Design a new clustering algorithm for partition clustering problem

Project Report submitted in partial fulfillment of the requirement for the degree of

Master of Technology

in

Computer Science & Engineering

under the Supervision of

Dr. Pardeep Kumar and Dr. Yugal Kumar

By

Pavika Bhardwaj (182201)

Jaypee University of Information Technology

Waknaghat, Solan – 173234, Himachal Pradesh

CERTIFICATE

This is to certify that project report entitled **"Design a new clustering algorithm for partition clustering problem"**, submitted by PavikaBhardwaj in partial fulfillment for the award of degree of Master of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been made under our supervision.

This report has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Pardeep Kumar

Date: 30th May,2020 **Dr. Pardeep Kumar**

(Associate Professor)

Date: 30th May,2020 **Dr.Yugal Kumar**

(Assistant Professor)

ACKNOWLEDGEMENTS

I am highly obliged to **Dr. Pardeep Kumar and Dr.Yugal Kumar** for the valuable information provided by them in this respective field. It has been a great privilege for me to have them as a mentor for this project. I take this opportunity to express my profound gratitude and deep regards to **Prof. Dr. Samir Dev Gupta, HOD (CSE&IT)** for his exemplary guidance and constant encouragement throughout the course of this project work. Lastly, I thank almighty, my parents, family and friends for their constant encouragement without whom this work would not have been completed.

Date: 30th May, 2020

Signature:

Pavika Bhardwaj

Table of Contents

List of Figures

List of Tables

Abstract

Machine learning is a set of techniques which allow a machine to act as human beings. There are various essential components of machine learning knowledge pyramid. It includes symbols, facts, data, information, knowledge, intelligence and wisdom. Machine learning algorithms are searched with optimization. They are predictive in nature. They are least dependent on the user. They are applied on huge collection of data. Data mining act as an application of it. Though K-Means is the simplest technique of clustering to be used, still it has certain drawbacks. This project mainly deals with using harris-hawk meta-heuristic optimization technique in clustering. They provide an edge over traditional partitioning techniques because of its successful implementation and high intensity. The project aims to obtain optimized cluster centres. “Hawks” represent the number of clusters needed. “Location of the rabbits” are represented as initial and final cluster centres. The proposed algorithm is further evaluated on two parameters namely accuracy and intra-cluster distance. It leads to high accuracy and low intra-cluster distance.

CHAPTER 1

INTRODUCTION

1.1 General

Clustering is an unsupervised machine learning technique in which unidentified class labels are used. It groups data objects in clusters with the help of distance measure. Distinct clusters are made whereby within each cluster alike objects are found. No data object is found to be similar in case of two different clusters. It is a good technique for discovering concealed patterns in the core data. The purpose of clustering is to find out dense and sparse regions in a dataset. It clusters data with high accuracy keeping I/O cost low. This means clusters must be more compact and each distinct cluster must be far apart from other clusters. Thus, a cluster is treated as an implicit class.

Clustering partitions huge datasets into groups according to their similarity, hence it is known as data segmentation in some applications. Clustering can be learnt simply by observing things around. Clustering has great potential in the fields of image processing, healthcare, bioinformatics, information retrieval, medicine and crime detection [8, 13]. Partitioning algorithms is the most elemental version of cluster analysis. Every cluster is distinct and heterogeneous in nature. K-means clustering is a popular partitioning algorithm on account of effortlessness and productivity.

Table 1.1: Pre-requirements of partitioning clustering

Pre-requirements of partitioning clustering	**General characteristics**
Number of clusters	How many clusters needed must be already defined by the user previously.
Boundary constraints	It comprises of upper and lower bound. Values must lie within this range.
Objective function	It can be single or multi objective function aimed to be minimized or maximized.

Although K-means is an admired clustering method which is extremely simple and efficient to use, still it has several shortcomings [5]. Meta-heuristic optimization algorithms are applied to obtain optimal solutions for clustering problems which helps to reduce the drawbacks of traditional partition clustering methods.

