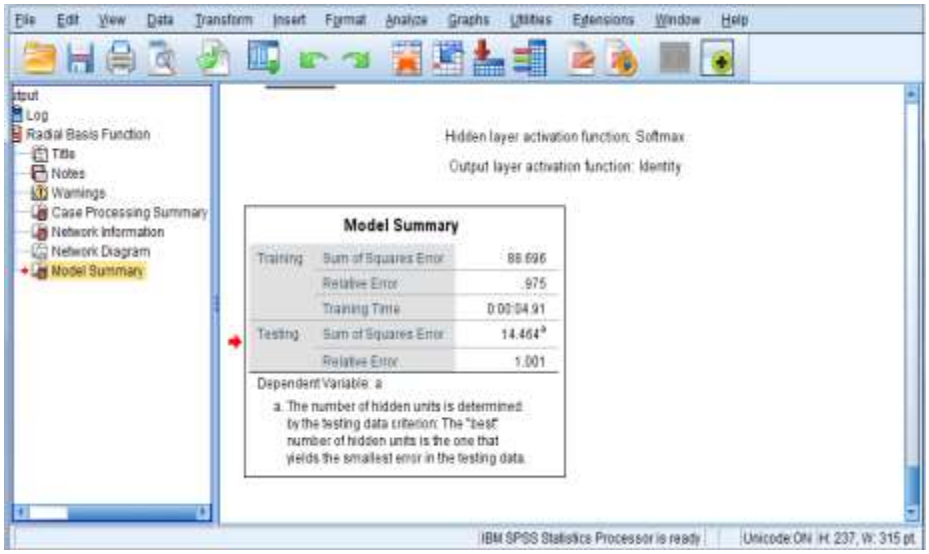


Figure 7.3 (b) Show the SSE summary of dataset (D2)

Iris Dataset (D3)

In figure-7.3 (a), show the case processing summary of total 150 (one hundred fifty) samples and reveal the report 72% (training) and 28% (testing) of 100 percentage of sample. The measure of SSE, the best performance is mentioned in figure- 7.3(b) of iris dataset (D3), and also represent in the relative error of training is 0.507, and relativity error of testing is 0.503 determined by RBF neural network.

Case Processing Summary

		N	Percent
Sample	Training	103	72.0%
	Testing	40	28.0%
Valid		143	100.0%
Excluded		7	
Total		150	

Figure 7.4 (a) Show the Case summary of dataset (D3)

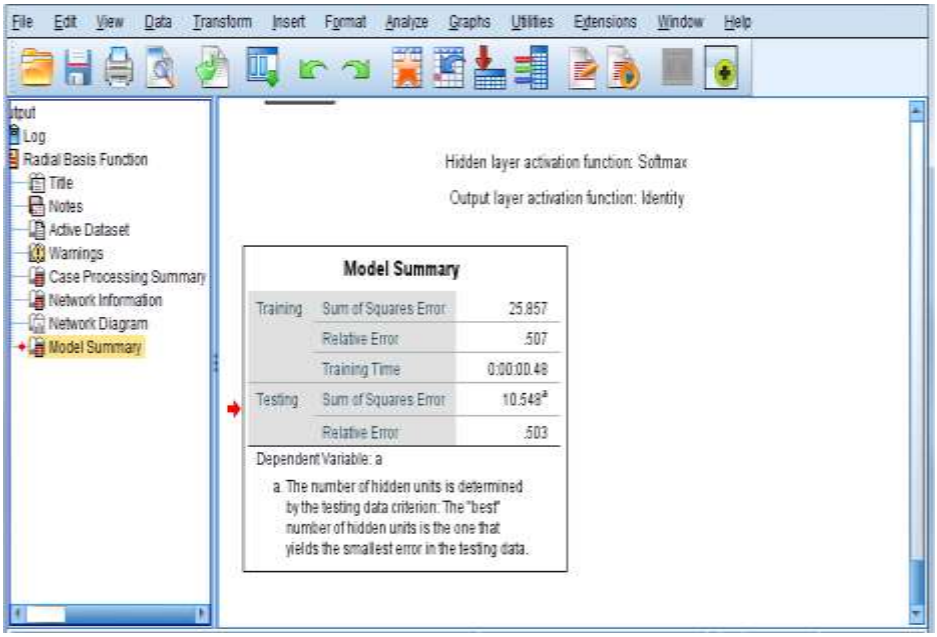


Figure 7.4 (b) Show the SSE summary of dataset (D3)

Wine Dataset (D4)

In figure-7.4 (a), show the case processing summary of total 179(one hundred seventy eight) samples, and reveal the report of 94.8 % (training) and 5.2% (testing) of 100 percentage sample. The measure of SSE, the best performance is mentioned in figure-7.4(b) of wine dataset (D4), and also represent in the relative error of training is 0.690, and relativity error of testing is 0.721 determined by RBF neural network.

Case Processing Summary

		N	Percent
Sample	Training	128	94.8%
	Testing	7	5.2%
Valid		135	100.0%
Excluded		43	
Total		178	

Figure 7.5 (a) Show the Case summary of dataset (D3)

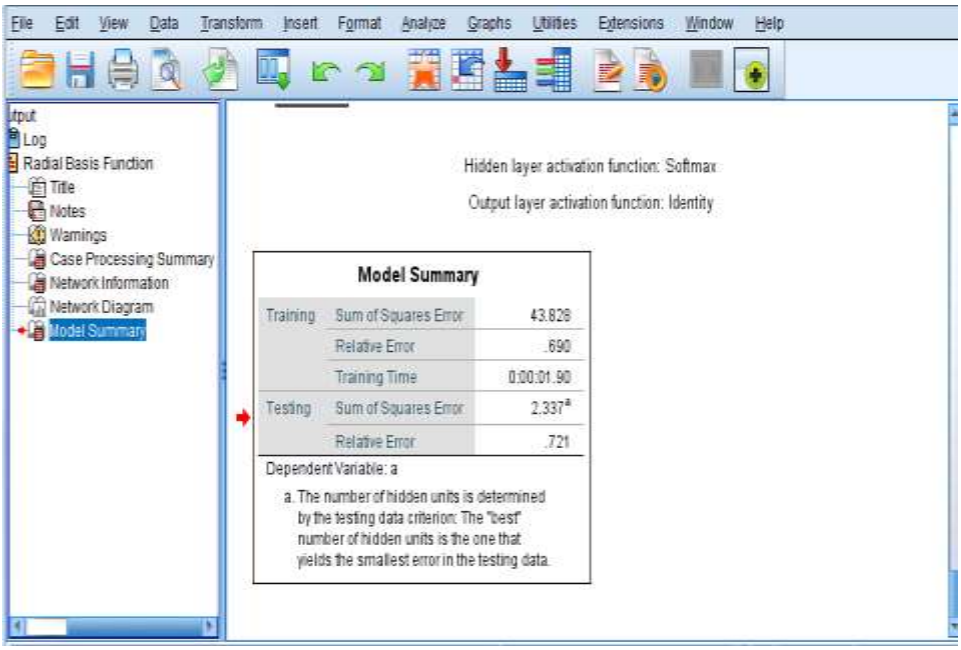


Figure 7.5 (b) Show the SSE summary of dataset (D4)

In Table-7.1, represent the comparative analysis of training and testing of four dataset namely heart disease dataset (D1), user knowledge modeling dataset (D2), iris dataset (D3), and wine dataset (D4).

Table 7.1 Comparative Analysis of SSE T-T of four Dataset

Dataset			
Heart Disease	Training	Sum of Squared Error	86.870
		Relative Error	0.896
		Training Time	0.00.00.62
	Testing	Sum of Squares Error	47.226

		Relative Error	1.022
User Knowledge Modeling	Training	Sum of Squared Error	88.696
		Relative Error	0.975
		Training Time	0.00.04.91
	Testing	Sum of Squares Error	14.464
		Relative Error	1.0201
Iris	Training	Sum of Squared Error	25.857
		Relative Error	0.507
		Training Time	0.00.00.48
	Testing	Sum of Squares Error	10.548
		Relative Error	0.503
Wine	Training	Sum of Squared Error	43.828
		Relative Error	0.690
		Training Time	0.00.01.90
	Testing	Sum of Squares Error	2.337
		Relative Error	0.721

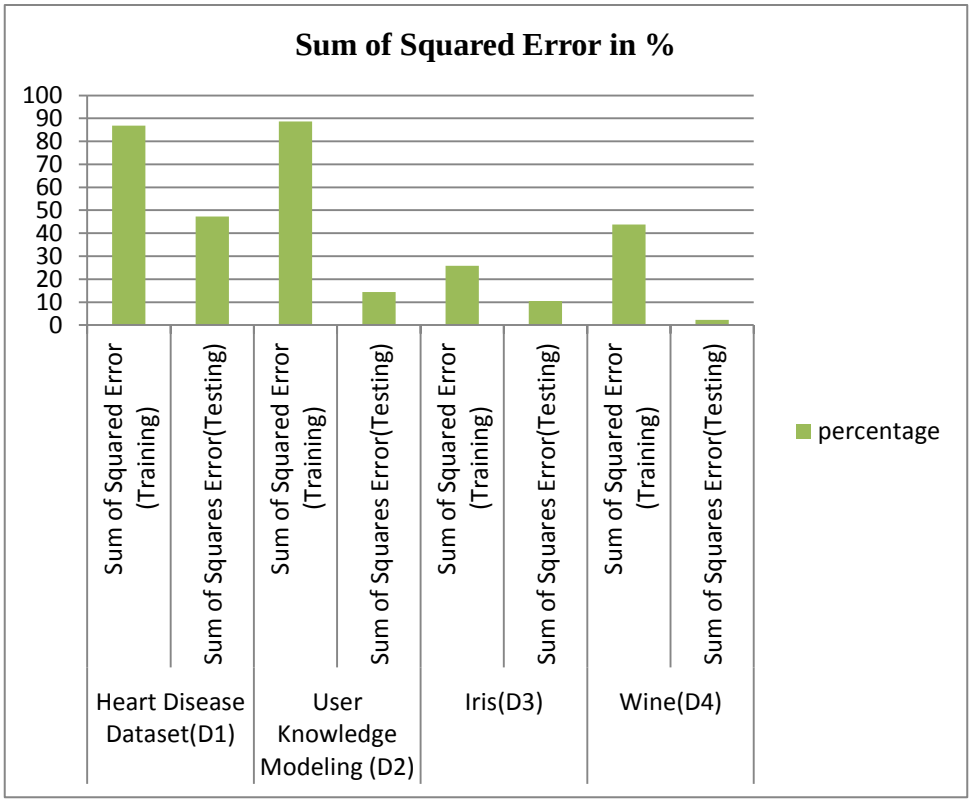


Figure 7.6 Illustrate SSE of four dataset

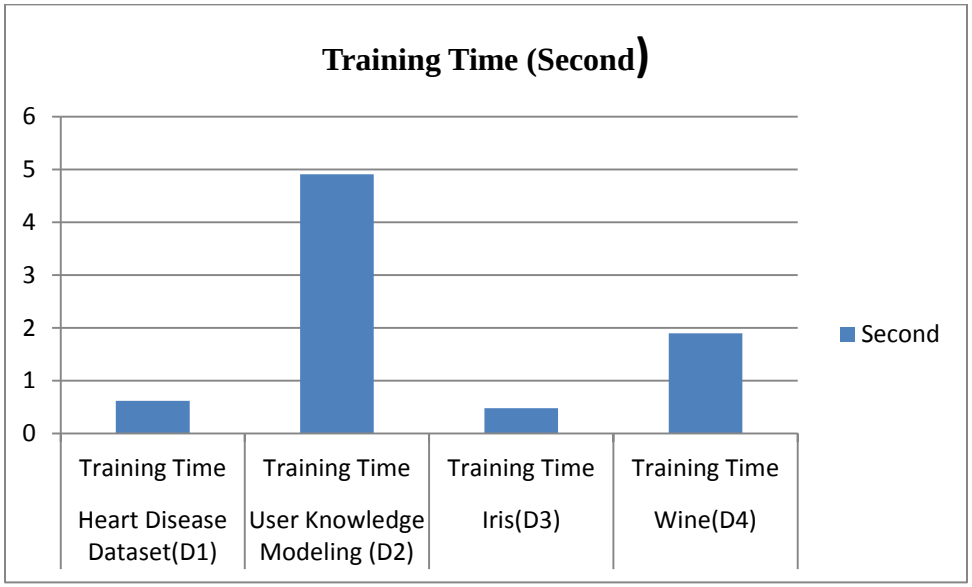


Figure 7.7 Illustrate training time of four dataset

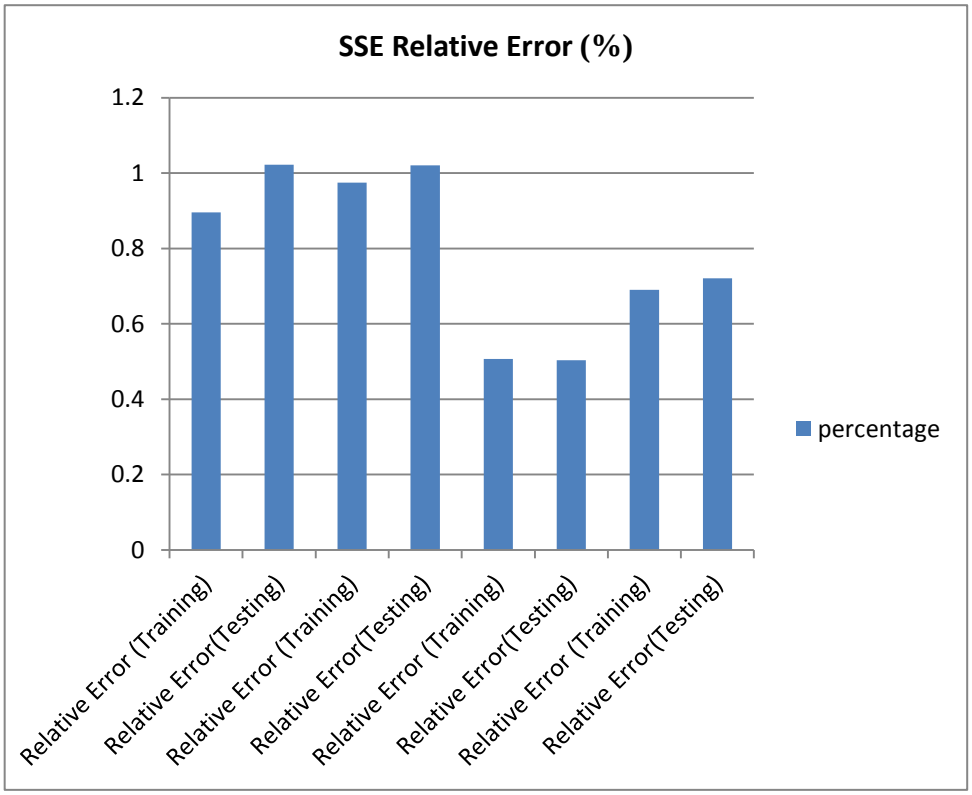


Figure 7.8 Illustrate SSE relative Error of four dataset

The processing summary for Dataset (D1), percentage of training at no. of instances (N=195) is 65.7 %, percentage of testing at no. of instances (N=102) is 34.3 % for total validity of data instances 297 out of 297.

The processing summary for Dataset (D2), percentage of training at no. of instances (N=183) is 84.7%, percentage of testing at no. of instances (N=33) is 15.3% for total validity of data instances 297.

The processing summary for Dataset (D3), percentage of training at no. of instances (N=103) is 72.00%, percentage of testing at no. of instances (N=40) is 28.00% for total validity of data instances 143 out of 150.

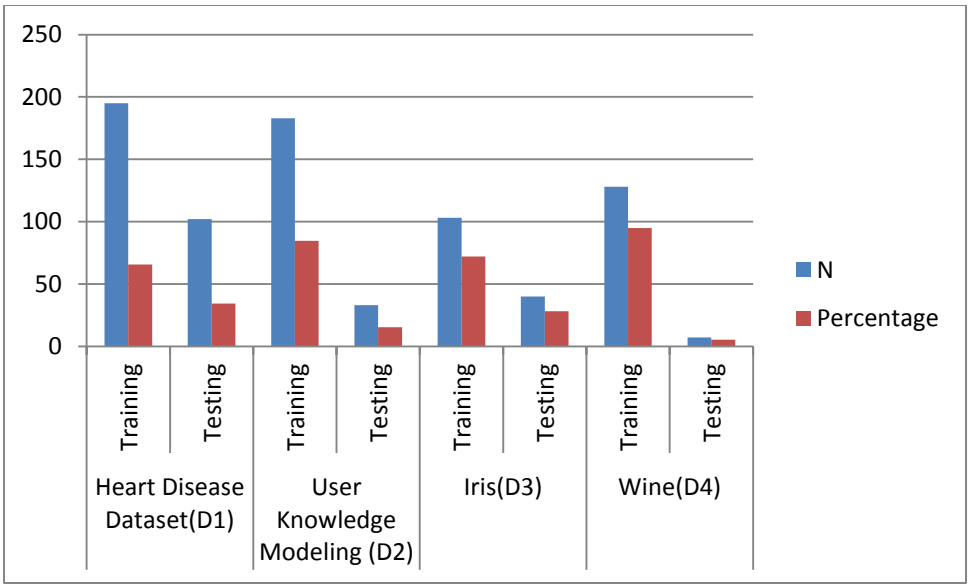


Figure 7.9 illustrate the percentage (%) of testing and training of four dataset

The processing summary for Data (D4), percentage of training at number of instances (N=128) is 94.8%, percentage of testing at number of instances (N=7) is 5.2% for total validity of data instances 135 out of 178.

7.5 Summary

Chapter-7, measure the sum of square error train and testing value using the RBFN kernel concept and achieve a better sum of square testing results based on the proposed work. This chapter, we estimate it metric by this technique on different factual datasets of Iris, Wine, and Heart Disease respectively find minimum sum of squared error.

CHAPTER-8

EXPERIMENTAL RESULTS AND DISCUSSION

In this chapter, we addressed the efficacy of the proposed method based on distinct performance attributes. The research findings are illustrated below.

8.1 Results Comparison for Measure Quality of Cluster

Measure the some metrics like that, intra cluster distance sum of squared error, CPU time , external metrics for four datasets Iris Dataset (D1), Wine Dataset (D2), User Knowledge Modeling (D3), and Heart Disease Dataset (D4). The proposed method and it is hybridized via PCA, PSO are used for the validation of clusters.

And also evaluated the performance is measured based on the external cluster quality metrics Precision, Recall, F-Measure (or F-Score), and Rand Index or Accuracy of four datasets.

The proposed methods are evaluated and it's output is calculated based on metrics. A comparison value of data clustering applies on the same datasets and validates the consequences of clusters.

8.1.1 Sum of Squared Error (SSE)

From Figure 8.1, illustrate the proposed approach offers better value for SSE than the current method. Data sets are draw along the X-axis in this graph and SSE values are draw along the Y-axis.

From the figure 8.1 it is shows that the proposed method has improved accuracy namely as:

Heart disease dataset (D1): Sum of Squared Error (SSE) of min-max method is 3.8% reduced, proposed AIEKM method is 7.6% reduced, and proposed methods hybridize via PCA is 25.5% reduced than k-means respectively.

User Knowledge Modeling dataset (D2): Sum of Squared Error (SSE) of min-max method is 4.1% reduced, proposed AIEKM method is 7.8% reduced, and proposed methods hybridize via PCA is 11.7% reduced than k-means respectively.

Iris dataset (D3): Sum of Squared Error (SSE) of min-max method is 5.9% reduced, proposed AIEKM method is 12.2% reduced, and proposed methods hybridize via PCA is 23.04 % reduced than k-means respectively.

Wine dataset (D4): Sum of Squared Error (SSE) of min-max method is 16.9% reduced, proposed AIEKM method is 18.1% reduced, and proposed methods hybridize via PCA is 23.9% reduced than k-means respectively.
Hence, the proposed method is better than the k-means existing algorithm.