1.2 Optimization algorithms

Optimization is an important component of machine learning algorithms. It is essential to have a sound knowledge of optimization frameworks. Framework of optimization algorithms consists of three core components namely objective function, collection of variables and a set of constraints. An objective function is a single numerical quantity which can be minimized or maximized. A collection of variables are the quantities which can be manipulated to optimize the objective function. A set of constraints are the restrictions on the values the variable can take. An optimization problem can be formulated through the following procedure:

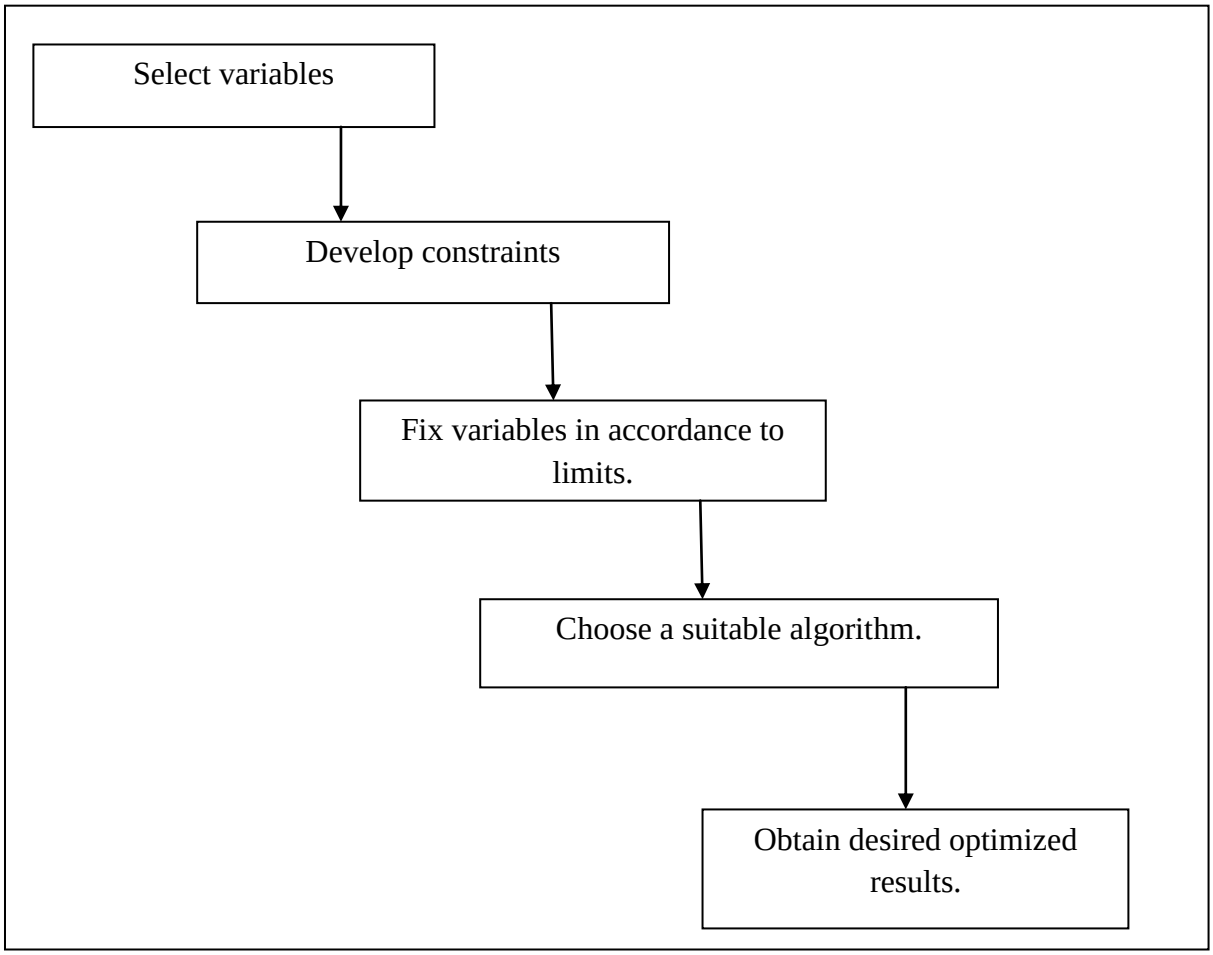

Fig 1.1: Formulation of an optimization problem.

1.2.1 Classification of optimization algorithms

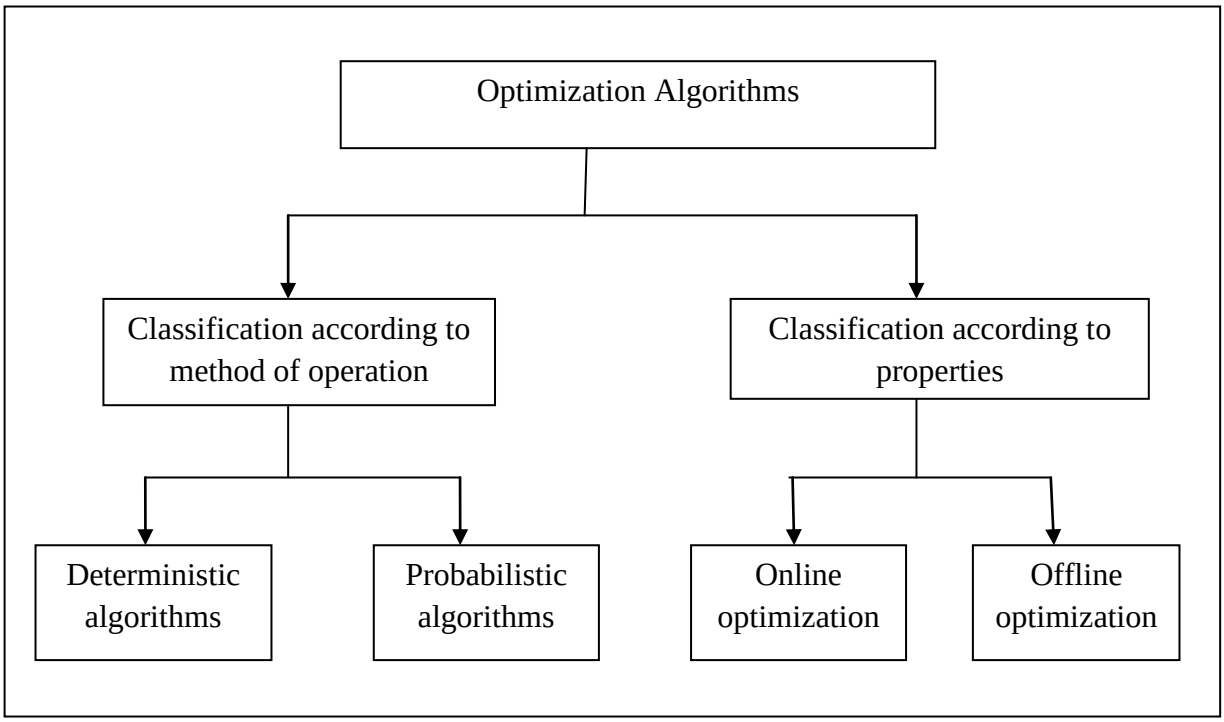

Fig 1.2: Taxonomy of optimization algorithms.

Optimization algorithms are classified in accordance to method of operation and properties. According to mode of operation, they are divided into two fundamental class namely deterministic algorithms and probabilistic algorithms. 'Inputs' are an essential component in the former one where there is no randomness involved. All kinds of meta-heuristic algorithms come under probabilistic algorithms. These algorithms rely on certain random values.

According to properties such as speed, optimization algorithms are further classified as online and offline optimization. Online optimization problems are tasks which should be solved in a quick time span ranging from milliseconds to few minutes. Examples include robot localization, load balancing and service composition for business processes. Offline optimization problems are tasks where time is not an important factor and a user is willing to wait for days to get optimal results.

1.3 Meta-heuristic algorithms

Meta-heuristic optimization algorithms are flexible. They follow a deviation-free mechanism and aim to avoid local optima [16, 17]. Structure of meta-heuristic algorithms can be easily manipulated. They utilize random initial solutions to avoid local optimum. Meta-heuristic algorithms have different solution strategies. It adapts itself according to the problem domain and has less computational power.