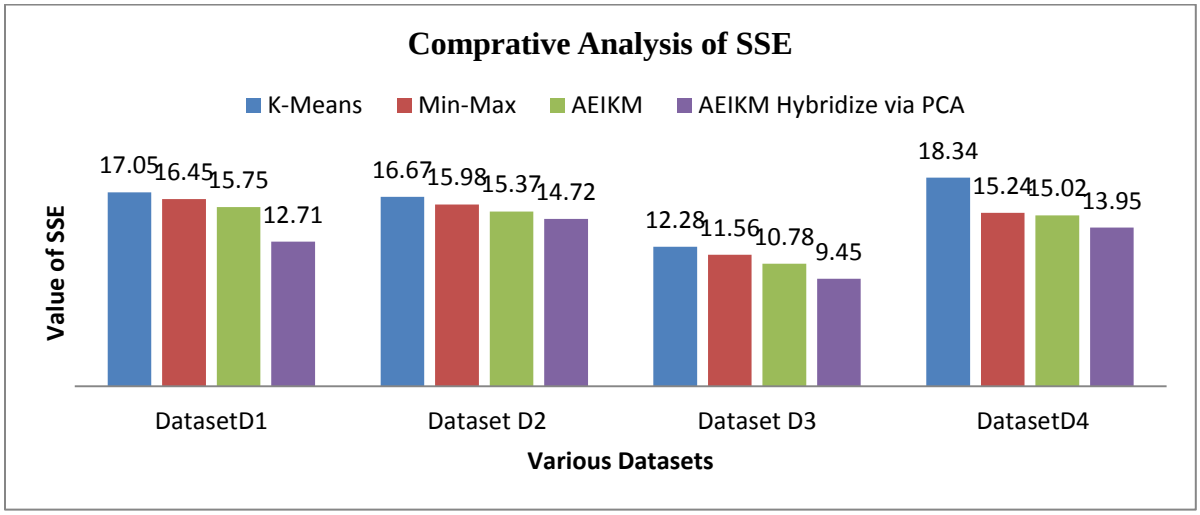


Figure 8.1 Comparative analyses of SSE by different methods

8.1.2 Intra Cluster Distance

From Figure 8.2, illustrate the proposed approach via PCA offers is better value of intra cluster distance than the current method. Data sets are draw along the X-axis methods in this graph and intra -cluster values are draw along the Y-axis.

From the figure 8.2 (a) it is shows that the comparative analysis of intra cluster distances as:

For Heart disease datasets D1: Proposed AEIKM method is 1.2%, 7.6%, 4.9%, 9.5%, 2.6%, 2.4%, and 12.9%, reduced than k-means for at K=2, 3, 4, 5, 6, 7, and 8 clusters respectively . And it proposed method hybridize via PCA at K=4 is 4.8% reduced than k-means.

For Wine Dataset D4: Proposed AEIKM method is 7.8%, 1.3%, 1.4%, 1.3%, 9.2%, and 4.9%, and 2.3% reduced than k-means for at K=2, 3, 4, 5, 6, 7, and 8 clusters respectively.spectively. And it proposed method hybridize via PCA at K=4 is 76% reduced than k-means.

Hence, the proposed method is better than the k-means existing algorithm.

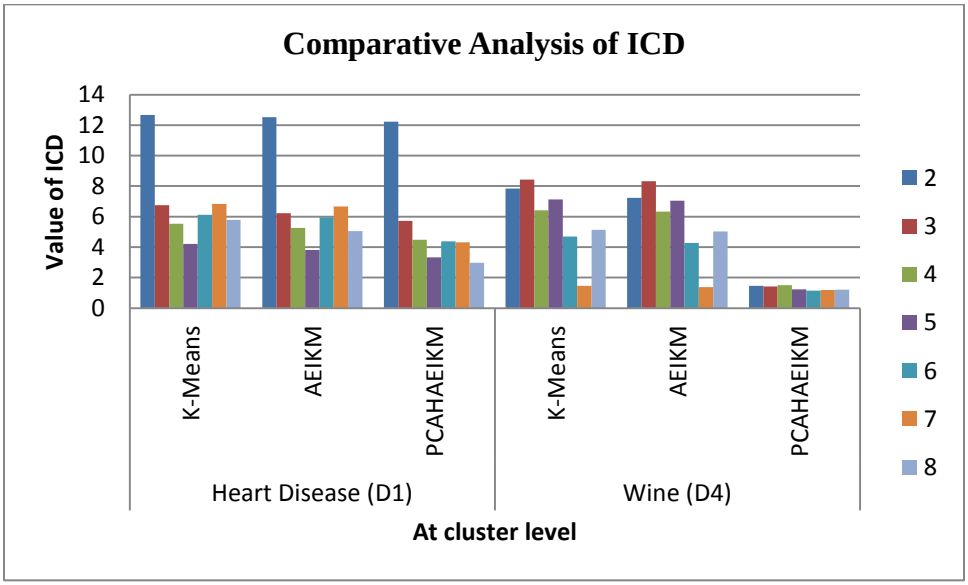


Figure 8.2 (a) Comparative analyses of ICD by methods

From the figure 8.2 (b) it is shows that the comparative analysis of intra cluster distances as:

For datasets D2: Proposed AEIKM method is 6.9%, 5.8%, 6%, 4.6%, 6%, 11.3%, and 10.6%, reduced than k-means for at K=2, 3, 4, 5, 6, 7, and 8 clusters respectively . And it proposed method hybridize via PCA at K=4 is 65% reduced than k-means.

For Dataset D3: Proposed AEIKM method is 4%, 2.12%, 2.11%, 5.8%, 5.9%, and 4.5%, reduced than k-means for at K=2, 3, 4, 5, 6, 7, and 8 clusters respectively.spectively. And it proposed method hybridize via PCA at K=4 is 33.7% reduced than k-means.

Hence, the proposed method is better than the k-means existing algorithm.

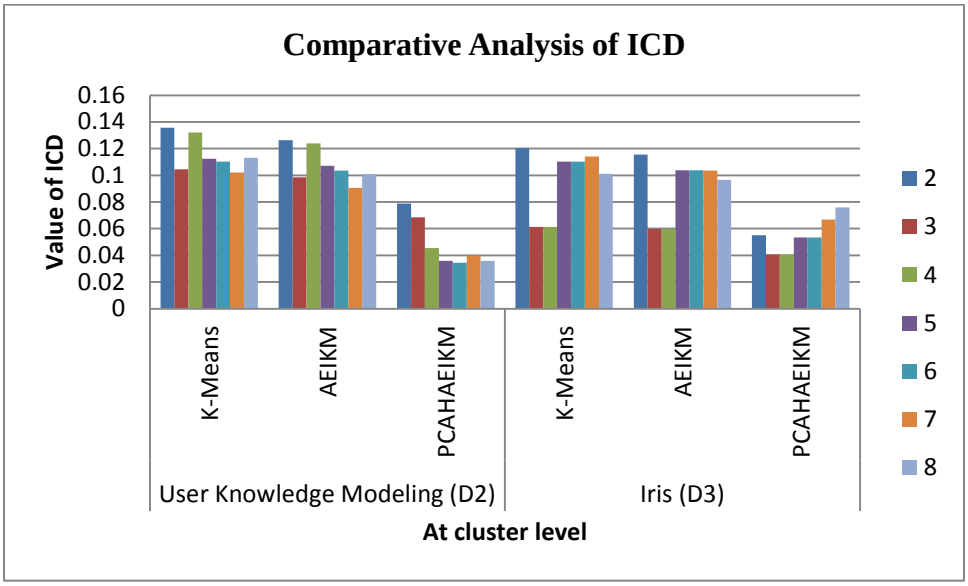


Figure 8.2 (b) Comparative analyses of ICD by methods

8.1.3 Execution Time Consumption by CPU

In Figure- 8.3(a), and Figure-8.3 (b), show the comparative analysis of the execution time in second at k=2, 3, 4, 5, 6, 7, 8 respectively. The execution time of CPU taken by proposed method are slightly decrease as compare than traditional K-Means that means CPU executing time is minimized.

From the 2^{nd} clusters onwards execution time of CPU for K-means method is changed up to 8^{th} clusters. The iterations is increased for huge dimensional given dataset, execution time is also increased. Another way proposed method hybridize via PCA, it is then minimized the dimensionality of large dataset are reduced results (lower exception time by CPU). The comparison of execution time (CPU time) of AIEMK proposed method, AEIKM via PCA, and traditional K-Means shown in figure 8.3 (a) for dataset D1 and Dataset D2.

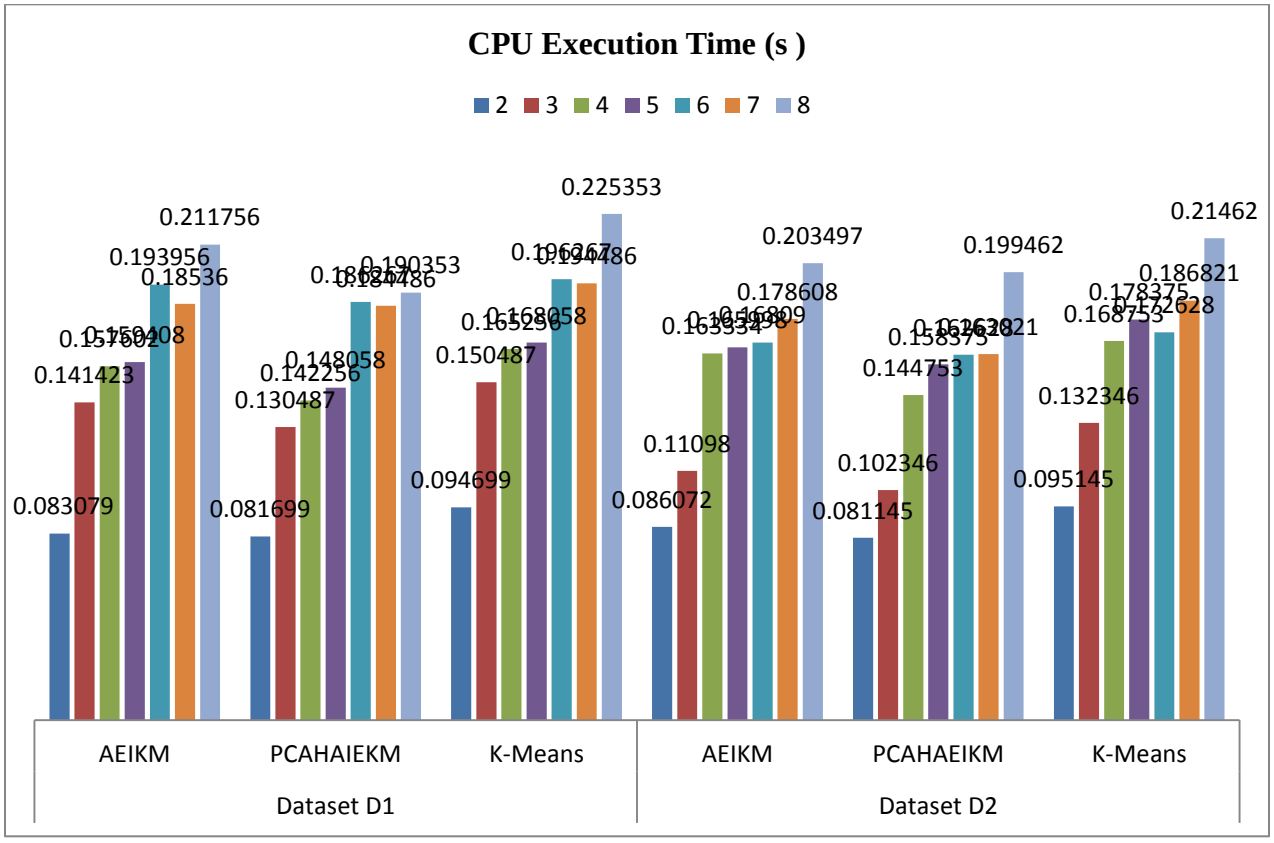


Figure 8.3 (a): Comparative analysis of CPU execution time of dataset D1 and D2

In in figure 8.3 (b) the comparison of executions time (CPU time) of AIEMK proposed method, AEIKM via PCA, and traditional K-Means for dataset D3 and Dataset D4.

Hence, it is observed that the proposed method and it is also hybridize via PCA minimized execution time and minimized the dimensionality of large dataset are reduced

results (lower exception time by CPU) shows in figure-8.3 (a), and figure-8.3 (b) respectively.

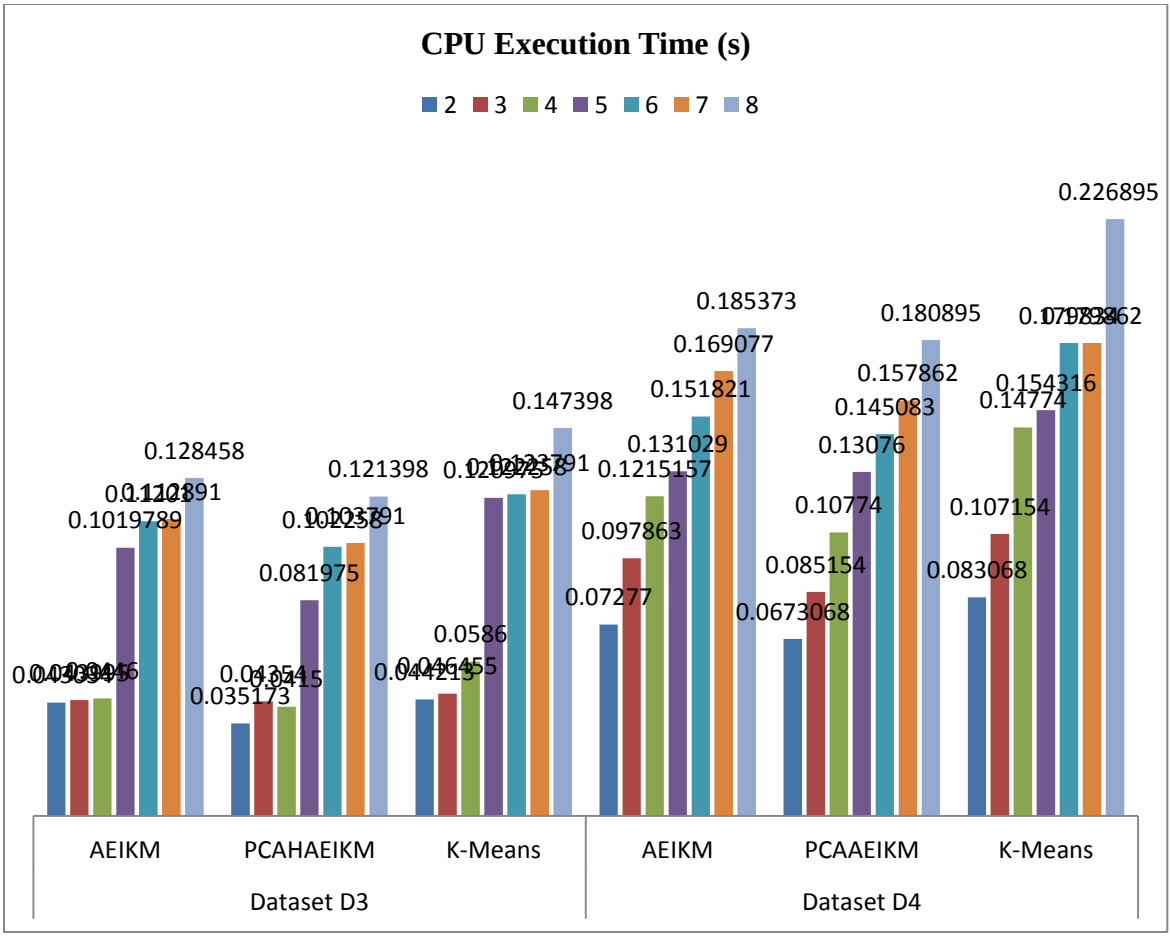


Figure 8.3 (b): Comparative analysis of CPU execution time of dataset D3 and D4

8.1.4 Comparison of External Metrics

In figure-8.4 illustrate the comparison of external metrics between AEIKM proposed method and old K-Means. The AIEKM proposed method has achieved the enhanced metrics over traditional (old) K-Means Method. From cluster number second (02) to onwards, AIEKM proposed method has found accuracy of more than old method and its performance at clusters labels.

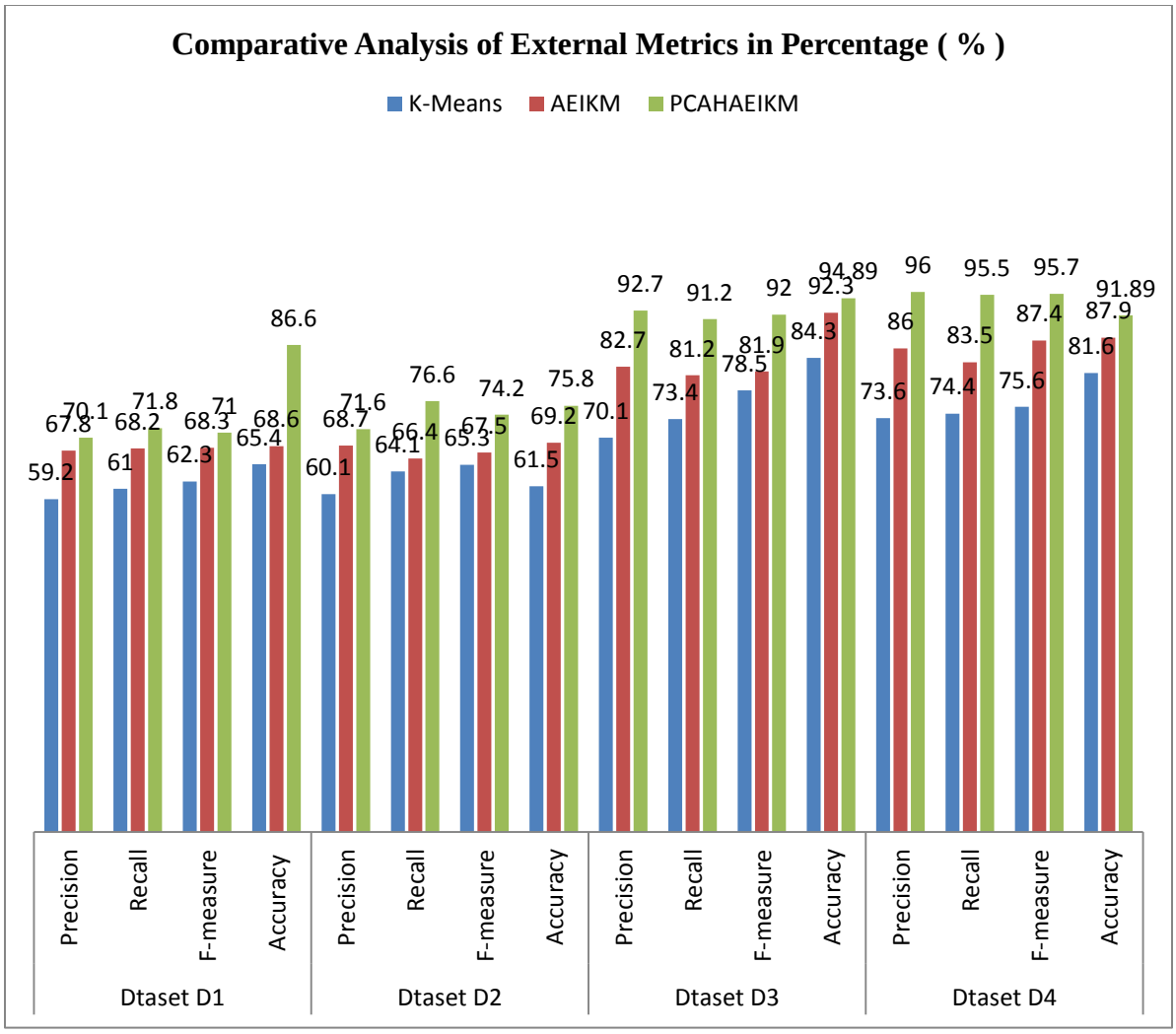


Figure 8.4 Analysis the external metrics by of K-Means, AEIKM, PCAHAEIKM methods

Precision

From Figure 8.4, illustrate the proposed approach offers better value for precision than the current method. Data sets are draw along the X-axis in this graph and precision values are draw along the Y-axis.

Recall

From Figure 8.4, illustrate the proposed approach offers better value for recall than the current method. Data sets are draw along the X-axis in this graph and recall values are draw along the Y-axis.

F-measure

From Figure 8.4, illustrate the proposed approach offers better value for F-measure than the current method. Data sets are draw along the X-axis in this graph and f-measure values are draw along the Y-axis.

Accuracy

From Figure 8.4, illustrate the proposed approach offers better value for accuracy than the current method. Data sets are draw along the X-axis in this graph and accuracy values are draw along the Y-axis.

From the figure 8.4 it is shows that the proposed method has improved accuracy namely as:

Heart disease dataset (D1) AIEKM proposed method is 4.8% more than old k-means, and AIEKM proposed method hybridize via PCA is 32.4% more than old k-means respectively.

User knowledge modeling dataset (D2) AIEKM proposed method is 12.52% more than old k-means, and AIEKM proposed method hybridize via PCA is 23.25% more than old k-means respectively.