Variants of meta-heuristic algorithms are single solution based and population based meta-heuristic algorithms. Only one solution gets processed in case of single solution based algorithm whereas a set of solution is much prevalent in case of the latter one. An optimizer generates random solutions over the problem in exploration phase and it focuses on quality of solutions in exploitation phase. Meta-heuristic algorithm mainly consists of searching steps such as exploration and exploitation phase [2].

1.3.1 Classes of meta-heuristic methods

Table 1.2: Classes of meta-heuristic methods

Paradigm	Methodology	Characteristics	Examples
Class 1	Evolution-based	Inspired from laws of biological evolution	Genetic algorithm, Biogeography based optimizer.
Class 2	Physics-based	It mimics the physical regulations of the universe	Ray Optimization and black hole.
Class 3	Swarm-based	It simulates all kind of animal or human behaviour.	Flower Pollination, Social Spider Optimization, Moth Swarm Algorithm.

1.3.2 Evolution based meta-heuristic methods

Evolution based meta-heuristic methods are able to deal with complex optimization problems due to its simplicity and flexibility. It derives its metaphor from biological evolution.

The fundamental building block of **genetic algorithm** is based on natural selection and natural genetics. It searches from multiple data points, not from a singular point. It utilizes information obtained from objective functions to find out the direction of the search. It employs probabilistic rules so that it can search uncertain areas to obtain global optimum. Tournament selection generates competitive parent strings for better convergence. Simulated binary crossover generates two children from two parents. Next, polynomial mutation occurs on individual strings of offspring.

Biogeography based optimization [6] studies geographical distribution of biological species in order to derive algorithms for optimization. Migration operator allows emigrating habitats share their good features with immigrating habitats. Mutation operator uses mutation rate to select a habitat. Elitism operator aims to maintain quality of population by keeping best habitats for next iteration. It is inspired by theory of Island Biogeography where relocation of species is represented as a mathematical model. Habitats are the desired output of this problem. High habitat suitability index means surroundings having more good species. BBO is applied to diverse application areas like image processing, wireless sensor networks [6].

1.3.3 Physics based meta-heuristic methods

Physics based meta-heuristic methods are based on physical convention of the universe.

Ray optimization [20] works on refraction property of light rays. This method draws its inspiration from transition of ray through which near-optimal solutions are obtained. Snell's law is the essential tool of this algorithm. The number of particles constitutes the 'variables'. First, determination of goal function for a solution vector takes place. The search space is filled with agents distributed all around. We assign best agent as global best and save the current position of each agent as best position. Next, each agent moves to a new position on the basis of movement vector. If an agent violates a boundary, then improve its position.

Black hole algorithm [8] is based on the phenomenon of black hole. Creation of black hole occurs when a huge star collapse. Black hole has high gravitational power through which matter gets squeezed into a tiny space. Boundary of the black hole is called event horizon. Any particle gets absorbed into the black hole if it comes near to the event horizon.

1.3.4 Swarm based meta-heuristic methods

Swarm based meta-heuristic methods are based on animal behaviour. **Flower pollination algorithm [19, 21, 23, 25]** is based on pollination process. It uses a representation of pollination where pollinators are used to spread pollen over the landscape. It takes its metaphor from flowers proliferation role in plants. It is a flexible, adaptable, scalable optimization method [19]. It is a technique which is initiated with random solutions. It consists of two operators such as local pollination operator and global pollination operator. Similarity in solution vectors is represented through flower constancy. The switch operator exchanges the improvement loop locally or globally.

Artificial bee colony [5] consists of employed bees, onlooker bees and scouts. They aim to optimize food search around their hive via communication. The employed bees perform following functions: search for location of food sources and carry forward this information to onlooker bees. The onlooker bees receive such data and exploits food sources. The scout bees explore food sources in different dimensions of the search space.