Iris dataset (D3) AIEKM proposed method is 9.48% more than old k-means, and AIEKM proposed method hybridize via PCA is12.5% more than old k-means respectively.

Wine dataset (D4) AIEKM proposed method is 7.72% more than old k-means, and AIEKM proposed method hybridize via PCA is 12.61% more than old k-means respectively.

So that, it is observed that the AIEMK, and hybridized concepts has enhanced the performance than traditional method.

Hence, the proposed methodology is better than the existing methodology.

8.1.5 Comparative Analysis of Fitness

In figure-8.5 show the comparison of fitness between K-Means and AEIKM proposed method. And also K-Means compare with AEIKM hybridize via PCA and AEIKM hybridize via PSO. Therefore, mentioned the performance being used the four dataset D1, dataset D2, dataset D3, and dataset D4 is like that in order K-Means < AEIKM<

PCAHAEIKM< PSOHAEIKM. The proposed method hybridize via PSO is better than rest methods.

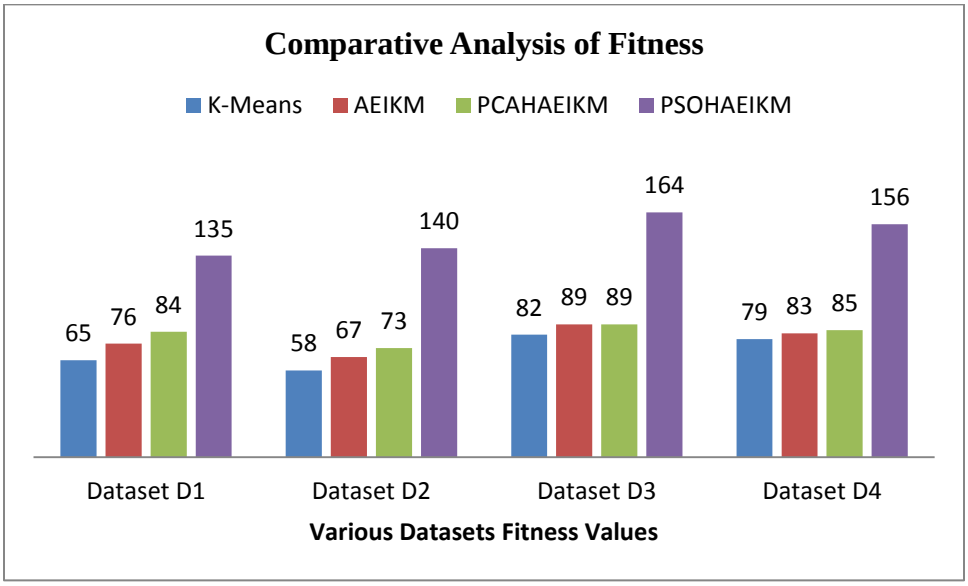


Figure 8.5 Analysis of fitness for various methods

Hence, the proposed methodology is better than the existing methodology.

8.2 Statistical Analysis

The statistical analysis of datasets is analyzed by SPSS and NCSS 2020 tools.

8.2.1 Analysis of Component reductions

In Table- 5.1, show the variance of Heart Disease Dataset (D1) and extractions of five components of same dataset because the eigenvalues of five components is greater than one, and eigenvalues are 3.097, 1.578, 1.261, 1.108, and 1.005 respectively. In Table-5.2, show the variance of User Knowledge Modeling (D2) extractions of two components of same dataset which eigenvalues is 1.382 and 1.173. In Table-5.3, show the variance of iris (D3) and extractions of one components of same dataset eigenvalue is 2.911. In Table-5.4, show the variance of iris (D3) and extractions of three components of same dataset which eigenvalues is 4.706, 2.497, and 1.446 respectively.

8.2.2 Comparative Analysis of Centroids

Measure the min distance between initial centroids of four datasets heart disease dataset (D1), user knowledge modeling (D2) iris Dataset (D3), and wine Dataset (D4). The k-means and this method is hybridized via PCA, they are used the validation of clusters at k. The comparative analysis of the min distance between initial centroid of cluster is by tools at K levels of cluster.

From Figure-8.6 (a) show the min distance between initial centroids at cluster level of user knowledge modeling (D2) at K=4 cluster level proposed method is reduced 24.3% as compare than existing method, and iris (D3) datasets at K=3 cluster level proposed method is reduced 34% than existing method.

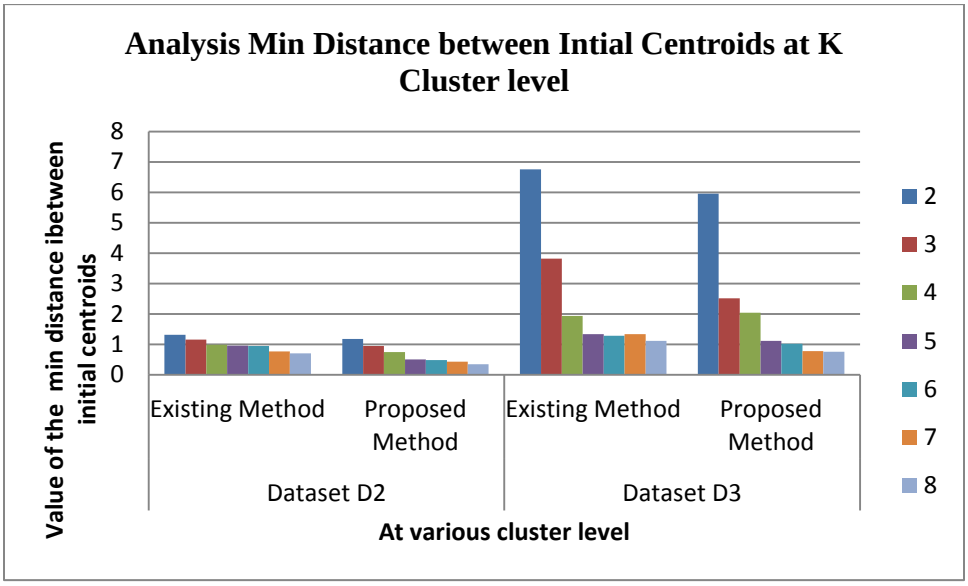


Figure 8.6 (a) Analysis of cluster centroids by SPSS

From Figure-8.6 (a) show the min distance between initial centroids at cluster level of heart disease (D1), and wine (D4) datasets at K=4 cluster level proposed method is reduced 59%, and 0.8% as compare than existing method.

Hence proposed hybridized concept is better than existing for minimized the distance between initial centroids because hybridized value is smaller than existing concept.

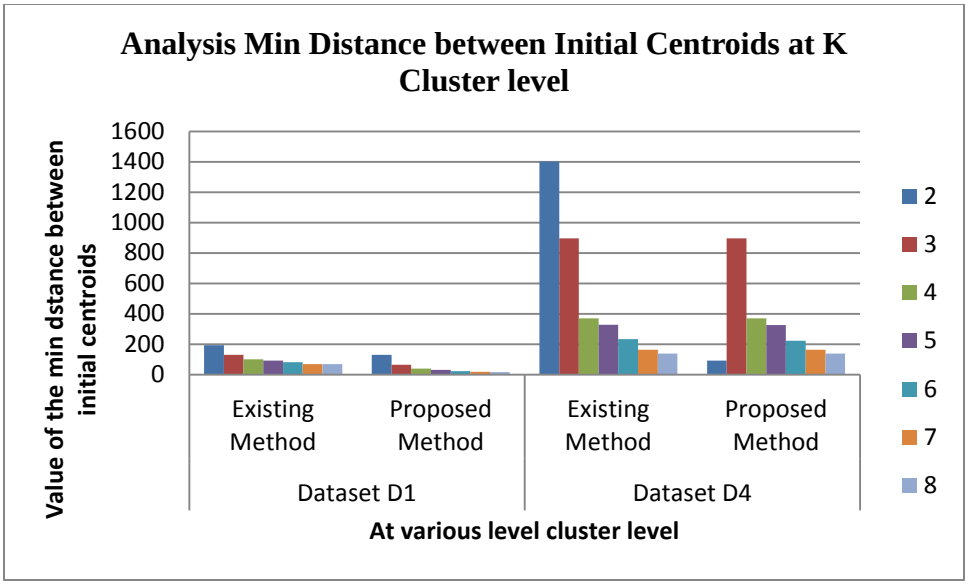


Figure 8.6 (b) Analysis of cluster centroids by SPSS

8.2.3 Analysis of F-ratio

In chapter-5, illustrated in Table 5.5 to Table-5.10 F-Ratio (or F-Score) at K=4 is applied k-means and also hybridized via PCA. They determined the outcome of the F-score is non-negative the created clusters are well separated to the others.

From the same Table analysis of F-Ratio by k-means at k=4, k-means using hybridized PCA for datasets D1, D2, D3, and D4 respectively. Illustrated that the measures value of F-Ratio which is greater than tabled value of 2.0838 at 5% with d. f. used df1= 3, df2= INF. Hence, estimated values of F-Ratio are $>$ F-Rato of tabled value (2.0838). It is not support to the Null-Hypothesis has differences of data means. We, conclude that statistical significant of no. of attributes for support to clusters creations.

8.2.4 Comparative Analysis of Silhouette

The silhouette value close to high then object is matched to itself cluster and very poorly matched to nearby clusters. Further, given maximum object have closed to high, then formation of cluster is suitable otherwise have few or more clusters.

In figue-8.7 illustrate silhouette values is evaluated by fuzzy approaches and also find this values via PCA the proposed methodology is shown the created cluster is significant, and silhouette value is reduced from cluster k= 2 to k=5 in both methods.

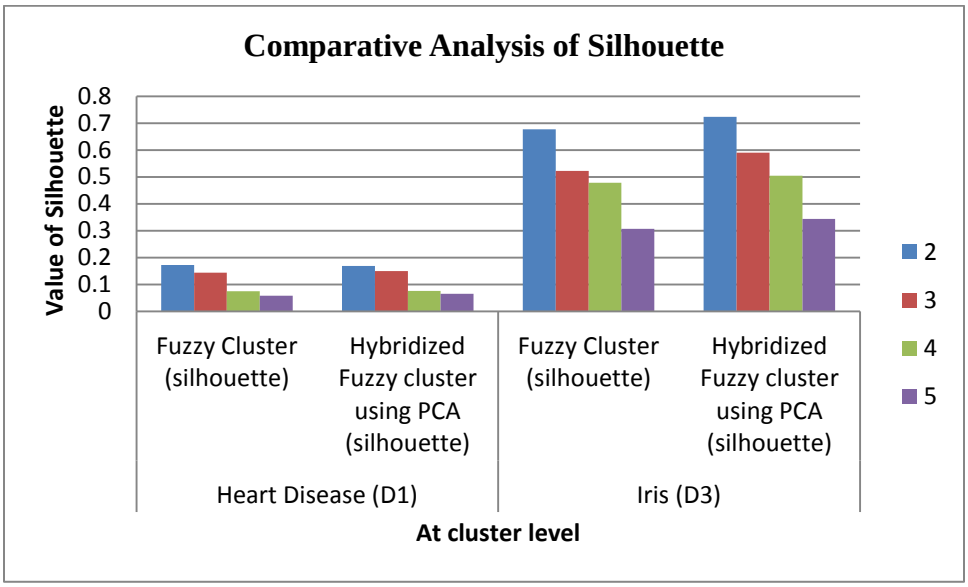


Figure 8.7 Analysis of Silhouette

8.3 Comparative Analysis the Fitness Function

Genetic Algorithm applies for simulation results of the kernel fisher's discriminant analysis. The kernel FDA is superior and more significant as compared to other methods. This performance is more favorable in the arrangement of datasets. it is mentioned that the fitness value of an objective function in terms of best fit and means, stopping criteria, and average distance between individual of the simulation process.

The relative examination criteria of objective function "KFDA" is lesser than an objective function of K-Means. The exit criteria are the selection when the number of generation produced touches the maximum (of population) value.

8.4 SSE by Machine Learning

In Table 8.1, show the comparative analysis of the SSE at k= 3 Iris, and k=4 rest three datasets. From the table show the SSE of K-means method by RBFNN. In figure 8.9, illustrate the analysis between various algorithms and proposed algorithm is better than rest.

Table 8.1 Comparative analysis of SSE with testing and trained by RBFNN

Dataset	Methods			
	k-means	proposed	RBFNN	
			Training	Testing
Dataset D1 K=4	17.05	15.75	86.870	47.226
Dataset D2, K=4	16.67	15.37	88.696	14.464
Dataset D3,K=3	12.28	10.78	25.857	10.548
Dataset D4,K=4	18.34	15.02	43.828	2.337

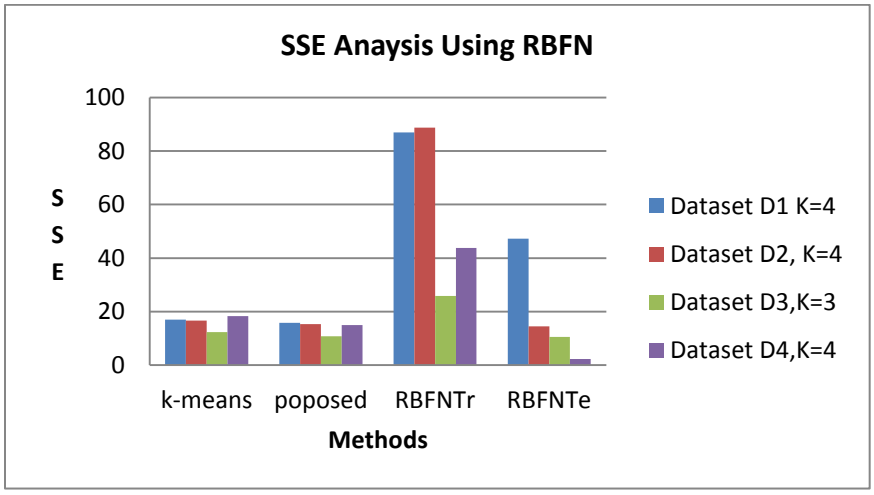


Figure 8.9 Analysis between various algorithms

8.5 Summary

From obtain the results; it's observed the AIEMK proposed method is better performance well for clustering for Dataset(D1), Dataset (D2), Dataset (D3), and Dataset (D4) in the terms of intra cluster distance, SSE, CPU time external metrics up to creates

more clusters. It is observed that the AIEMK proposed method found the better results. We have analyzed by SPSS tool and NCSS tool the min distance between initial centroids of proposed concept is good than existing method using the reduction component, F-RATIO of the original k-means and it is hybridized via PCA is more significant. Compactness of the created clusters all results reveal the significant and good quality clusters by silhouette index.

It is compare the fitness value of an objective function of KFDA is more significant than fitness of objective function of k-means like in terms of best fit and means by genetic algorithm. In addition, the evaluated values of SSE are more significant by RBFN and compare the proposed AEIKM method.

CHAPTER- 9

SUMMARY AND CONCLUSIONS

The outcome of this research effort is presented in this chapter, along with recommendations that can be implemented in the future.

9.1 Conclusions

The efficient data clustering is improved step by step analysis. A feature of data selection is an enhancement, improvement the original K-means method of clustering. In this thesis we deigned an efficient framework of clustering approach for consists four key components follows as mentioned second , third, fourth, and fifth objective respectively. In first component (i.e. second objective) develop the concept to enhance quality of clustering in terms of metrics. In this objective we measure the metrics of clusters quality and enhanced the features by suggested techniques. The new results are derived for various terms for using four datasets. These results proved that the proposed AEIKM method produces high-quality clusters created in lesser time, reduced intra-cluster distance, and minimize the sum of squared errors, but also its method hybridized via PCA has more advanced. The clustering accuracy of this technique is significantly high while compared to the existing techniques. The main objectives namely the determination of enhanced the internal and external metrics in the data clustering process are accomplished by implementing the proposed algorithm. The advantages of the algorithm are, it can handle better data objects of datasets and it can deal with numerous types of real-world datasets capably with the smallest processing time. Measure the fitness of fitness function $F_i^K(X)$by applied PSO concept. The comparative analysis of AEIKM method is between PCAHAEIKM method, and PSOHAEIKM method respectively. And also finally, found the outcome is proposed method hybridized via PSO is better than the rest methods.

The dataset is more diverse, larger, and has high speed in the progression of new technology. Its data is gathered from the different resources, but its dimensionalities and sizes have been more explored in technology development. We have proposed objectives are discussed in chapter-1and mentioned my objective is to handle and find better results as related to the traditional method.

In second component (i.e. third objective) we analyze traditional k-means with various approaches of clustering method by SPSS 1.7, and NCSS 2020 Software tools. In statistical Analysis: The statistical analysis lessens the n^{th}dimensional dataset to applying the PCA ideas on the considered dataset. In the chapter -5, shown are the results with reduced no. of components, and also illustrated the comparative analysis of initial centroids value at k cluster is more significance of the proposed methodology. Further, we have analyzed the F-RATIO or F-SCORE of the original k-means, and it is hybridized via PCA. In the fuzzy approaches to applied find the silhouette the compactness of the created clusters all results reveal the significant and good quality clusters.

In third component (i.e. fourth objective) analyze fitness of objective function by Genetics Algorithm. We have work done used Genetic Algorithm (GA) for simulation results of fitness functions of KFDA and k-means. The fitness function of KFDA is better and more significant as related to function of k-means procedures. In chapter-6, it is mentioned that fitness value of an objective function like in terms of best fit and means. And also the exit criteria are the selection when the no. of generation produced touches the max value of population.

In fourth component (i.e. fifth objective) we, analyze the sum of squared error metrics of clustering metrics by RBNN approach of ANN. We worked on new concepts following the proposed framework and employed the RBFNN approach of ANN in the thesis. RBFNN is touching the appropriate sum of squared error outcomes for extra assistance with the good quality of generated clusters. PCA, PSO, and fuzzy approaches are all concern of the hybridized notion. To improve the cluster quality performance by recognizing and selecting attributes.

9.2 Observation and Limitation

In clustering, here are more arises the difficulty for n-dimensional and huge capacity of data. As an outcome, it is required to critical investigate the data parallelism with dispersed across numerous processors, and along with the usage of memory (secondary storage) to handle the massive amount of public data.

It consists of various types for data like structured data, “unstructured data”, and semi-structured data. Therefore, still, some difficulties of clustering have nearby reduced the dimensionality, but a huge capacity of data has become more complex for its analysis.

9.3 Future Work and Future Scope

Nowadays the evolution of research works in area of mining for huge genuine dataset to explore metric of clustering. The PCA and PSO algorithm are influenced on the clustering efficiency. In future research work by the PSO algorithm will be achieved best fitness value of implementing on a choice of fitness functions. In advance, PCA is implemented on a huge dataset to the reduction of dimension. PSO idea is useful to the general problem with calculating the fitness in the future research work, but the limitation of fast convergence.

This work has been expanded to include pattern matching for a diversity of languages, numerals, and contours, as well as a comparison of several handwritten letters. In adding, the GA concept is applied to public datasets to construct clusters related to defined objective functions, and live datasets.

REFERENCES

[1] Haq, E. U., Huarong, X., & Khattak, M. I. (2017). A Review of Various Clustering Techniques. Empirical Research Press Ltd.

[2] Patel, B., & Gondaliya, C. (2017). Student Performance Analysis Using Data Mining Technique, International Journal of Computer Science and Mobile Computing, 6(5), 64-71

[3] Bae, Y., Kim, Y. S., Rhee, F. C. H., Kim, Y. T., & Tao, C. W. (2017). Editorial Message: Special Issue on Fuzzy System in Data Mining and Knowledge Discovery: Modeling and Application. International Journal of Fuzzy Systems, 19(4), 1157-1157.

[4] Wang, X., & Bai, Y. (2016). A modified Min Max-means algorithm based on PSO. Computational intelligence and neuroscience, 2016.

[5] Jalil, A. M., Hafidi, I., Alami, L., & Khouribga, E. N. S. A. (2016). Comparative study of clustering algorithms in text mining context. IJIMAI, 3(7), 42-45.

[6] Alsayat, A., & El-Sayed, H. (2016). Social media analysis using optimized K-Means clustering. In 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 61-66). IEEE.

[7] Rathore, P., & Shukla, D. (2015). Analysis and performance improvement of K-means clustering in big data environment. In 2015 International Conference on Communication Networks (ICCN) (pp. 43-46). IEEE.

[8] Haraty, R. A., Dimishkieh, M., & Masud, M. (2015). An enhanced k-means clustering algorithm for pattern discovery in healthcare data. International Journal of distributed sensor networks, *11*(6), 615740.