Particle swarm optimization [15] is extremely popular meta-heuristic algorithm. PSO follows one-way information sharing mechanism. It consists of two parameters namely P-Best and G-Best. P-Best refers to the personal best position, whereas G-Best refers to the best position in a swarm. Each particle keeps information about the best position it has gained so far. In each step, each particle is moved towards the best particle with a changed velocity and added randomness. It is guided by notion of fitness. Its search strategy is based on velocity and position updating.

Social spider optimization [22]is based on cooperation among social spiders. Every spider is assigned a weight irrespective of its gender. Vibration is the medium through which spiders communicate with each other. Every spider feels three vibrations from other spiders. These include vibration from nearest spider having a higher fitness, best spider in the swarm and nearest female spider. Dominant male spiders show better fitness than non-dominant ones. Mating operator is the last step of this algorithm. Dominant male spiders mate with female

ones within a mating radius. Fitness value of new spider produced are calculated and compared with worst population. Spider having better fitness value replaces less fit spider.

Moth swarm algorithm [24]is a swarm based algorithm inspired from behaviour of moths. Here, the position of light source is the most feasible solution of the optimization algorithm. Pathfinder moths, prospector moths and on-lookers moths are an essential component in this algorithm. Pathfinder moths search for food. Prospector moths use spiral method to intensify the search. Moths with low fitness values are known as on-lookers moth. They must probe effectively around the hotspots of the prospectors. The aim of moths is to drift towards the moonlight.

CHAPTER 2

LITERATURE REVIEW

2.1 Literature work

This section presents a brief summary on the basis of literature.

Baalamurugan and Bhanu (2018) have introduced Efficient Stud Krill Herd (ESKH-C) algorithm for solving clustering problem. The proposed algorithm uses stud selection and crossover operator. This operator makes the solutions more refined. It chooses solution of better quality for each krill. It is an optimisation approach which aims to minimize the fitness functions. One fitness function is the lowest value of distance between krill and source of the food. Another fitness function is the smallest value of distance between krill and largest concentration of herd. The performance of algorithm is tested on two synthetic databases and five real datasets. Simulation results are evaluated using various validity measures such as Jaccard, Rand, Beta and Distance index and compared with well-known algorithms. Authors claim that the proposed algorithm provides good results for the datasets.

Pal (2017) have proposed a novel meta-heuristic clustering method called BBOKM. It uses exploitation and exploration capabilities of BBO and K-means for data clustering. The novel method initializes the population by K-Means algorithm. Intra-cluster distance is chosen as a performance measuring criteria. The proposed algorithm has been tested on eleven datasets namely Iris, Wine Glass, CMC, Cancer, Heart, Lung Cancer and Vertebral. It has been compared with three evolutionary algorithms. Simulation results showed that BBOKM algorithm work well for many datasets. Exception lies in case of cancer dataset where differential evolution algorithm showed better results.

Hatamlou (2013) have proposed clustering by black hole optimization. Six well-known datasets are used to evaluate the performance of proposed algorithm. Intra-cluster distance and error rate are the validity measures used for evaluation. The author claimed that black hole (BH) algorithm performs better than other algorithms. It results into great quality solutions and value of standard deviation is small.

Lukasik (2017) have proposed data clustering with grasshopper optimization algorithms. GOA technique implements two components of grasshoppers movement strategies. First component is the interaction of grasshoppers while in larvae form and in insect form. Second

component is the tendency to move towards the source of food. For the experiments, a set of standard real and synthetic clustering benchmark datasets are used. Results showed that clustering based on this optimization technique provided better and high accurate results as compared to standard k-means. GOA-based clustering outperforms than K-means on majority of the datasets.

Marinakis (2008) have proposed to combine genetic algorithm and GRASP technique and apply it in clustering problems. Genetic algorithm is used for feature selection which reduces redundant features. Later GRASP (greedy randomized adaptive search procedure) is applied for clustering problem. In the first phase, number of features are activated using genetic algorithm. In order to implement selection mechanism, roulette wheel selection is used. An individual having greater fitness will have a lager sector and lower fitness will have small sector. 1-point crossover is used in crossover phase of the algorithm. Parents get separated and offspring takes one part each from both the parents. Later, fitness function for the offspring is calculated. Its performance is tested on nine benchmark datasets. Comparison is drawn with Tabu search algorithm. Results show that the proposed technique provides excellent results in terms of high accuracy.