[9] Boobord, F., Othman, Z., & Abubakar, A. (2015). PCAWK: A Hybridized Clustering Algorithm Based on PCA and WK-means for Large Size of Dataset. Int. J. Advance Soft Compu. Appl, *7*(3).

[10] Kamel, N., Ouchen, I., & Baali, K. (2014). A sampling-pso-k-means algorithm for document clustering. In Genetic and evolutionary computing (pp. 45-54). Springer.

[11] Eslamnezhad, M., & Varjani, A. Y. (2014). Intrusion detection based on Min Max K-means clustering. In7'th International Symposium on Telecommunications (IST'2014) (pp. 804-808). IEEE.

[12] Ganganath, N., Cheng, C. T., & Chi, K. T. (2014). Data clustering with cluster size constraints using a modified k-means algorithm. In 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (pp. 158-161). IEEE.

[13] Anusha, M., & Sathiaseelan, J. G. R. (2014). An enhanced K-means genetic algorithm for optimal clustering. In 2014 IEEE International Conference on Computational Intelligence and Computing Research (pp. 1-5). IEEE.

[14] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[15] Suganya, R., & Shanthi, R. (2012). Fuzzy c-means algorithm-a review. International Journal of Scientific and Research Publications, *2*(11),

[16] Xu, R., & Wunsch, D. C. (2005). Survey of clustering algorithms. In 2005 Transaction on Neural Networks. IEEE.

[17] Hongxia, P., Xiuye, W., & Jinying, H. (2010). Fault feature extraction based on KPCA optimized by PSO algorithm. In 2010 8th IEEE International Conference on Industrial Informatics (pp. 102-107). IEEE.

[18] Xu, X., Liu, H., Li, L. and Yao, M. (2018). A comparison of outlier detection techniques for high dimensional data. International Journal of Computational Intelligence Systems, 11(1), 652-662.

[19] Anusuya V. and Lattha P. (2011). Clustering of datasets using PSO-K-Means and PCA-K-Means. International Journal of computational Intelligence and Informatics, 1(3), (2011), 180-184.

[20] Alashwal H., et al.(2019). The application of unsupervised clustering methods to Alzheimer's disease: Frontiers in Computational Neuroscience, 13, 31. https://doi.org/10.3389/fncom.2019.00031

[21] Aryuni M., Madyatmadaja E. D., and Miranda E.(2018). Customer segmentation in XYZ bank using k-means and k-medoids clustering. In 2018 International

Conference on Information Management and Technology (ICIMTech),(pp. 412-416).IEEE.

[22] Yelda M., Pathakota S. R. and Srinivasa T. M. (2010). Enhancing K-Means clustering algorithm with improved initial center. International Journal of Computer Science and Information Technologies, 1(2), 121-125.

[23] Tzortzi G. and Likas A.(2014). The min-max k-means clustering algorithm. Pattern Recognition, 47(7), 2505-2516.

[24] Ali H. H. and Kadhum L. E. (2017).K-Means clustering algorithm applications in data mining and pattern recognition. International Journal of Science and Research, 6(8), (2017), 1577-1584.

[25] Kumar S. and Kaur S.(2017). Modified k-means clustering algorithm for disease prediction. International Journal of Engineering and Techniques, 3(3), 195-201.

[26] Hossain M. Z., Akhtar M. N., Ahmad R. B. and Rahman M. (2019.) A dynamic k-means clustering for data mining. Indonesian Journal of Electrical Engineering and Computer Science, 13(2), (2019), 521-526.

[27] Ma L., Gu L., Li, B. Ma, Y. and Wang J. (2015). An improved K-Means algorithm based on map reduced and grid.. International Journal of Grid and Distributed Computing, 8(1).
DOI:10.14257/IJGDC.2015.8.18

[28] Allam M. N..(2016). Particle swarm optimization: algorithm and its code in MATLAB. Research Gate, 8,), 1-10.

[29] Pena, J.M., Lozano, J. A., & Larranaga, P. (1999).An empirical comparison of four initialization methods for the K-Means algorithm. Pattern recognition letters, 20(10), 1027-1040.

[30] Celebi, M. E., Kingravi, H. A. , & Vela, P. A. (2013). A comparative study of efficient initialization methods for the K-Means clustering algorithm. Expert System with Applications, 40(1), 200-210.

[31] Han, X., Quan, L., Xiong, X., Almeter, M., Xiang, J., & Lan, Y. (2013).A novel data clustering algorithm based on modified gravitational search algorithm. Engineering Applications of Artificial Intelligence, 61, 1-7.

[32] Scholkopf, B., Smola, A., & Muller, R. (1997). Kernel principal component analysis. In International Conference on Artificial Neural Networks, (pp. 583-588). Springer, Berlin, Heidelberg.

[33] Mostafa, A. et.al. (2015).CT liver segmentation using artificial bee colony optimization. Procedia Computer Science, 60, 1622-1630.

[34] Xinfeng, W., Jing, Q., &, Guanjun, L. (2007). Kernel function optimization in kernel principle component analysis and its application to feature extraction of Gear faults. Journal of vibration, Management & diagnosis, 27(1), 62-64.

[35] Mirkin, B.(2012).Clustering: a data recovery approach. CRC Press.

[36] Vesterstrom, J., & Thomsen, R. (2004).A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithm on numerical Benchmark problems. In Evolutionary Computation *CEC2004* Congress on evolutionary computation (vol.2, pp. 1980- 1987). IEEE.

[37] Ratnaweera, A., Halgamuge, S. K., & Watson, H. C. (2004).Self organizing hierarchical particle swarm optimizer with time varying acceleration coefficients. IEEE Transaction evolutionary computation, 8(3), 240-255.

[38] Liu, X., & Fu, H.(2010). An efficient clustering algorithm with ant colony. JPC, 5(4), 598-605.

[39] Poli, R., Kennedy, J., & Blackwell, T. (2007).Particle swarm optimization. Swarm Intelligence, 1(1), 33-57.

[40] Majhi, S. K., & Biswal, S. (2018).Optimal cluster analysis using hybrid K-Means and ant lion optimizer. Kabala International Journal of modern science, 4(4), 347- 360.

[41] Mary, C., & Raja, S. K. (2009).Refinement of clusters from K-Means with ant colony optimization. Journal of theoretical & applied information technology, 6.

[42] Ren, Q., & Zhuo, X.(2011).Application of an improved K-Means algorithm in gene expression data analysis. In 2011 IEEE International Conference on System Biology (ISB), (PP.87-91).IEEE.

[43] Ma, L., Gu, L. , Li, B., Ma, Y., & Wang, J.(2015). An improved K-Means algorithm based on map reduced and grid," International Journal of Grid & Distributed computing, 8(1), 2015.

[44] Shi-Wei, L., & Xiao-Dong, Q.(2010).Date clustering using principal component analysis and particle swarm optimization. In 2010 5th International conference on computer science & education, (pp.493-497). IEEE.

[45] Sethi, C., & Mishra, G. (2013).A linear PCA based hybrid K-Means PSO algorithm for clustering large dataset. International Journal of scientific & Engineering Research, 4(6), 1559-1566.

[46] Komaraswamy, G., & Wahi, A. (2011).Improving the cluster performance by combining PSO and K-Means algorithm," ICTACT Journal on soft computing, 1(4).

[47] Nazeer, K. A., & Sebastian, M. P. (2009).Improving the accuracy and efficiency of K-Means clustering algorithms. Proceeding of the world congress on engineering, (vol.1, and pp.1-3) London: Association of Engineers.

[48] Olson, D.L.(2007). Data Mining in Business Services. Service Business, 1(3), 181-193.

[49]. Kudyba, S., & Hoptroff, R.(2001). Data Mining and Business Intelligence: A Guided to Productivity.IGI Global.

[50] Tsai, C. F., Tsai, C. W., Wu, H.C., & Yang, T.(2004). ACODF: A Novel Data Custer Approach for Data Mining in Large Databases. Journal of Systems and Software, 73(1), 133- 145.

[51] Zhang, N., Leatham, K., Xiong, J., & Zhong, J., (2018). PCA K-Means Based Clustering Algorithm for High Dimensional and Overlapping Spectra Signals. In 2018 Ninth International Conference on Intelligent Control and Information Processing(ICICIP) (pp.349-345). IEEE.

[52] Jamal, A., Handayani, A.,Septiandri, A. A.,Ripmiatin, E., &Effendi, Y.(2018). Dimensionality Reduction using PCA and K-Means Clustering for the Breast Cancer Prediction. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 192-201

[Online]: DOI: https://doi.org/10.24843/LKJITI.2018.v09.i03.p08

[53] Alkhavrat, M., Aljnidi , M., & Aljoumaa, K. (2020). A Comparative Dimensionality Reduction Study in Telecom Customer Segmentation using Deep Learning and PCA. Journal of Big Data, 7(1), 9.
[Online]: DOI https://doi.org/10.1186/s40537-020-0286-0

[54] Al-Zubai, I. M., Jafa, A., & Aljoumaa, K.(2019). Predicting Customer's Gender and Age Depending on Mobile Phone Data. Journal of Big Data, 6(1),18.
[Online]: https://doi.org/10.1186/s40537-019-0180-9

[55] Peres-Neto, P. R., Jackson, D., &Somers, K. M.(2005). How Many Principal Components? Stopping Rules for Determining the Number for Non -trivial Axes Revisited. Computational Statistics & Data Analysis, 49(4), 974-997.

[56] Wang, D., Cui, P., & Zhu, W. (2016). Structural Deep Network Embedding. In Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1225-1234).

[57] Ding, C., & He, X.(2004). K-Means Clustering via Principal Component Analysis. In Proceedings of the 21st International Conference on Machine Learning (pp.29).
[Online]: https://doi.org/10.1145/1015330.1015408

[58] Hermawanto, D.(2013). Genetic Algorithm for Solving Simple Mathematical Equality Problem. *arXiv preprint arXiv*. 2013; 1308.4675.
https://arxiv.org/ftp/arxiv/papers/1308/1308.4675.pdf.

[59] Al Malki A., Rizk M. M., El-Shorbagy M. A., Mousa A. (2016). A.Hybrid genetic algorithm with k-means for clustering problems. Open Journal of Optimization. 5(02), 71.

[60] Oujezsky V., Horvath. (2018). Traffic similarity observation using a genetic algorithm and clustering. Technologies, 6(4), 103.

[61] Kemsley E. K. (1966). Discriminant Analysis of High Dimensional Data: A Compression of Principal Component Analysis and Partial Least Square Data Reduction Methods. Chemo Metrics and Intelligent Laboratory Systems, 33, 47-61.

[62] Sayed E. H., Gabbar H. A., Miyazaki S. (2009).Improved Evolving Kernel of Fisher's Discriminant Analysis for Classification Problem. Journal of Applied Sciences, 9(12), 2313-23.

[63] Min W., Siqing Y.(2010).Improved K-Means clustering based on Genetic Algorithm. In 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). IEEE. 6, V6-636. doi:10.1109/ICCASM.2010.5620383.

[64] Yong Y., Xincheng G.(2012).A New Minority kind of Sample Sampling Method Based on Genetic Algorithm and K-Means Cluster. In 2012 7th International Conference on Computer Science & Education (ICCSE).IEEE.126-129. doi:10.1109/ICCSE.2012.6295041.

[65] Kachitvichyanukul V. (2012).Comparison of Three Evolutionary Algorithms: GA, PSO and DE. Industrial Engineering & Management Systems, 11(3), 215-223.

[66] Babaie S. S., Mahdi E. E. O., Firoozan T. (2015).A Novel Combined Approach of K-Means and Genetic Algorithm to Cluster Cultural Goods in Household Budget. In Proceeding of 4th International Conference on Frontiers in Intelligence Computing: Theory and Applications (FICTA). Springer, New Delhi. 273-283. https://doi.org/10.1007/978-81-322-2695-6_24.

[67] Bhatia S. (2014).New Improved Technique for Initial Cluster Centers of K-Means Clustering using Genetic Algorithm. In International Conference for Convergence for Technology (ICCT 2014), IEEE, 1-4. doi:10.1109/12CT.2014.7092112.

[68] Rahman M. A., Islam M. Z. (2014).A Hybrid Clustering Technique Combining a Novel Genetic Algorithm with K-Means. Knowledge Based Systems, 71, 345-365.

[69] Lu Z., Zhang K., He J., Niu Y. (2016).Applying K-Means Clustering and Genetic Algorithm for Solving MTSP. In International Conference on Bio-Inspired Computing: Theories and Applications, Springer, Singapore, 278-284. https://doi.org/10.1007/978-981-10-3614-9_34.

[70] Aibinu A. M., Salau H. B., Rahman N. A., Ackachukwu C. M. (2016).A Novel Clustering Based Genetic Algorithm for Route Optimization. Engineering Science and Technology an International Journal, 19(4), 2022-2034.

[71] Zeebaree D. Q., Haron H., Abdulazeez A. M., Zeebaree S. R. (2017).Combination of K-Means Clustering with Genetic Algorithm: A Review. International Journal of Applied Engineering Research, 12(24), 14238-14245.

[72] Krishnasamy, G., Kulkarni, A. J., Paramesran, R. (2014).A hybrid approach for data clustering based on modified cohort intelligence and k-means. Expert Systems with Applications, 41(13), 6009-6016.

[73] Krishna K., Murty M. N. (1999).Genetic k-means algorithm. IEEE Transactions on Systems, Man, and (Cybernetics), 29(3), 433-349.

[74] Prashanth N. A., Sujatha P.(2018).Comparision Between PSO and Genetic Algorithms and for Optimizing of Permanent Magnet Synchronous Generator(PMSG) Machine Design. International Journal of Engineering & Technology, 7(3.3), 77-81.

[75] Kennedy J., Eberhart R.(1995).Particle Swarm Optimization. In Proceeding of ICNN95 International Conference on Neural Networks. IEEE, 4, 1942-1948. doi:10.1109/ICNN.1995.488968.

[76] Tharwat A., Gaber T., Hassanien A. E., Elnaghi B. E.(2017).Particle Swarm Optimization: A Tutorial. In Hand book of Research on Machine Learning Innovations and Trends, IGI Global, 614-635. doi:10.4018/978-1-5225-2229-4.ch026.

[77] Van der Merwe D., Engelbrecht A. P. (2003).Data clustering using Particle Swarm Optimization. The 2003 congress on Evolution Computation CEC'03, IEEE. 1, 215-220. doi:10.1109/CEC.2003.1299577.

[78] Shi Y., Eberhart R. (1998).A Modified Particle Swarm Optimizer. Applied Mathematics and Computation, 189(5), 69-73.

[79] Zhao W., Zu W., Zeng H.(2009).A modified Particle Swarm Optimization via Particle visual Modeling Analysis. Computers & Mathematics with Applications, 57(11-12), 2022-2029.

[80] Hongxia P., Xiuye W., Jinying H. (2010).Study of Fault Extraction Based on KPCA Optimized by PSO Algorithm. In International Conference on Fuzzy Systems. IEEE. 1-6.
doi :10/1109/FUZZY.2010.5583947.

[81] Xinchao Z. (2010).A Perturbed Particle Swarm Algorithm for Numerical Optimization. Applied Soft Computing, 10(1), 119-124.

[82] Gen M., Cheng R.(1997).Genetic Algorithm and Engineering Design. John Wiley & Sons, Inc., New York.

[83] Wei X., Pan H., Wang F.(2009).Feature Extraction Based on Kernel Principal Component Analysis Optimized by PSO Algorithm. Journal of Vibration, Measurement and Diagnosis, 29(9), 162-166.

[84] He Y., Wang Z. (2018).Regularized Kernel Function Parameter of KPCA Using WPSO-FDA for Feature Extraction and Fault Recognition of Gearbox. Journal of Vibroengineering, 209(1), 225-239.

[85] Chang D. X., Zhang X. D. Zheng C. W.(2009).A Genetic Algorithm with GENE Rearrangement for K-Means Clustering. Pattern Recognition, 42(7), 1210-1222.

[86] Dabbura I. (2018).K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Towards Data Science, Saatavissa: https://twardsdatascience.com/k-means-clustering-algorithm-evaluation-methods-and-drawbacks-aa03e644b48a

[87] De Wit E., Van Doremalen N., Farzarano D., Munster V. J.(2016).SARS and MERS: recent insights into emerging coronaviruses. Nature Reviews Microbiology, 14(8), 523.
Doi:10.1038/nrmicro.2016.81

[88] Chan J. F. W., Yuan S., Kok K. H., To K. K. W., Chu H., Yang J.,..Yuen K. Y.(2020).A familial cluster of pneumonia associated with the 2019 novel

coronavirus indicating person to person transmission: a study of a family cluster. The Lancet, 395(10223), 514-523.
https://doi.org/10.1016/s0140-6736 (20)30154-9

[89] Otter J. A., Donskey C, Yezli S., Douthwaite S., Goldenberg S., Weber D.(2016).Transmission SARS and MERS coronaviruses and influenza virus in healthcare settings: the possible role of dry surface contamination. Journal of Hospital Infection, 92(3), 235-250.
Doi:10.1016/j.jhin.2015.08.027

[90] Dowell S. F., Simmerman J. M., Erdman D. D., Wu J. S. J., Chaovavanich A., Javadi M., ..and, Ho M.(2004). S.Severe acute respiratory syndrome coronavirus on hospital surface. Clinical Infectious Diseases, 39(5), 652-657.
Doi:10.1086/422652

[91] Kampf G., Todt D., Pfaender S., and Steinmann E. (2020).Persistence of coronaviruses on inanimate surfaces and their inactivation with biocidal agents. Journal of Hospital Infection, 104, 246-251.

[92] Bezdek, J. C., and Pal, N. R. (1995).Cluster validation with generalized Dunn's indices. In proceedings 1995 Second New Zealand International Two Stream Conference on Artificial Neural Networks and Expert Systems(pp. 190-193). IEEE.
Doi:10.1109/ANNES.1995.499469

[93] Ilc, N.(2012). Modified Dunn's clustering validity index based on graph theory. Przeglqd Electrotechniczny,88(2),126-131.
[Online].Available: http://pe.org.pl/articles/2012/2/36.pdf

[94] Bora D. J., and Gupta D. A. (2014). A comparative study between fuzzy clustering algorithm and hard clustering algorithm. *arXiv preprint arXiv*: 1404.6059.

[95] Steinbach, M., Kumar, V., and Tan, P.(2005). Cluster analysis: Basic concepts and algorithms. Introduction to data mining, 1st edition, Pearson Addison Wesley.

[96] Rousseau, P. J.: Silhouettes.(1987). A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.
Doi:10.1016/0377-0427(87)90125-7.

[97] "Silhouette (Clustering)", Dec, 28; 2020. [Online]. Available: https://en.wikipedia.org/wiki/silhouette_(clustering)

[98] Rewashes M. and Ralescu A. (2012). Fuzzy cluster validity with generalized silhouette. In annual Meeting of the North American Fuzzy Foundation Processing Society (NAFIPS), 2012, (pp.16).IEEE.
[Online]. Available: http://ceur-ws.org/vol-841/submission_33.pdf

[99] Dunn J. C.(1973).A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. J. Cyberner 3, 32- 57.

[100] Selim S. Z. and Kamel M. S. (1992). On the mathematical and numerical properties of the fuzzy c-means algorithm. Fuzzy sets and System, 49(2), 181-191.

[101] https://www.cs.bham.ac.uk/~jxb/INC/l14.pdf

[102] Wettschereck, D., and Dietterich, T. (1991). Improving the performance of radial basis function networks by learning center locations. In NIPS (Vol. 4, pp. 1133-1140).

[103] Chun-tao, M., Kun, W., and Li-yong, Z. (2009, March). A new training algorithm for RBF neural network based on PSO and simulation study. In 2009 WRI World Congress on Computer Science and Information Engineering (Vol. 4, pp. 641-645). IEEE.

[104] de Lacerda, E. G., de Carvalho, A. C., and Ludermir, T. B. (2002). Model selection via genetic algorithms for RBF networks. Journal of Intelligent & Fuzzy Systems, 13(2-4), 111-122.

[105] Ziyang, Z., Zhisheng, W., Yong, H., and Mingzhi, G. (2008, July). Learning method of RBF network based on FCM and ACO. In 2008 Chinese Control and Decision Conference (pp. 102-105). IEEE.

[106] Dash, C. S. K., Behera, A. K., Dehuri, S., and Cho, S. B. (2016). Radial basis function neural networks: a topical state-of-the-art survey. Open Computer Science, 1(open-issue).

[107]. Kumar, A., Usha, T. A., & Sivaranjani , C. A.(2016). A comparative study on K-Means and genetic algorithm for data clustering. International Journal of Engineering Research and Development, 12(11), 01-09.