Tang (2012) have proposed multiple bio-inspired algorithms to be merged with K-Means. In every algorithm, exploration phase for global optimum is different. In C- wolf, exploration phase is enabled by random escapes. C-Firefly, C-Cuckoo and C-bat algorithm enable exploration through levy flight. The results of proposed methods are compared with K-Means. The proposed algorithm is compared with six real-time datasets. Results showed that each one of the new algorithms is able to achieve good and even partitions.

Shanthi (2018) have proposed clustering based on crow search algorithm. It overcomes K-means local optimum problem. Fitness function of CSAK-Means algorithm is Mean Square Error Criterion. The performance of proposed CSAK means method is evaluated on six benchmark datasets. Its performance is further evaluated on internal and external measures like Silhouette, purity, rand index and F-measure. It is compared with other well-known algorithms. The author claimed that CSAK algorithm outperforms than other algorithm.

Zhou (2017) have proposed social-spider optimization algorithm based on simplex method for clustering. Simplex method helps to increase variation of the population. Fitness function is calculated and weight to a spider is assigned. We calculate three vibrations from each spider. We update location of male and female spiders and further calculate mating radius. If

the spiders are within mating radius range, then new spiders are created. Otherwise we update location of worst spider by simplex method. We obtain spider having best value of fitness. The proposed algorithm is further tested with two artificial datasets and nine real-time datasets. The experimental results have been compared to six state-of-art algorithms. Author claimed that the proposed algorithm performs better than the other algorithms.

Yang (2017) have proposed moth swarm algorithm for clustering. For path-finders moth, we sort according to the fitness value. For every prospector moth, we create new fitness values and new light sources. For every onlooker moth, we produce Gaussian walks. In the end, these 3 kinds of moths are the required clusters. Performance of the proposed algorithm is tested with one artificial and three real datasets. Author claimed that the proposed algorithm has a high efficiency and helps to solve complex optimization algorithms.

Senthilnath (2019) have proposed clustering in flower pollination algorithm. This approach extracts useful information in terms of optimal cluster centres. Cross-pollination obtains global solutions. Self–pollination aims to find local data solutions. Three standard UCI datasets and multispectral crop type dataset are used to validate its robustness. The performance of proposed algorithm is compared with multiple known algorithms. The algorithm is evaluated on certain validity measures such as classification error percentage, time complexity and statistical significance. The author claimed that the proposed algorithm has lowest error value and provides great convergence.

Hatamlou (2012) have proposed gravitational search algorithm to be integrated with K-Means for good clustering. The proposed algorithm constitutes three steps. First, it applies k-means algorithm on desired dataset to obtain optimized cluster centres. An initial population of solutions is obtained in the second step which contains candidate solutions using minimum, mean and maximum of the dataset. GSA is employed for determining optimal solutions. Five real datasets are tested on the proposed algorithm. Its performance is compared with other defined algorithms.

Boushaki (2018)proposed clustering through chaotic cuckoo search algorithm. It is inspired by quantum theory. A chaotic map is used to initialize population. The property of non-repetition accelerates the search by exploring the search space in an efficient manner. It is ensured that cuckoos remain inside the available search space. Cuckoos should be bounded within the search space available. They should not move outside the search space. The performance of the proposed algorithm is tested on six real datasets. It is further compared

with eight well known algorithms. The author claimed that the new algorithm performs superior than the others.

Mageshkumar (2018)have proposed a fusion of ant colony optimization and ant lion optimization for clustering. ACO algorithm is used to generate initial random solutions for the candidates. In the inter-mediate stage, ALO algorithm is used where best ant lion is selected as elite. Iterated local search algorithm is used in the final stage for improving the quality of clusters obtained.