[108] Chaita, J. (2016). Implementing & Improvisation of K-means Clustering Algorithm. International Journal of Computer Science and Mobile Computing, *5*, 191-203.

[109] Capso, M., Perez, A., and Lozano, J. A.(2020). An efficient k-means clustering algorithm for tall dataset. Data Mining and Knowledge Discovery, 34, 776-811. https://doi.org/10.1007/s10618-020-00678-9

[110] Aslam, A., Qamar, U., Kan, R. A., and Saqib, P.(2020). Improving k-means method by finding initial centroid points. In 2020 22nd International Conference on Advanced Communication Technology(ICACT) (pp. 624-627). IEEE.

[111] Aydin, N., & Yurdakul, G.(2020). Assessing countries performances against COVID-19 via WSIDEA and machine learning algorithms. Applied Soft Computing, 97, 106792.

[112] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020).The k-means algorithm. A Comprehensive Survey and Performance Evaluation.Electronics, 9(8), 1295.

[113] Baruri, R., Ghosh, A., Banerjee, R., Mandal, A., and Halder, T.(2019). An empirical evaluation of k-means clustering technique and comparision. In 2019 International Conference on Mchine Learning. Big Data Cloud and Parallel Computing (COMITCon) (pp. 470-475). IEEE

[114] Verma, and A. Zisserman (2008). A statistical approach to material classification using image patches exemplar. IEEE Transaction on pattern analysis and Machine Intelligence, 31(11) , 2031-2047.

[115] Coates, Ng and H. Lee. (2011). An analysis of single layer networks in unsupervised feature learning. In Proceedings of 14 th International Conference

on Artificial Intelligence and Statistics. (pp. 2015-223). JMLR Workshop and Conference Proceeding.

[116] Bishop, C. (1991). Improving the generalization properties of radial basis function neural networks. Neural Computation, 3(4), 579-588.

[117] Park and Sandlberg, I. W. (1993).Approximation and radial basis function networks. Neural Computation, 5(2), 305-316.

LIST OF PUBLICATIONS

1. Muhammad Kalamuddin Ahamad, Ajay Kumar Bharti," An Effective Technique on Clustering in Perspective of Huge Data Set," International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020.

2. Muhammad Kalamuddin Ahamad, Dr. Ajay Kumar Bharti, "Comparative Analysis The Fitness Function of K-Means and Kernel Fisher's Discriminant Analysis (KFDA) With Genetic Algorithm," European Journal of Molecular & Clinical Medicine(EJMCM), ISSN 2515-8260 Volume 7 Issue 11, 2020. **Scopus Index** International Journal.

3. Muhammad Kalamuddin Ahamad, Dr. Ajay Kumar Bharti,"Analysis the Cluster Performance OF Real Dataset Using SPSS TOOL With K-Means Approach VIA PCA," Advances in Mathematics: Scientific Journal (AMSJ),**10** (1), pp. 535–542, 2021. ISSN: 1857-8365 (printed); 1857-8438 (electronic), **Scopus Index** International Journal. https://doi.org/10.37418/amsj.10.1.53 Spec. Iss. on ICIRPS-2020

4. Muhammad Kalamuddin Ahamad, Dr. Ajay Kumar Bharti," Fitness Analysis of Cluster for Real World COVID-19 Dataset Using Fuzzy Clustering,", Journal of Huazhong University of Science and Technology, 50(02), Feb 2021. ISSN: 1671-4512 **Scopus Index** International Journals.

5. Muhammad Kalamuddin Ahamad, Dr. Ajay Kumar Bharti," Prevention from the COVID-19 in India: Fuzzy Logic Approach," In International Conference on Advanced Computing and Innovative Technologies in Engineering (ICACITE) 2021, Added in 20 April IEEE Explorer, Scopus Database, doi:10.1109/ICACITE51222.2021.9404575

6. Muhammad Kalamuddin Ahamad, Dr. Ajay Kumar Bharti,"Safety Assessment of health care Covid-19 in India: Assisted the Fuzzy logic Approach (FIS)"In "Recent Advancement in Information and Communication Technology (RAICT-2020)," organized by Integral University, Lucknow 05 - 07 November, 2020 , Attended

- **COMMUNICATED/ACCEPTED:**
    1. Validation of clustering based framework using unsupervised Machine learning, SAMS 2021, **ACCEPTED**
    2. An Efficient Improved K-Means to Measure the Cluster Performance Using Hybridized Technique

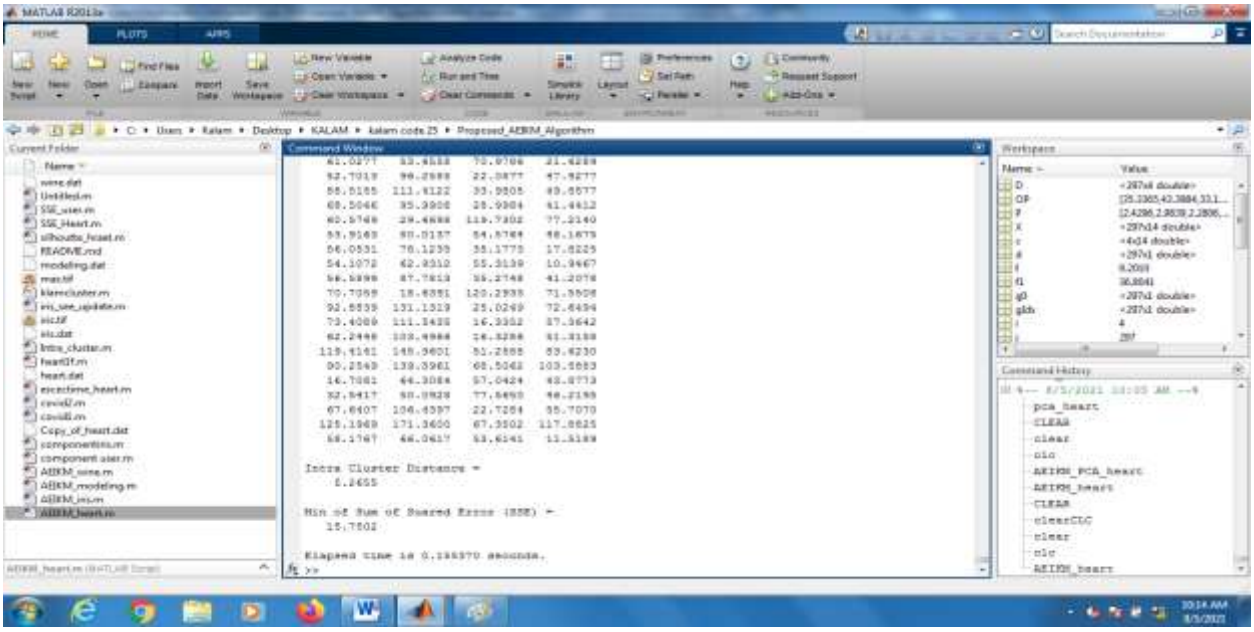
APPENDIX-I

Result Snapshot (D1), Tables

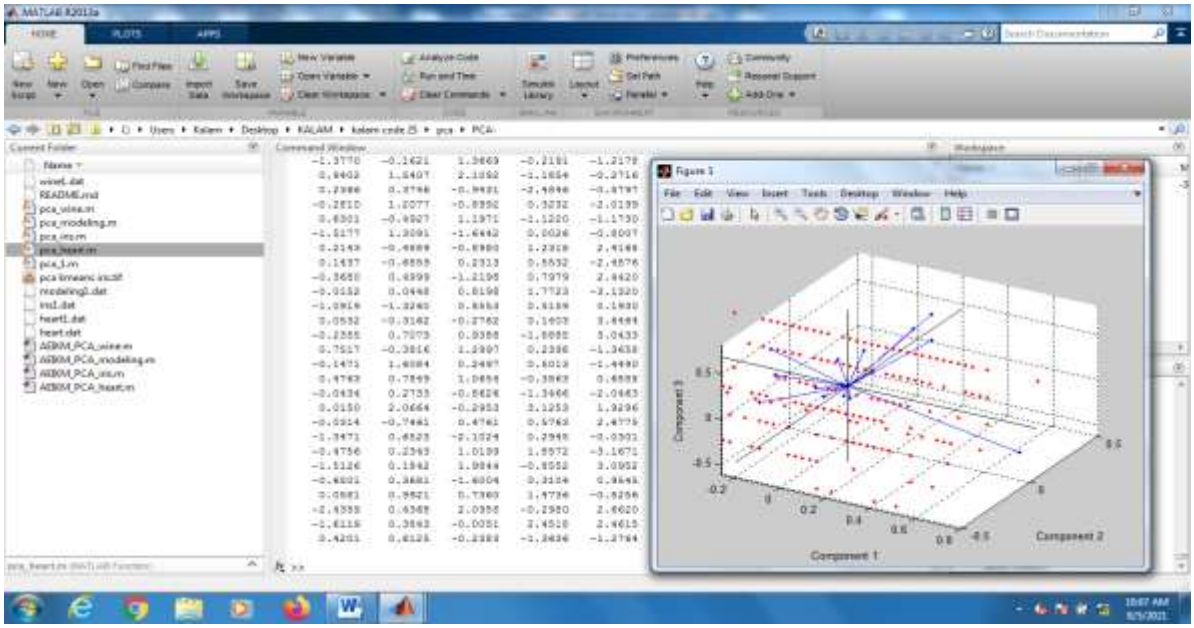
Result Snapshot (D1)

For heart disease dataset (D1)

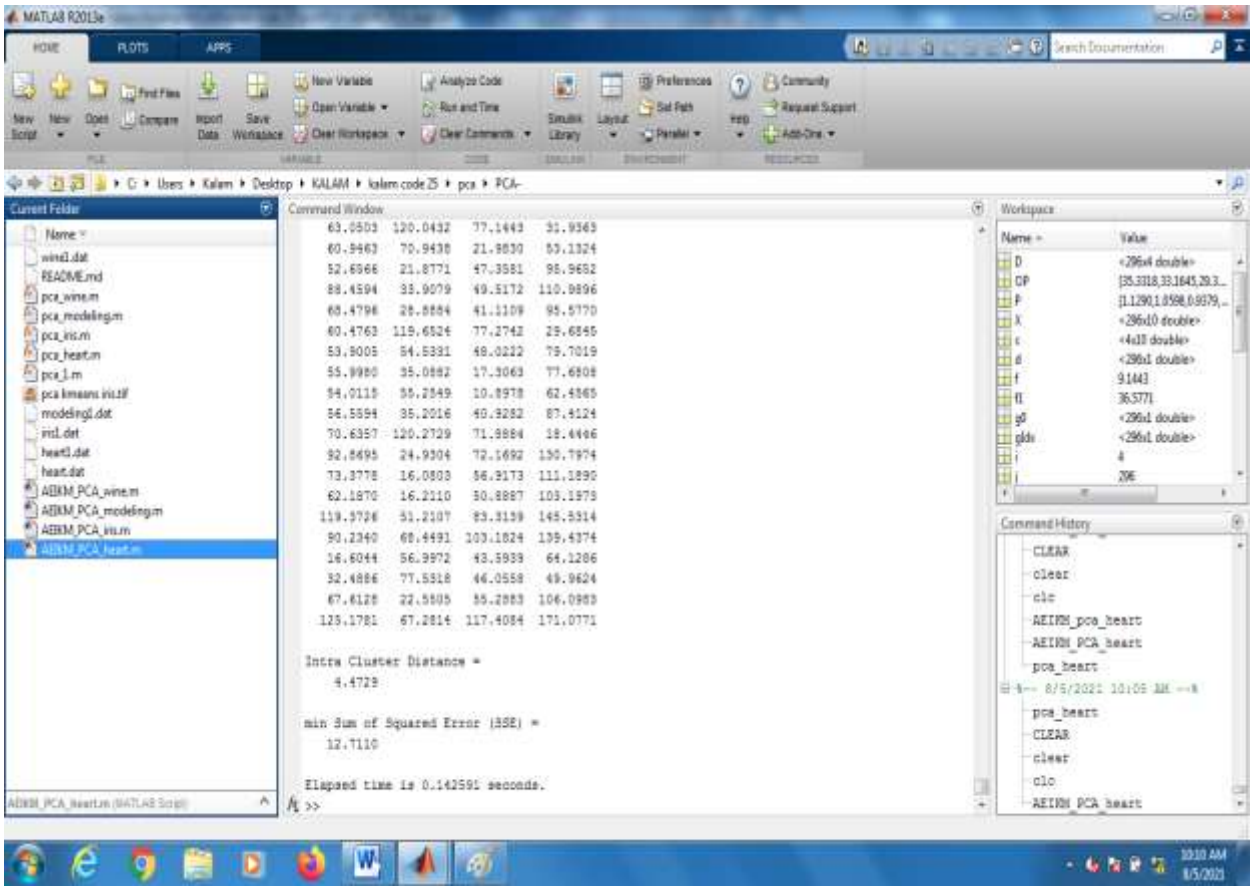
- **Proposed AEIKM Algorithm: Heart Disease Dataset D1 at K=4**



- **Reduction Components by PCA: Heart Disease Dataset D1**



- **AEIKM Hybridized via PCA: Heart Disease Dataset D1 at K=4**



TABLES

Table 1 Comparative Analysis of the Intra Cluster Distance D2 and D3

Datasets	Methods	No. of Clusters						
		2	3	4	5	6	7	8
User Knowledge Modeling (D2)	K-Means	0.1356	0.1045	0.132	0.1123	0.1102	0.1022	0.113
	AEIKM	0.1262	0.0984	0.124	0.1071	0.1035	0.0906	0.101
	PCAHAEIKM	0.0789	0.0686	0.0455	0.036	0.0345	0.0396	0.0358
Iris (D3)	K-Means	0.1204	0.0612	0.0612	0.1102	0.1103	0.114	0.1011
	AEIKM	0.1155	0.0599	0.0599	0.1038	0.1038	0.1034	0.0965
	PCAHAEIKM	0.055	0.0407	0.0407	0.0535	0.0535	0.0668	0.0759

Table 2 Comparative Analysis of the Intra Cluster Distance D1 and D4

Datasets	Methods	No. of Clusters						
		2	**3**	**4**	**5**	**6**	**7**	**8**
Heart Disease (D1)	K-Means	12.6723	6.743	5.534	4.21	6.1233	6.83	5.792
	AEIKM	12.5261	6.2281	5.2655	3.8188	5.9623	6.6601	5.0407
	PCAHAEIKM	12.2363	5.7264	4.4729	3.3277	4.3881	4.3067	2.9712
Wine (D4)	K-Means	7.845	8.424	6.4122	7.1301	4.7002	1.4523	5.1402
	AEIKM	7.229	8.3187	6.3211	7.0383	4.2669	1.3806	5.0233
	PCAHAEIKM	1.4639	1.4108	1.5033	1.2289	1.1368	1.1791	1.2112

Table 3 Comparative Analysis of the Execution Time in Second

Datasets	Methods	No. of Clusters						
		2	3	4	5	6	7	8
D1	AEIKM	0.083079	0.141423	0.157602	0.159408	0.193956	0.18536	0.211756
	PCAHAIEKM	0.081699	0.130487	0.142256	0.148058	0.186267	0.184486	0.190353
	K-Means	0.094699	0.150487	0.165256	0.168058	0.196267	0.194486	0.225353
D2	AEIKM	0.086072	0.11098	0.163334	0.165998	0.16809	0.178608	0.203497
	PCAHAEIKM	0.081145	0.102346	0.144753	0.158375	0.162628	0.163021	0.199462
	K-Means	0.095145	0.132346	0.168753	0.178375	0.172628	0.186821	0.21462
D3	AEIKM	0.043034	0.043995	0.0446	0.1019789	0.11201	0.112891	0.128458
	PCAHAEIKM	0.035173	0.04354	0.0415	0.081975	0.102258	0.103791	0.121398
	K-Means	0.044213	0.046455	0.0586	0.120975	0.122258	0.123791	0.147398
D4	AEIKM	0.07277	0.097863	0.1215157	0.131029	0.151821	0.169077	0.185373
	PCAAEIKM	0.0673068	0.085154	0.10774	0.13076	0.145083	0.157862	0.180895
	K-Means	0.083068	0.107154	0.14774	0.154316	0.179834	0.179862	0.226895

Heart Disease Dataset: D1,

User Knowledge Modeling Dataset=D2,

Iris Dataset=D3,

Wine Dataset=D4

Table 4 Comparative Analysis of Fitness

Datasets	K-Means	AEIKM	PCAHAEIKM	PSOHAEIKM
D1	65	76	84	135
D2	58	67	73	140
D3	82	89	89	164
D4	79	83	85	156

Heart Disease Dataset: D1,

User Knowledge Modeling Dataset=D2,

Iris Dataset=D3,

Wine Dataset=D4

Heart Disease dataset D1

Table 5 Comparative Analysis of the Variance Explained

Total Variance Explained						
Component	**Initial Eigenvalues(λ)**			**Extraction Sums of Squared Loadings**		
	Values of λ	% of Variance(σ^2)	Cumulative %	Total	% of Variance	Cumulative %
1	3.097	23.823	23.823	3.097	23.823	23.823
2	1.578	12.139	35.962	1.578	12.139	35.962
3	1.261	9.702	45.664	1.261	9.702	45.664
4	1.108	8.524	54.188	1.108	8.524	54.188
5	1.005	7.728	61.916	1.005	7.728	61.916
6	0.877	6.750	68.666	-	-	-
7	0.837	6.438	75.104	-	-	-
8	0.752	5.784	80.888	-	-	-
9	0.683	5.253	86.140	-	-	-
10	0.559	4.303	90.443	-	-	-
11	0.464	3.566	94.010	--	-	-
12	0.417	3.210	97.219	-	-	-
13	0.361	2.781	97.778ss	-	-	-
14	0.311	2.222	100	-	-	-

User Knowledge Modeling Dataset D2

Table 6Comparative Analysis of the Variance Explained

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.382	27.640	27.640	1.382	27.640	27.640
2	1.173	23.451	51.091	1.173	23.451	51.091
3	.973	19.454	70.545			
4	.915	18.305	88.850			
5	.558	11.150	100.000			

Extraction Method: Principal Component Analysis.

Iris Dataset D3

Table 7 Comparative Analysis of the Variance Explained Dataset D3

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.921	73.016	73.016	2.921	73.016	73.016
2	.918	22.949	95.965			
3	.141	3.518	99.484			
4	.021	.516	100.000			

Extraction Method: Principal Component Analysis.

Wine Dataset

Table 8 Comparative Analysis of the Variance Explained Dataset D4

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.706	36.199	36.199	4.706	36.199	36.199
2	2.497	19.207	55.406	2.497	19.207	55.406
3	1.446	11.124	66.530	1.446	11.124	66.530
4	.919	7.069	73.599			
5	.853	6.563	80.162			
6	.642	4.936	85.098			
7	.551	4.239	89.337			
8	.348	2.681	92.018			
9	.289	2.222	94.240			
10	.251	1.930	96.170			
11	.226	1.737	97.907			
12	.169	1.298	99.205			
13	.103	.795	100.000			

Extraction Method: Principal Component Analysis.

F-Ratio

Herat Disease Dataset

Table 9 Comparative Analysis of the F-Ration Dataset D1

NCSS 2021, v21.0.1 2/22/2021 10:30:22 PM 4

K-Means Cluster Analysis Report

Dataset C:\Users\Kalam\Desktop\k\heart.NCSS
Computer-Generated Random Seed: 8102271

F-Ratio Section

Variables	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
Age_in_years	3	293	932.3322	73.19019	12.74
X_1___male__0___female_	3	293	15.74588	0.06051897	260.18
Chest_pain_type	3	293	16.55704	0.7709596	21.48
Resting_blood_pressure__in_mm_Hg_on_admission_to_the_hospital_	3	293	1838.158	299.9271	6.13
Serum_cholestoral_in_mg_dl	3	293	2435.536	1823.511	1.34
Fbs_fasting_blood_sugar__gt__120_mg_dl___1___true__0___false_	3	293	10.62702	0.01670084	636.32
Resting_electrocardiographic_results	3	293	3.565037	0.9634864	3.70
Maximum_heart_rate_achieved	3	293	17432.14	353.2181	49.35
Exercise_induced_angina__1___yes__0___no_	3	293	5.684576	0.1647308	34.51
ST_depression_induced_by_exercise_relative_to_rest	3	293	32.60195	0.6360666	51.26
The_slope_of_the_peak_exercise_ST_segment	3	293	7.715275	0.3070717	25.13
Number_of_major_vessels__0_3__colored_by_flourosopy	3	293	20.50617	0.6807207	30.12
X3___normal__6___fixed_defect__7___reversable_defect	3	293	133.1428	2.433525	54.71
TargetX1_or_0	3	293	78.62064	0.7347307	107.01

Fuzzy

Table 10 Comparative Analysis of Silhouette Dataset D1

Summary Section

Number Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	1336.969288	0.172883	0.5013	0.0025	0.4602	0.9204
3	887.187572	0.143425	0.3422	0.0133	0.5547	0.8320
4	665.798047	0.074896	0.2558	0.0077	0.6614	0.8819
5	532.288185	0.058233	0.2056	0.0070	0.7160	0.8950

Hybridized via PCA

Table 11 Comparative Analysis of the F-Ration Dataset D1

K-Means Cluster Analysis Report

Dataset C:\Users\Kalam\Desktop\k\HEART _PCA.NCSS
Computer-Generated Random Seed: 7444534

Cluster Standard Deviations (Continued)

Variables	Cluster3	Cluster4
age_in_years	6.373155	6.280147
X_1__male__0__female_	0.492375	0.4606464
chest_pain_type	0.9450056	0.9987797
resting_blood_pressure__in_mm_Hg_on_admission_to_the_hospital	18.47572	19.85297
serum_cholestoral_in_mg_dl	39.15715	43.632
fbs_fasting_blood_sugar__gt__120_mg_dl__1__true__0__false	0	0
resting_electrocardiographic_results	0.9201119	0.9877296
maximum_heart_rate_achieved	19.23597	22.09729
exercise_induced_angina__1__yes__0__no_	0.1542807	0.4800915
Count	83	41

F-Ratio Section

Variables	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
age_in_years	3	293	3469.018	47.2173	73.47
X_1__male__0__female_	3	293	3.123318	0.1897602	16.46
chest_pain_type	3	293	18.03772	0.7557991	23.87
resting_blood_pressure__in_mm_Hg_on_admission_to_the_hospital	3	293	4021.135	277.5758	14.49
serum_cholestoral_in_mg_dl	3	293	14395.83	1701.051	8.46
fbs_fasting_blood_sugar__gt__120_mg_dl__1__true__0__false	3	293	11.60596	0.006677548	1738.06
resting_electrocardiographic_results	3	293	10.91227	0.8882588	12.29
maximum_heart_rate_achieved	3	293	15887.33	369.0353	43.05
exercise_induced_angina__1__yes__0__no_	3	293	15.50666	0.06416339	241.67

Iris Dataset d3

Table 12 Comparative Analysis of the F-Ration Dataset D3

NCSS 2021, v21.0.1

K-Means Cluster Analysis Report

Dataset C:\Users\Kalam\Desktop\k\iris.NCSS
Computer-Generated Random Seed: 5281802

Cluster Standard Deviations

Variables	Cluster1	Cluster2	Cluster3	Cluster4
sepal_length	0.3484172	0.459272	0.4806861	0.3602097
petal_length	0.1736709	0.7284569	0.5535599	0.4803273
sepal_width	0.3487587	0.1965063	0.2692811	0.1958863
petal_width	0.1080123	0.2754375	0.2886609	0.2838223
Count	49	22	29	50

F-Ratio Section

Variables	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
sepal_length	3	146	26.36147	0.1581092	166.73
petal_length	3	146	143.7958	0.222441	646.44
sepal_width	3	146	5.817593	0.07232754	80.43
petal_width	3	146	26.11542	0.0577636	452.11

Hybridized via PCA

Table 13 Comparative Analysis of the F-Ration Dataset D3

NCSS 2021, v21.0.1

K-Means Cluster Analysis Report

Dataset C:\Users\Kalam\Desktop\k\IRIS_PCA.NCSS
Computer-Generated Random Seed: 7145863

Cluster Standard Deviations

Variables	Cluster1	Cluster2	Cluster3	Cluster4
sepal_width	0.2466001	0.2109502	0.2552318	0.2700362
petall_width	0.6115317	0.1472556	0.5810422	0.1855715
petall_length	0.3042691	0.09176629	0.2666405	0.1080655
Count	46	20	54	30

F-Ratio Section

Variables	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
sepal_width	3	146	6.287773	0.06266631	100.34
petall_width	3	146	142.577	0.2474842	576.11
petall_length	3	146	26.11561	0.0577596	452.14

Wine Dataset

Table 14 Comparative Analysis of the F-Ration Dataset D4

K-Means Cluster Analysis Report

Dataset C:\Users\Kalam\Desktop\k\wine.NCSS
Computer-Generated Random Seed: 7108727

F-Ratio Section

Variables	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio
Alcohol	3	174	23.57343	0.263987	89.30
Malic_acid	3	174	25.37881	0.8319673	30.50
Ash	3	174	0.6399308	0.06552901	9.77
Acl	3	174	220.5429	7.54251	29.24
Mg	3	174	2544.698	163.6323	15.55
Phenols	3	174	14.91474	0.1412921	105.56
Flavanoids	3	174	46.77177	0.208511	224.31
Nonflavanoid_phenols	3	174	0.3086926	0.01043339	29.59
Proanth	3	174	8.227118	0.191396	42.98
Color_int	3	174	186.1578	2.257495	82.46
Hue	3	174	1.679994	0.02418032	69.48
OD	3	174	21.21365	0.147025	144.29
Proline	3	174	4138109	29529.79	140.13

Wine Dataset

Table 15 Comparative Analysis of the Silhouette Report Dataset D4

Fuzzy Clustering Report

Dataset	C:\Users\Kalam\Desktop\k\wine.NCSS
Variables	Alcohol to Proline
Distance Type	Euclidean
Scale Type	None

Summary Section

Number Clusters	Average Distance	Average Silhouette	F(U)	Fc(U)	D(U)	Dc(U)
2	3597.023745	0.632629	0.6798	0.3597	0.1119	0.2238
3	2208.593553	0.541184	0.5264	0.2895	0.2149	0.3223
4	1573.473612	0.512456	0.4784	0.3046	0.2101	0.2802
5	1197.959414	0.497245	0.4110	0.2638	0.2802	0.3502

Procedure Input Settings

Autosaved Template File
C:\Users\Kalam\Documents\NCSS 2021\Procedure Templates\Autosave\Fuzzy Clustering - Autosaved 2021 2 10-19 49 46.t7

APPENDIX-II

Copies of Manuscripts

ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **10** (2021), no.1, 535–542
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.10.1.53 Spec. Iss. on ICIRPS-2020

ANALYSIS THE CLUSTER PERFORMANCE OF REAL DATASET USING SPSS TOOL WITH K-MEANS APPROACH VIA PCA

Muhammad Kalamuddin Ahamad[1] and Ajay Kumar Bharti

ABSTRACT. Partitioning problems are handled by the idea of cluster and this technique which plays the essential work in mining of data from the given dataset. The K-Means cluster is well accepted theory to apply on huge datasets, but has some drawbacks. The factual dataset is taken from the repository of data used for clustering. Furthermore, as getting the outcome of this procedure is essential to resolve the limitations and quality enhanced of cluster by apply the Principal Component Analysis (PCA) on the dataset. In paper we have demonstrate the results by experimental for factual datasets with dissimilarities. We have worked to validate the experimental significant for the clusters metric and component size minimized for different dataset during the processing on SPSS tool on the basis of eigenvalues. In this research paper we also discussed the comparative analysis of distance between initial centroid of wine and disease of heart dataset at the level of cluster k=2 and k=3.

1. INTRODUCTION

The emerging new trends, technology and growth of business through the internet services, another way said enriched the huge amount of resources of

[1]*corresponding author*

2020 *Mathematics Subject Classification.* 68T09, 94A16, 91C20.

Key words and phrases. K-Means, Principal component analysis, Dimensionality,Centroid, Eigen values.

Submitted: 18.11.2020; *Accepted:* 25.12.2020; *Published:* 22.01.2021.

the dataset such as storage of databases, audio, video, graphics, and images. In addition to the data consists the own characteristics like that consistent, structured, unstructured, uncertain, mixed, enormous, self-motivated and more complexes analyze and considerate of the people. Therefore, in the research field of data mining the more important to how investigate and self-mining of implicit, unidentified and more vital knowledge it can manage the support like in administrative behaviors. Therefore, its removal and detection of knowledge from the business database is helpful better quality of clustering. In the extraction and analyzed of data from the huge dataset to apply the statistical tool with employed the concepts of artificial intelligence. The data mining research field is very supportive in e-business and its applications to require a demo of functionality for industries [1], applications in business like as promoting, marketing, advertising [2]. There are author discussed the k-means clustering approaches and also proposed the performance in [3]. In this research paper the PCA techniques utilize on numerical attributes on the database the noisy feature reduces the dimensions of problem consider as the dataset but improve the cluster quality on the basis of the distance between initial centroids. This paper is arranged as follows manner. In Section 1- Introduction, Section 2- Literature Review, Section 3- Proposed Research Methodology, in the section-4 Brief the experimental effects. In the Section -5 discuss the conclusion and future scope.

1.1. **Mathematical illustration of coefficient Matrix.**

Definition 1.1. *Consider that the set of data value X is consisting nonempty set members of attribute m, extraction of data sample n then determine the mean and normalized of all existing members of attribute respectively. Illustrate in the mathematical form as equation-1 and normalized equation follow as equation-2*

$$\mu = 1/n(\sum_{i=1}^{n} x_i) = 0,$$

$$N = 1/(n-1)(\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij})^2) = 1,$$

where represent the value of x_{ij} is normalized at j=1,2,3,. . . ,m.

Definition 1.2. *Consider the consisting of two dimensional datasets of nonempty attribute X_{ij}, where i= 1to n, and j=1 to m, further associated the covariance*

matrix Coefficient at the m attribute is

$$\begin{bmatrix} Cov_{11} & Cov_{12} & \cdots & Cov_{1m} \\ Cov_{21} & Cov_{22} & \cdots & Cov_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Cov_{n1} & Cov_{n2} & \cdots & Cov_{nm} \end{bmatrix},$$

where the notation Cov_{ij} is the covariance coefficient between the X_i and X_j and also it's mentioned by C.

1.2. Evaluate the Covariance Matrix and Eigenvalues. Find the characteristic value of the characteristic equation $|\lambda I–C = 0|$and find the eigenvalues λ_j. This is sorted as $\lambda_1, \lambda_2, \ldots, \lambda_m$, and further find the orthogonal eigenvector. In this research paper simulated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ related to pth principal component$(p \leq m)$.Therefore the contribution eigenvalue rate is higher than value 1.00 to pick from the considering dataset.

2. Literature review

The cluster is creating to reveal the more in sequence of co- linearity, multi-co-linearity, and regression and correlation between the attribute of the dataset. Finding the total consequence in statistical examination of data is to reflect several points to have common characteristics. In the consequences of large dimensional dataset to bothered the mining process of information and not well shape clusters. In this research paper, we illustrate with the lower dimensionality of data by principal component analysis.

The author used PCA concept is reduced dimensions for changing the original data for mining and classifies it's by using the k-means. Find out the consequences to illustrate the more accuracy of reduced dimensionality of large dataset for analysis [4]. Worldwide cause deaths are in among the woman with breast cancer, and prediction of its disease to possibility of treatment otherwise very risky for health author discussed in [5]. PCA analysis tool the summarized of giving regular set patterns; evaluate the deviation of different variables, covariance and performance of dataset [8].

In telecom business the characteristics of huge datasets for main motive to discover such as reduced of real dataset, minimize the users clustering, analysis [6]. The statistics of dataset dimensionality has set of attribute and such variety

of data used in the research study and mining concepts are employed in this field like as telecommunication industries for helping the administrative strategy [7]. The significant concept of network is represented in a lower dimension to protect the structure of network node explained in [9]. The cluster property is explained and also discussed removal of k-1 term of covariance matrix and PCA project the higher to lower dimensional space, data placement in lower space graph and applied the clustering algorithm k-means [10].

2.1. **K-Means Algorithm.** This algorithm to apply on a considering d-Dimensional dataset, Choose k-prototype centroid at random from data point, Create the early dividing of cluster by assigning the object to the closest Centroid, and finally create the cluster[4-5].

2.2. **Principal Component Analysis (PCA).** The Principal component analysis is can only extract a linear projection of the data. Consider the consisting of data like as $X = x_1, x_2, \ldots, x_M$ are M vectors authors explained in[3,10]. PCA is described consisting of few steps as follows.

Step 1: Initially determine mean of data the given data set as

$$\mu = 1/M(\sum_{i=1}^{M} x_i).$$

Step 2: In the second step find subtract of mean from each individual data element Subtraction, therefore represent in the mathematical term as

$$\bar{x} = \sum_{i=1}^{M} (x_i \breve{} \mu).$$

Step 3: Measure the matrix of covariance C as

$$C = 1/M(\sum_{i=1}^{M} (\bar{x_i})(\bar{x_i})^T).$$

Step 4: Compute the eigenvalue and eigenvector $CX = \lambda X$, where $\lambda = \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ are eigenvalues and C is covariance matrix.

Step 5: Reduced the dimension of dataset.

Steps 6: Return the reduced dataset for clustering process

3. Proposed research methodology

The dataset like disease of heart, and wine are generally available on UCI Irving repository of machine learning archive. These factual datasets are retrieve from path of repository is mentioned as https://archive.ics.uci.edu.ml/datasets. The dataset is contained the instances 297,178, number of attribute 14, 13 and multivariate type of data characteristics of heart disease and wine respectively. The proposed algorithm are discussed as following:

3.1. **Procedure of PCA on Tools.**

Statistical analyses of various datasets are iris, wine and heart disease. This data set is large the make cluster initially not good cluster quality. The PCA is good concept the reduction of dimensionality of the dataset. Examines of a dataset constructs to reduced dimensions. The adopted procedure in this research paper as following

Step 1: Initially set the name of variables or attributes, then after filling the data values into data view filed,

Step 2: Create the structured data set in 2D,

Step 3: Go to analyze the 2D structure data set until return the eigenvalues,

i. Select the dimension reduction tool factor,

ii. Select the coefficient of component from descriptive field,

iii. Select the Extraction (Method) generate the variance matrix based on eigenvalues, eigenvalues set >1 and maximum iteration of convergence set at 25

iv. Press OK,

Step 4: Return the eigenvalues,

Step 5: Stop the procedure

3.2. **KMWPCA Algorithm.**

Step 1: Consider the arranged dataset, and after applied the reduction tools from SPSS,

Step 2: Extraction of some component by using the Step 1,

Step 3: Consider the element of variance and initial eigenvalues,

Step 4: To projecting data in lower dimensional subspace, getting reduced the dimension of real datasets,

Step 5: After reduction component of dataset to analyze K-Mean to achieve the distance between initial centroid for best clustering,

Step 6: Stop the process.

4. Discuss the Experimental Results

In research studies a concept to analyzed PCA implementation on SPSS Statistics 17.0 tools. Measure the eigenvalues of heart disease dataset by PCA and it's implemented on this given tools shown the results in table-1 and also analysis the simulating result is illustrated of same dataset in figure-1.

Table 1: Measure the eigenvalues at cluster k=3

Total Variance Explained at K=3

	Initial Eigenvalues(λ)			Extraction Sums of Squared Loadings		
Component	Total	% of Variance(σ^2)	Cumulative %	Total	% of Variance	Cumulative %
1	3.097	23.823	23.823	3.097	23.823	23.823
2	1.578	12.139	35.962	1.578	12.139	35.962
3	1.261	9.702	45.664	1.261	9.702	45.664
4	1.108	8.524	54.188	1.108	8.524	54.188
5	1.005	7.728	61.916	1.005	7.728	61.916
6	.877	6.750	68.666	-	-	-
7	.837	6.438	75.104	-	-	-
8	.752	5.784	80.888	-	-	-
9	.683	5.253	86.140	-	-	-
10	.559	4.303	90.443	-	-	-
11	.464	3.566	94.010	--	-	-
12	.417	3.210	97.219	-	-	-
13	.361	2.781	97.778ss			
14	.311	2.222	100	-	-	-

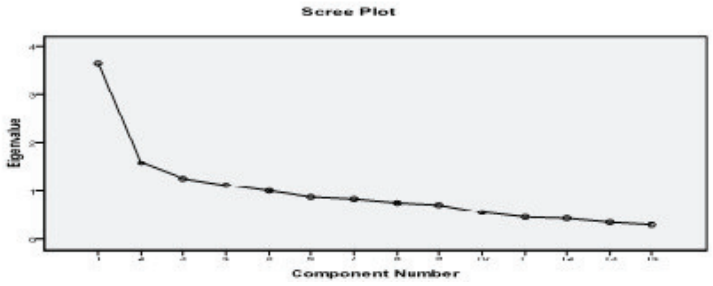
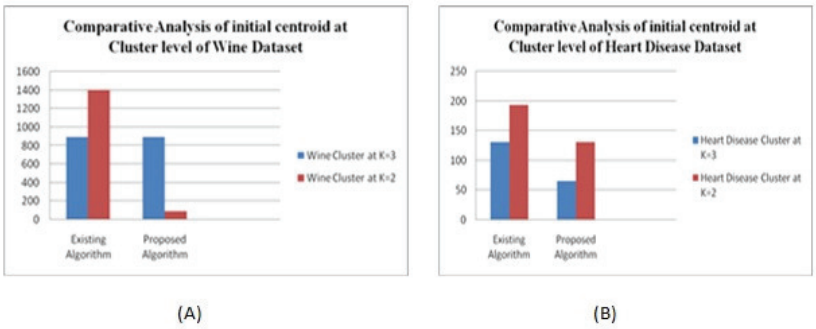


Figure 1. Analysis the extraction component from Heart Disease Dataset with eigenvalues

To measure the distance between initial centroid of given two dataset wine and heart disease respectively shown the result in below table-2. In figure-2, illustrate the comparatively analysis of existing(k-means algorithm) and proposed algorithm(k-means with PCA).

Table 2: Measure the distance between initial centroids of cluster

Datasets	Level of Cluster	Minimum distance between initial centroids	
		Existing Algorithm	Proposed Algorithm
Wine	K=3	895.845	895.844
	K=2	1402.192	92.143
Heart Disease	K=3	130.603	65.023
	K=2	193.811	131.031



(A) (B)

FIGURE 2. (A)Comparative analysis of centroids between existing and proposed algorithm(Wine dataset), (B)Comparative analysis of centroids between existing and proposed alogrithm(Heart disease dataset

In the figure-2, show the minimum distance of initial centriods at level of cluster k=3,wine, and heart disease datasets are smaller as compared than at level of cluster k=2.In this research paper, the proposed algorithm(k-means with PCA) is better than the existing algorithm(k-means) on the basis of minimum distance between the initial centroids and obtain good cluster.

5. CONCLUSION AND FUTURE SCOPE

We discuss and address in this study work lessening the component reduction and it's performance by the principal component analysis measured using SPSS statistics 17.0 tools. The dataset of heart disease are simulated on the reduction component tool. It is Lessening the dimension of the dataset by threshold value on the estimated eigenvalues. Furthermore, the simulation method is more significant and used successfully for partitioning the huge dimensionality for factual existing dataset and helpful for validating of clustering performance. In addition this paper shows the study of comparative study for existing algorithm with proposed algorithm at cluster level. If the cluster level increases then

the small distance between initial centeroid deceases and gets a well-defined cluster. This concept is to identify the appropriate causes for disease and support treatment of heart ailment, and also relevant information extract of wine dataset.

REFERENCES

[1] D.L. OLSON: *Data Mining in Business Services*, Service Business, **1**(3) (2007), 181-193.

[2] S. KUDYBA, R. HOPTROFF: *Data Mining and Business Intelligence: A Guided to Productivity*, IGI Global, 2001.

[3] C.F. TSAI, C.W. TSAI, H.C. WU, T. YANG: *ACODF: A Novel Data Custer Approach for Data Mining in Large Databases*, Journal of Systems and Software, **73**(1) (2004), 133-145.

[4] N. ZHANG, K. LEATHAM, J. XIONG, J. ZHONG: *PCA K-Means Based Clustering Algorithm for High Dimensional and Overlapping Spectra Signals*, In 2018 Ninth International Conference on Intelligent Control and Information Processing(ICICIP), (2008), 349-345.

[5] A. JAMAL, A. HANDAYANI, A.A. SEPTIANDRI, E. RIPMIATAIN, Y. EFFENDI: *Dimensionality Reduction using PCA and K-Means Clustering for the Breast Cancer Prediction*, Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, (2018), 192-201.

[6] M. ALKHAVRAT, M. ALJNIDI, K. ALJOUMAA: *A Comparative Dimensionality Reduction Study in Telecom Customer Segmentation using Deep Learning and PCA*, Journal of Big Data, **7**(1) (2020), art.id.9.

[7] I.M. AL-ZUBAI, A. JAFA, K. ALJOUMAA: *Predicting Customer's Gender and Age Depending on Mobile Phone Data*, Journal of Big Data, **6**(1) (2019), art.no.18.

[8] P.R. PERES-NETO, D. JACHSON, K.M. SOMERS: *How Many Principal Components? Stopping Rules for Determining the Number for Non -trivial Axes Revisited*, Computational Statistics and Data Analysis, **49**(4) (2005), 974-997.

[9] D. WANG, P. CUI, W. ZH: *Structural Deep Network Embedding*, In Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016), 1225-1234.

[10] C. DING, X. HE: *K-Means Clustering via Principal Component Analysis*, In Proceedings of the 21st International Conference on Machine Learning, (2004), art.no.9.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, MAHARISHI UNIVERSITY OF INFORMATION TECHNOLOGY & INTEGRAL UNIVERSITY, LUCKNOW, UTTAR PRADESH, INDIA.

Email address: ahamad_kalam@rediffmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, MAHARISHI UNIVERSITY OF INFORMATION TECHNOLOGY LUCKNOW, UTTAR PRADESH, INDIA.

Email address: ajay_bharti@hotmail.com

European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7 Issue 11, 2020

COMPARATIVE ANALYSIS THE FITNESS FUNCTION OF K-MEANS AND KERNEL FISHER'S DISCRIMINANT ANALYSIS (KFDA) WITH GENETIC ALGORITHM

Muhammad Kalamuddin Ahamad*

Department of Computer Science & Engineering,
Maharishi University of Information Technology (MUIT)
(& Integral University) Lucknow,Uttar Pradesh, India

Ajay Kumar Bharti

Department of Computer Science & Engineering
Maharishi University of Information Technology (MUIT), Lucknow,
Uttar Pradesh, India

Abstract*: *In the field of research, the growth of data mining using the k-means technique is well accepted, which involves the extraction of data from datasets with some limitations. To overcome the drawback of this technique we employed the kernel concepts and resolved the cluster inadequacy of separability. We have proposed an optimization technique to include a fisher's discriminant analysis into the kernel of particle swarm optimization concept with GA (Genetic Algorithm) to evaluate fitness function. The fitness function value is required to select offspring for the next generation. The consequence was to reduce the noise and enhance the performance of clustering. The GA (Genetic Algorithm) was employed to optimize the objective of the fitness function by providing the input parameter. The kernel technique performs more fault identification features than principal component analysis. Results found are more beneficial by this method like fitness value, stopping criteria, and the average distance between individuals. In this research paper, we discuss the comparative analysis with the objective function of k-means and kernel fisher's discriminant analysis in the domain of the large dataset. The fitness value of proposed KFDA is smaller than k-means fitness.

Keywords: K-Means, Cluster, Kernel Principal Component Analysis, Fitness function, Genetic Algorithm (GA).

1. Introduction

K-means technique is required to input data for creating the cluster, but it is a more tedious task to find the cluster. In GA, it creates automatically and picks a gene and their numbers of a gene are randomly generated. If the users are identified the correct gene at the initial population then the latter creates a good quality of cluster. With the help of fitness function and arrangement of gene and its operator makes the good quality of cluster centroids.

A genetic algorithm's idea explains the calculation of numerical easily solvable like a mathematical problem that has been presented in detail [1]. There are some procedures available for mining features and classification of multivariate datasets. In k-means, a null cluster is created with initial centroids its main drawback, but GA is applied heuristic search based on the natural selection, and the suggested hybrid k-means using GA removes the difficulty of creating empty clutters [2], and a chromosome is created from clustering k-means center[3]. The principal component analysis works a very crucial role classification of the dataset, but fisher's discriminant analysis produced to improve the result as compared to principal component analysis. KPCA is a nonlinear result of PCA and similar manner KFDA is the same as the FDA. In the linear nature of the dataset, classification occurs not better, but the kernel idea handles the non-linearity problem [4-5]. Kernel concept handles nonlinear problems to overcome few difficulties using of optimization technique for evolutionary computing. In this research paper we discussed a comparative analysis of the performance of fitness function.

The traditional k-means technique is generally required for clustering of huge data set because it's a very simple concept and more convergence of the data. This algorithm is more responding to the first centroid of the cluster. The cluster of huge data set is generally affected by the data point. This algorithm has few drawbacks, but the genetic algorithm is used to overcome the responsiveness of the first cluster centroid, reduced some data point's impact and gets more accuracy and high-quality of the cluster [6].This research paper is prepared as follows. In Section- 1 introduction, Section-2 discuss the brief of traditional K-Means algorithm, Genetic Algorithm (GA), Kernel of fisher discriminates of Particle

Swarm Optimization (PSO), Section-3 brief of the proposed work, Section- 4 explain the result in detail, and Section-5 brief the conclusions.

2. Literature Review

Author Yong et al., presented his research work at the 12th international conference ICCSE, IEE, to identify the problem and how we can enhance the minority class performance. They have assessed the criteria of the mixed dataset of two classes, there is one minority and secondly majority basis on the right and false classification. It is reflecting the performance of its classification and validates the conclusion by KNN with a Supporting Vector Machine (SVM) [7]. We have studied the analysis of three algorithms discussed in detail by the author [8] among them first genetic algorithm, second differential evolution, and third particle swarm optimization. Also, the genetic algorithm is more benefited for separate optimization over two algorithms.

Evaluate the most favorable result of difficulty facing by every family at present to how can handle the financial plan for accessing the cluster of economic and community behavior founded on K-Means and genetic algorithms [9] and producing many secure clusters for huge data set presented[10]. These methods combine to describe the manifold TSP (Traveling Salesman Problem), also create the high quality of cluster with GA techniques [11-12]. The route optimizing problem and congregate the global result in expressions of the accuracy, time of computing, and convergence speed for online real application [13], and more applications are discussed the GA (Genetic Algorithm) via K-Means [14], and survey [15].

There are n feature of dataset fall into clusters K for condition k less than n then the objective function set to minimize. The selected the cluster center is more carefully, this process is repeated going to the cluster center till does not vary, and aim of this technique minimized objective function and reduced the squared errors [16].

Genetic k-means algorithm it is special kind of clustering technique on distance based mutation, GKA is faster technique than rest algorithm using in cluster [17]. By comparing genetic algorithm and PSO we found the result PSO is better because it's confined a local and global searching at the similar time. The reflection of PSO is poor and undignified for smaller population size, but enduring to the time bound PSO is good [18].

2.1. Particle Swarm Optimization

A paper published in 1995 at the international conference on the evolution of computation. Introducing this research paper in the conference, then after a change is the scenario of using its paper of PSO conceptual theory to handle the various kinds of complex optimizer problem. This is a very easy and attractive concept to felicitate the global searching process [19-20].

In this method, the populace of the effect is recognized as a swarm of the particles and carried out the result indicated as the particle. Further, all particles have velocity and position. The Particle is being in the move to another position with velocity. When occurring the next position is the paramount imaginable position, then which is required to update both its location and velocity in presented [21, 23, and 25], and this procedure is repetitive until found the criteria. The process of this technique is represented in fig.-1.

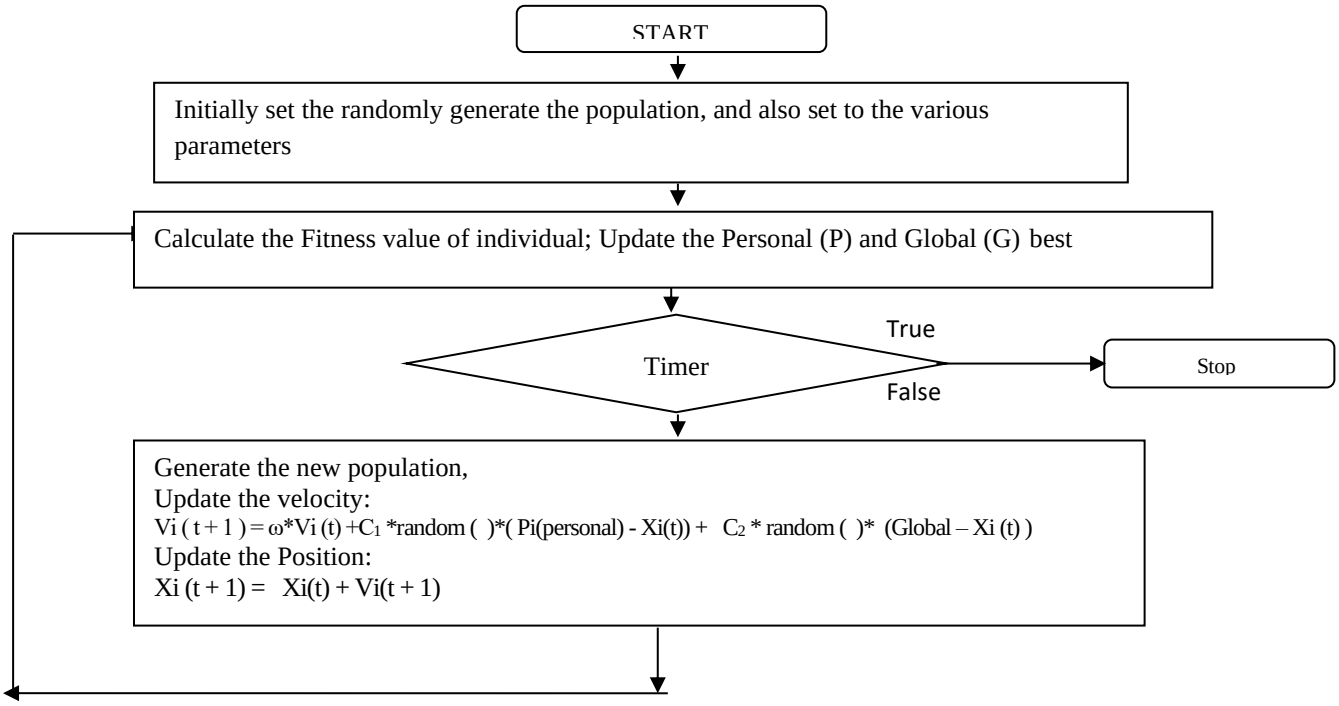


Fig.1 Flow chart the procedure of particle swarm optimization

2.2 Kernel Trick of KPCA

There are given a sample set of a data $A = \{a_1, a_2, a_3, \dots\dots, a_n\}$ and every member of a data point belonging in the domain field T^N. Nonlinear mapping represented as

$$\emptyset: R^N \to T, \quad (1)$$

Where, $a = \emptyset(a)$

Mercer's conditions

$$K(a_i, a_j) = \emptyset(a_i)^{transpose} \emptyset(a_j) \quad (2)$$

Where

R = Set of the domain

N= Number of attribute and its value 1, 2, 3…, n.

∅= represent the nonlinear function of mapping

T = the range of function

$K(a_i, a_j) = Kernel$ Function of input space

The established model kernel function optimization by the author Hongxia et al. is discussed in detail [25]. Consider the two datasets like

$$A_1 = \{a_{11}, a_{12}, a_{13}, \dots\dots\dots, a_{1i}\}, \text{ And}$$
$$A_2 = \{a_{21}, a_{22}, a_{23}, \dots\dots\dots, a_{2j}\},$$

From dataset A_1

$$\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \emptyset(a_{1i}) \quad (3)$$

From dataset A_2

$$\mu_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \emptyset(a_{2j}) \quad (4)$$

From equation (1) and (2)

$Ds = |\mu_1 - \mu_2|^2$

$= |\mu_1 - \mu_2|^{transpose} |\mu_1 - \mu_2|$

$$= |\frac{1}{n_1}\sum_{i=1}^{n_1}\emptyset(a_{1i}) - \frac{1}{n_2}\sum_{j=1}^{n_2}\emptyset(a_{2j})|^{tranpose}|\frac{1}{n_1}\sum_{i=1}^{n_1}\emptyset(a_{1i}) - \frac{1}{n_2}\sum_{j=1}^{n_2}\emptyset(a_{2j})|$$

$$= \frac{1}{n_1} * \frac{1}{n_2}(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\emptyset(a_{1i})\emptyset(a_{2j})) - 2*\frac{1}{n_1} * \frac{1}{n_2}(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\emptyset(a_{1i})\emptyset(a_{2j}))+\frac{1}{n_2} * \frac{1}{n_2}(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\emptyset(a_{1i})\emptyset(a_{2j}))$$

$$= \frac{1}{n_1} * \frac{1}{n_2}(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K(a_{1i}, a_{1j})) - 2*\frac{1}{n_1} * \frac{1}{n_2}(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K(a_{1i}, a_{2j}))$$

$$+\frac{1}{n_2}\frac{1}{n_2}(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K(a_{2i}, a_{2j})) \quad (5)$$

Where
μ_1= mean vector of one feature space F_1
μ_2= mean vector of one feature space F_2
F = feature space
n_1, and n_2 = size of dataset
Ds= square distance between the mean two spacesF_1, F_2

Determine the dispersion of two samples

$$\text{df1} = \sum_{i=1}^{n_1}|\emptyset(a_{1i}) - \mu_1|^2 = \sum_{i=1}^{n_1}\emptyset(a_{1i})^{transpose}\emptyset(a_{1i}) - n_1\mu_1^{transpose}\mu_1$$

$$= \sum_{i=1}^{n_1}K(a_{1i}, a_{1i}) - \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K(x_{1i,}\, a_{1j}) \quad (6)$$

$$\text{df2} = \sum_{j=1}^{n_2}|\emptyset(a_{2j}) - \mu_2|^{transpose}$$

$$= \sum_{j=1}^{n_2}\emptyset(a_{2j})^{transpose}\emptyset(a_{2j}) - n_2\mu_2^{transpose}\mu_2$$

$$= \sum_{j=1}^{n_2}K(a_{2j}, a_{2j}) - \frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K(a_{2i,}\, a_{2j}) \quad (7)$$

Where,
df1= dispersion within the sample of F1feature space
df2= dispersion within the sample of F2feature space
n_1 And n_2 = size of sample

2.3 Description of GA (Genetic Algorithm)

To develop the concept of a genetic algorithm by Goldberg who has inspired the idea of evolution theory proposed by C. Darwin's. In this theory, C. Darwin quotes the survival of an organ can be maintained through the procedure of crossover, reproduction, and also mutation. The evolution concept useful to the computational algorithm is identified usually to trend as alike objective function. A solution generated by a genetic algorithm is acknowledged as a chromosome, but collected works of these chromosomes are called the population. These chromosomes are compared from the Genes and find its either numerical value, value of binary stream, symbol value, or character depending on the complicatedness. These chromosomes are going through the procedure called fitness function, and find the appropriateness of problems generated by the GA. The higher fitness values of chromosomes have more possibility to prefer in the subsequent generation [1, 26, and 29]. The details about the procedure of this algorithm are available to propose the techniques by researcher Holland (1975) and by Goldberg (1989).

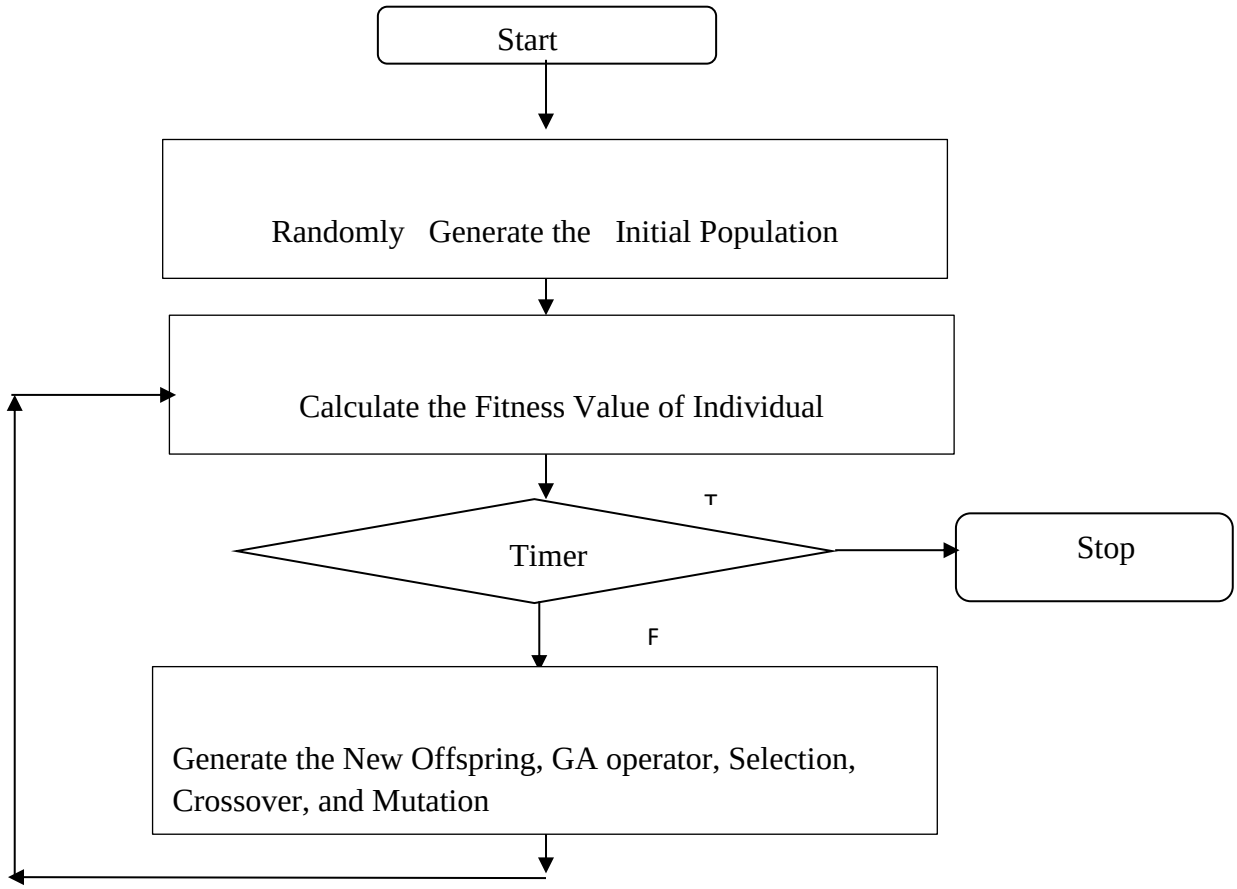


Fig. 2 The flow chart procedure of the genetic algorithm

There are discussed the operators of a GA as follows:

Selection Operator: In this concept offer the preference of a value of greatest fitness of the chromosome allowed to go for the following generation.

Crossover Operator: In this concept matching the individuals. The selecting of two individuals by the theory of selection operator and apply the crossover operator can be rearranged at the site creating the new individual known as the offspring.

Mutation Operator: In this concept inserts the genes in the offspring getting by crossover operator to maintain the size of a population to keep avoid the early convergence.

The procedures of generic algorithm summarize as follows:

Step 1: Set to initialize the value of the population with random values.
Step 2: The fitness function evaluate the number population
Step 3: Until the convergence repeat
I: Select the descendants from the population
II: Crossover and generate the new offspring
III: Apply the mutation of new offspring
IV: Determine the fitness of the new offspring

The flow chart of this algorithm is illustrated in above in fig. – 2

3. Proposed Work

The consequence of the objective function is to find from the kernel trick of fisher's discriminant analysis. In this function some parameter are required for the performance like as means vector of feature space and square of the distance. Given the size of datasets containing the two factors firstly row or instances and secondly columns or attribute.

We proposed an index of performance that defined the fitness function KFDA by PSO and compare it with the objective function of K-Means employ the genetic algorithm.

By using the maximum iteration and inertia define the relation in [19, 22, and 28], find the fitness function, and it is valuable for the separation between max and min of classifications by ω parameter. The ω is one parameter set at the min point of Fisher's Determinant Analysis (FDA) and it can vary if it changes the ω parameter. An objective function is set to optimized by PSO [27], from equation (5), (6), and (7)

$$F_{fitness\_function} = \frac{(df1+df2)}{Ds} \quad (8)$$

Author Dabbura, define the objective function of K-Means used in minimize the squared error [30],

$$F(x) = \sum_{j=1}^{K} \sum_{i=1}^{n} \left| a_i^j - c_j \right|^2 \quad (9)$$

Where,
c_j= centroids for j cluster,
K= number of cluster,
n= number of object,
a_i^j= i^{th} object in k^{th} cluster

The objective functions from the equation (8) and (9), to simulate defined the objective function by genetic algorithm for using the optimal tool for optimizing in MATLAB and set some parameter mentioned in below Table-1.

In this paper mentioned the exit criteria for pick up that produced the number of generations to be reached maximum (of a population) value the parameter set of option to measuring performance.

Table 1: Set the Optional Parameters Measuring the Performance

Parameters	Range/Value
Set Mutation	0.8
Generation of Random Number	[0,1]
Use Default Population	20
Size of Population	1000
Variables	2, 5
Set Size of Variable	10
Type of Population	Double vector

4. Result and Discussion

In research studies, to develop the concept of proposed fitness function and comparative analysis this function with the objective function of K-Means apply the concept of genetic algorithm. The fitness function is implemented on MATLAB Ra12013a optimizing tool of genetic algorithm.

The experimental set up of determining fitness function in problem solver Genetic Algorithm(GA), and fitness function defined @ problem fitness and set the number of the variable 05 (five).

For Objective function of Kernel FDA
Fitness function= @kalam_fitness, a number of iteration is 51 at number of variable is 02 (two) on the run solver view the result as following the fig.-3.

The optimization of running,
Objective function value: 2.81424828271717591E-4= 0.000281427

Optimization Terminated: Average change in fitness value less than option.

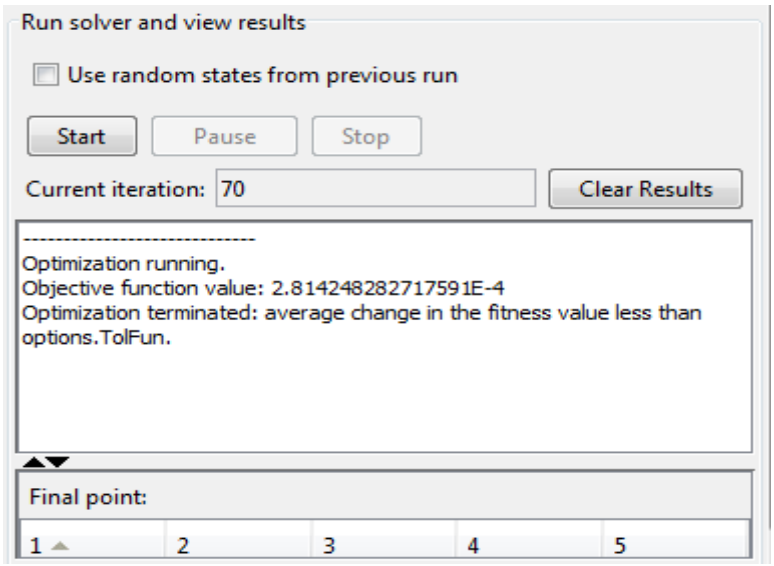


Fig. 3 Value of the objective function of KFDA at five variables

For Objective function of k-means
Fitness function= @kalam1_fitness, number of iteration is 51 at number of variable is 2 on the run solver view the result as following the fig.-4.

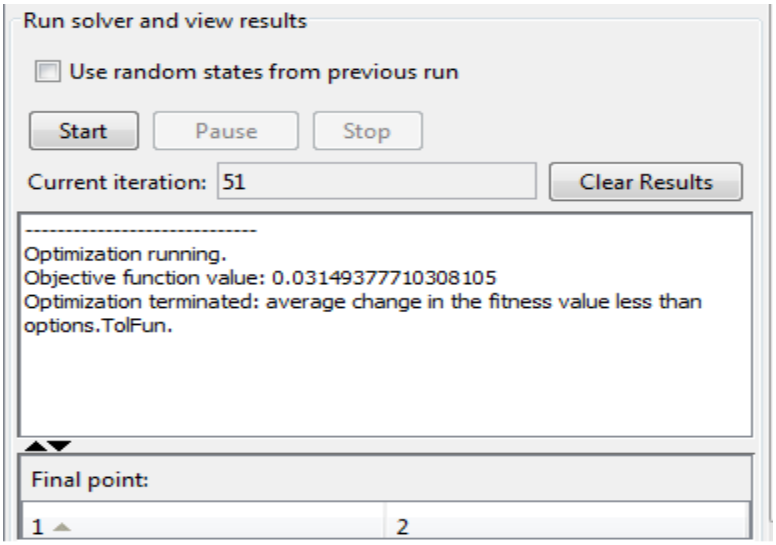


Fig. 4 Value of the objective function of k-means at two variables

The optimization of running,
Objective function value: 0.031493777710308150 = 0.0314938
Optimization Terminated: Average change in fitness value less than optimum.

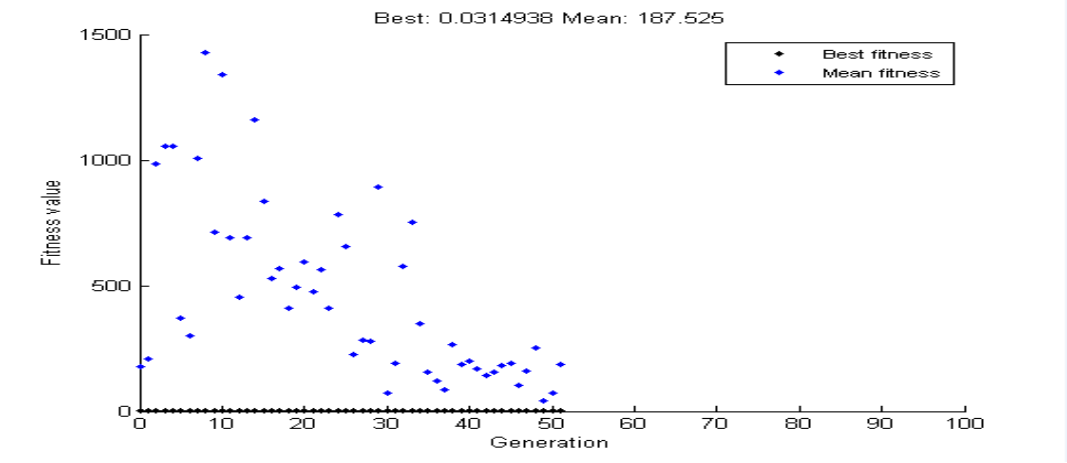
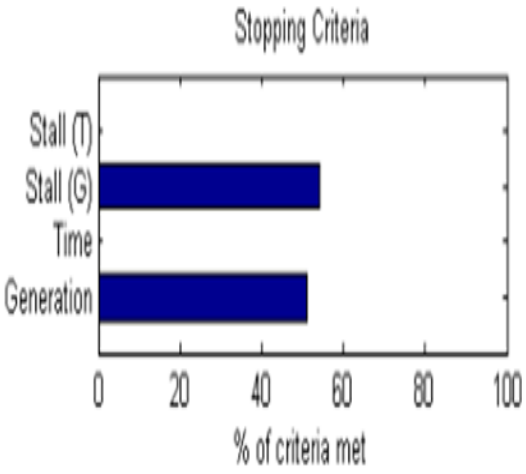


Fig. 5 Values of best fitness and mean for K-Means



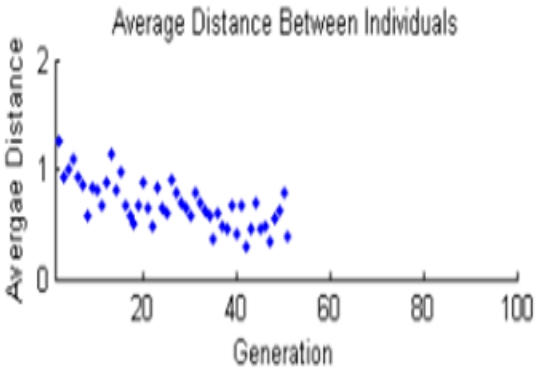
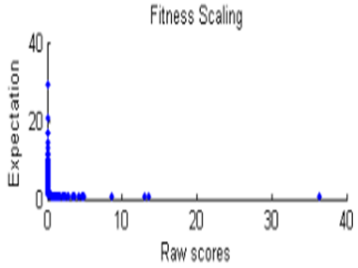


Fig. 6 Stopping criteria of K-Means

Fig. 7 Average distance between indvidual of K-Means



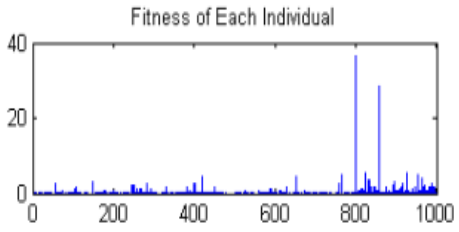


Fig. 8 Fitness scaling of K-Means

Fig. 9 Fitness of each indvidual of K-Means

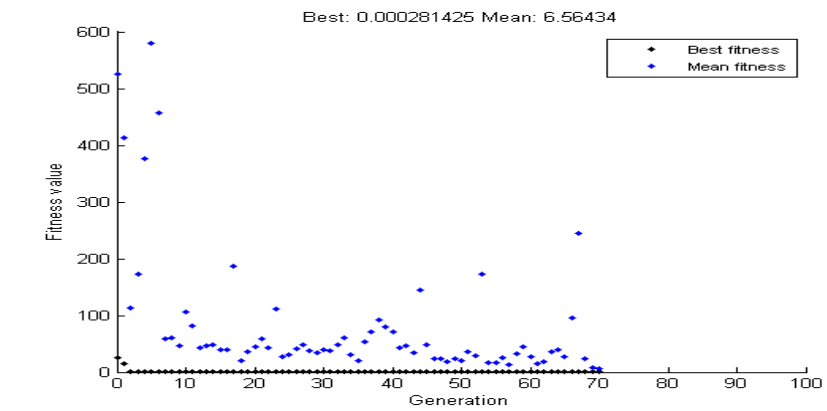


Fig. 10 Best fitness of KFDA

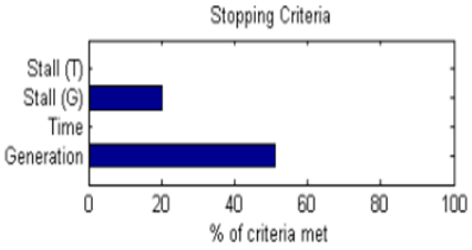


Fig. 11 Stopping criteria of KFDA

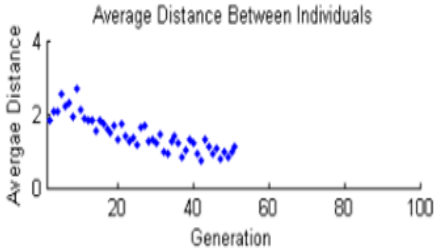


Fig. 12 Average distance between individual of KFDA

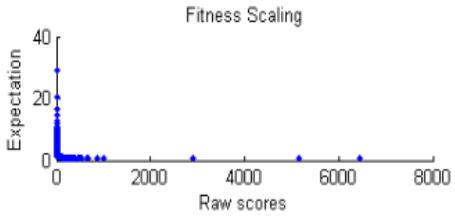


Fig.13 Fitness scaling of KFDA

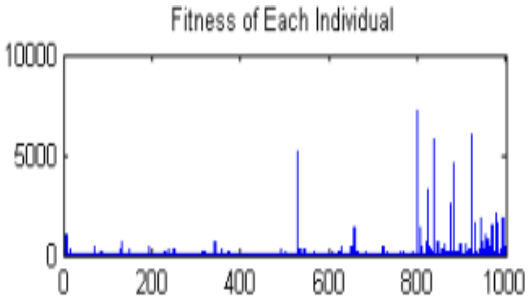


Fig. 14 Fitness of each indvidual of KFDA

Table 2 Analysis the objective function between k-means and proposed KFDA

Criteria of K-Means objective function	**Criteria of Proposed KFDA (PKFDA) objective function**
1. Fitness value = 0.0314938 2.Mean value=219.562 3.Expectation fitness scaling =30 4.Fitness of each Individual at 800 lies between 36 and 39 5.Stopping % criteria met S(G) is below 80 6.Stopping % criteria met S(T) is above 50 7.Average Distance between Individual approximate above 01(one)	1. Fitness value =0.000281427 2.Mean value=197.442 3.Expectation fitness scaling =35 4.Fitness of each Individual at 800 above 500 5.Stopping % criteria met S(G) is above 50 6.Stopping % criteria met S(T) is 20 7. Average Distance between Individual approximate above 02(two)

Therefore, we have set the fixed population for some attribute of a dataset: the double size of which represent as uniformly, operator crossover mutation set at 0.8 vector, size of the population: default value at 20, set the initial random generation rang [0, 1], and fitness scaling (R). In addition, to select a stochastic function and constraints mutations depending on the basis of fitness function but where the crossover function is scattered. There are the parameters n_1 and n_2 both, were set at 10. The fitness value, mean value, stopping % criteria met S(G) and S(T), average distance between individual, expectation fitness scaling, and fitness of each individual, are mentioned in the fig.-5, fig.-6, fig.-7, fig.-8, fig.-9 respectively of the k-means algorithm. In this paper, we proposed the objective function KFDA of more significant fitness value, mean value, stopping % criteria met S(G) and S(T), average distance between individual, expectation fitness scaling, and fitness of each individual are mentioned in the fig.-10, fig.-11, fig.-12, fig.-13, and fig.-14 respectively. And also analyze of comparative evaluation proposed objective function KFDA is more significant and preferable than the objective function of K-Means shown in Table 2.

5. Conclusions

Nowadays, the trend in the research work is focused on the clustering problem of datasets. In this paper, we have used the concept of kernel trick. The kernel idea is to enhance the performance of various types of datasets. In addition, the genetic algorithm applies for simulation results of the kernel fisher's discriminant analysis. The generated offspring will be selected for the next generation and supplied as fitness function values that are focused on the simulation process. The kernel FDA is superior and more significant as compared to other methods. This performance is more favorable in the classification of datasets. In this research paper, it is mentioned that the fitness value of an objective function in terms of best fit and means, stopping criteria, and average distance between individual of the simulation process. The comparative analysis criteria of objective function KFDA is smaller than an objective function of K-Means. The exit criteria are the selection when the number of generation produced reaches the maximum (of population) value.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for publication.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing the original draft, preparation, writing and review editing, visualization, have been done by the first author. The supervision and project administration have been done by the second author.

References:

[1] Hermawanto D.Genetic Algorithm for Solving Simple Mathematical Equality Problem. *arXiv preprint arXiv*. 2013; 1308.4675.
https://arxiv.org/ftp/arxiv/papers/1308/1308.4675.pdf.

[2] Al Malki A., Rizk M. M., El-Shorbagy M. A., Mousa A. A.Hybrid genetic algorithm with k-means for clustering problems. Open Journal of Optimization. 2016; 5(02); 71.

[3] Oujezsky V., Horvath, Traffic similarity observation using a genetic algorithm and clustering. Technologies. 2018; 6(4); 103.

[4] Kemsley E. K. Discriminant Analysis of High Dimensional Data: A Compression of Principal Component Analysis and Partial Least Square Data Reduction Methods. Chemo Metrics and Intelligent Laboratory Systems.1966; 33; 47-61.

[5] Sayed E. H., Gabbar H. A., Miyazaki S. :Improved Evolving Kernel of Fisher's Discriminant Analysis for Classification Problem. Journal of Applied Sciences. 2009*;* 9(12); 2313-23.

[6] Min W., Siqing Y.Improved K-Means clustering based on Genetic Algorithm. In 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). IEEE. 2010; 6; V6-636. doi:10.1109/ICCASM.2010.5620383.

[7] Yong Y., Xincheng G.A New Minority kind of Sample Sampling Method Based on Genetic Algorithm and K-Means Cluster. In 2012 7th International Conference on Computer Science & Education (ICCSE).IEEE.2012;126-129.doi:10.1109/ICCSE.2012.6295041.

[8] Kachitvichyanukul V. Comparison of Three Evolutionary Algorithms: GA, PSO and DE. Industrial Engineering & Management Systems.2012*;* 11(3); 215-223

[9] Babaie S. S., Mahdi E. E. O., Firoozan T. A Novel Combined Approach of K-Means and Genetic Algorithm to Cluster Cultural Goods in Household Budget. In Proceeding of 4th International Conference on Frontiers in Intelligence Computing: Theory and Applications (FICTA). Springer, New Delhi. 2015; 273-283. https://doi.org/10.1007/978-81-322-2695-6_24.

[10] Bhatia S. New Improved Technique for Initial Cluster Centers of K-Means Clustering using Genetic Algorithm. In International Conference for Convergence for Technology (ICCT 2014), IEEE .2014; 1-4. doi:10.1109/12CT.2014.7092112.

[11] Rahman M. A., Islam M. Z. A Hybrid Clustering Technique Combining a Novel Genetic Algorithm with K-Means. Knowledge Based Systems.2014; 71; 345-365.

[12] Lu Z., Zhang K., He J., Niu Y. Applying K-Means Clustering and Genetic Algorithm for Solving MTSP. In International Conference on Bio-Inspired Computing: Theories and Applications, Springer, Singapore.2016; 278-284. https://doi.org/10.1007/978-981-10-3614-9_34.

[13] Aibinu A. M., Salau H. B., Rahman N. A., Ackachukwu C. M. A Novel Clustering Based Genetic Algorithm for Route Optimization. Engineering Science and Technology an International Journal.2016; 19(4); 2022-2034.

[14] Zeebaree D. Q., Haron H., Abdulazeez A. M., Zeebaree S. R. Combination of K-Means Clustering with Genetic Algorithm: A Review. International Journal of Applied Engineering Research.2017*;* 12(24); 14238-14245.

[15] Ahamad M. K., Bharti A. K. An Effective Technique on Clustering in Perspective of Huge Dataset. International Journal of Recent Technology and Engineering (IJRTE). 2020; 8(6); 4485-4491.

[16] Krishnasamy, G., Kulkarni, A. J., Paramesran, R.. A hybrid approach for data clustering based on modified cohort intelligence and k-means. Expert Systems With Applications. 2014; 41(13),6009-6016.

[17] Krishna K., Murty M. N. Genetic k-means algorithm. IEEE Transactions on Systems, Man, and (Cybernetics). 1999; 29(3); 433-349.

[18] Prashanth N. A., Sujatha P.Comparision Between PSO and Genetic Algorithms and for Optimizing of Permanent Magnet Synchronous Generator(PMSG) Machine Design. International Journal of Engineering & Technology.2018; 7(3.3); 77-81.

[19] Kennedy J., Eberhart R.: Particle Swarm Optimization. In Proceeding of ICNN95 International Conference on Neural Networks. IEEE. 1995; 4; 1942-1948. doi:10.1109/ICNN.1995.488968.

[20] Tharwat A., Gaber T., Hassanien A. E., Elnaghi B. E.Particle Swarm Optimization: A Tutorial. In Hand book of Research on Machine Learning Innovations and Trends, IGI Global. 2017; 614-635. doi:10.4018/978-1-5225-2229-4.ch026.

[21] Van der Merwe D., Engelbrecht A. P. Data clustering using Particle Swarm Optimization. The 2003 congress on Evolution Computation CEC'03, IEEE. 2003; 1; 215-220.doi:10.1109/CEC.2003.1299577.

[22] Shi Y., Eberhart R. A Modified Particle Swarm Optimizer. Applied Mathematics and Computation.1998; 189(5); 69-73.

[23] Zhao W., Zu W., Zeng H.: A modified Particle Swarm Optimization via Particle visual Modeling Analysis. Computers & Mathematics with Applications.2009*;* 57(11-12); 2022-2029.

[24] Hongxia P., Xiuye W., Jinying H. Study of Fault Extraction Based on KPCA Optimized by PSO Algorithm. In International Conference on Fuzzy Systems. IEEE. July 2010;1-6.doi :10/1109/FUZZY.2010.5583947.

[25] Xinchao Z. A Perturbed Particle Swarm Algorithm for Numerical Optimization. Applied Soft Computing. 2010; 10(1); 119-124.

[26] Gen M., Cheng R.: Genetic Algorithm and Engineering Design. John Wiley & Sons, Inc., New York. 1997.

[27] Wei X., Pan H., Wang F. Feature Extraction Based on Kernel Principal Component Analysis Optimized by PSO Algorithm. Journal of Vibration, Measurement and Diagnosis. 2009; 29(9); 162-166.

[28] He Y., Wang Z. Regularized Kernel Function Parameter of KPCA Using WPSO-FDA for Feature Extraction and Fault Recognition of Gearbox. *Journal of Vibroengineering.2018;* 209(1); 225-239.

[29] Chang D. X., Zhang X. D. Zheng C. W.A Genetic Algorithm with GENE Rearrangement for K-Means Clustering. Pattern Recognition.2009*;* 42(7); 1210-1222.

[30] Dabbura I. K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Towards Data Science. 2018. Saatavissa: https://twardsdatascience.com/k-means-clustering-algorithm-evaluation-methods-and-drawbacks-aa03e644b48a