2.2 A Comprehensive review table

Table 2.1 Summary of literature work

References	**Source**	**Year**	**Methodology**	**Performance evaluation**
Ref [3]	Springer	2018	Proposed Efficient Stud Krill Herd Clustering algorithm to solve clustering problem.	Jaccard, Rand, Beta and Distance index
Ref [4]	Springer	2011	Proposed harmony search optimization algorithm to solve initialization issue of clustering algorithms.	Seven real datasets and two artificial datasets.
Ref [5]	Elsevier	2010	Proposed artificial bee colony algorithm for clustering.	Compared with ACO,GA,SA,TS and KNM-PSO.
Ref [6]	IEEE Transactions	2017	Introduced a novel hybrid meta-heuristic technique called BBOKM.	Sum of intra-cluster distance, Friedman test and Holm test.

Ref [7]	Pertanika Journals	2016	Proposed differential search clustering method.	Precision, recall and G-measure.
Ref [8]	Elsevier	2013	Proposed black hole optimization approach for data clustering.	Sum of intra-cluster distance and error rate.
Ref [9]	IEEE Transactions	2017	Proposed data clustering with grasshopper optimization algorithms.	Mean and standard deviation values of Rand Index
Ref [10]	Elsevier	2018	Proposed novel SOS (symbiotic organism search) algorithm for clustering.	Mean and standard deviation values.
Ref [11]	Springer	2019	Proposed coral reef optimization with substrate layers (CRO-SLC) for data clustering.	Sum of squared error (SSE).
Ref [12]	Springer	2008	Proposed hybrid stochastic genetic GRASP algorithm for data clustering.	Compared with Tabu Search algorithm.

Ref [13]	Springer	2020	Proposed the GWOTS algorithm for clustering.	SSE, purity and entropy.
Ref [14]	IEEE Transactions	2012	Proposed integration of bio-inspired optimization algorithms into k-means clustering.	Objective function value and CPU time.
Ref [15]	Elsevier	2007	Proposed combinatorial particle swarm based optimization technique for clustering approach.	Variance ratio criterion (VRC) and squared error (SE).
Ref [16]	Springer	2018	Proposed clustering algorithm for crow - search algorithm	Silhouette, Purity, Normalized Mutual Information, Rand Index and F-measure.
Ref [17]	Elsevier	2019	Proposed pathfinder algorithm as meta-heuristic optimizer	Tested on unimodal, multimodal and composite functions.

Ref [18]	Elsevier	2019	Proposed collective decision optimization algorithm for training artificial neural networks.	Mean error value and standard deviation to evaluate search capability.
Ref [22]	Elsevier	2017	Proposed social spider optimization algorithm for clustering analysis.	Purity
Ref [24]	Springer	2017	Proposed moth swarm algorithm for clustering analysis.	Comparison is done with other algorithms such as GWO, FPA, CS, ABC and K-Means.
Ref [25]	Springer	2019	Proposed flower pollination algorithm as a standalone approach for data clustering.	Classification error percentage, time complexity and statistical significance.
Ref [26]	Elsevier	2012	Proposed a hybrid data clustering algorithm based on k-means and gravitational search algorithm (GSA-KM).	Sum of intra-cluster distances, number of fitness function evaluations.

Ref [27]	Elsevier	2018	Proposed quantum chaotic cuckoo search algorithm for data clustering.	Sum of intra cluster, error rate and Student's t - test.
Ref [28]	Springer	2018	Proposed Ant lion hybrid meta-heuristic algorithm for data clustering.	Intra-cluster distance, distance index, beta index and random coefficient.

CHAPTER 3

PROPOSED STATEMENT

3.1 Problem description

Data mining is a method of extracting useful information from vast, incomplete and unprocessed data [1]. It uses qualitative and quantitative techniques to discover concealed relationships among data items. It is an inter-disciplinary sub field of computer science [1]. Classification and clustering are an important method in data mining. These methods are widely used to determine unseen patterns in data mining. Classification is a two-step process consisting of learning step and classification step. In the learning step, training data are analysed by classification algorithm and further classification rules are being made. In the classification step, test data are used to estimate the accuracy of the classification rules. If accuracy is acceptable, rules can be applied to reach a particular decision.

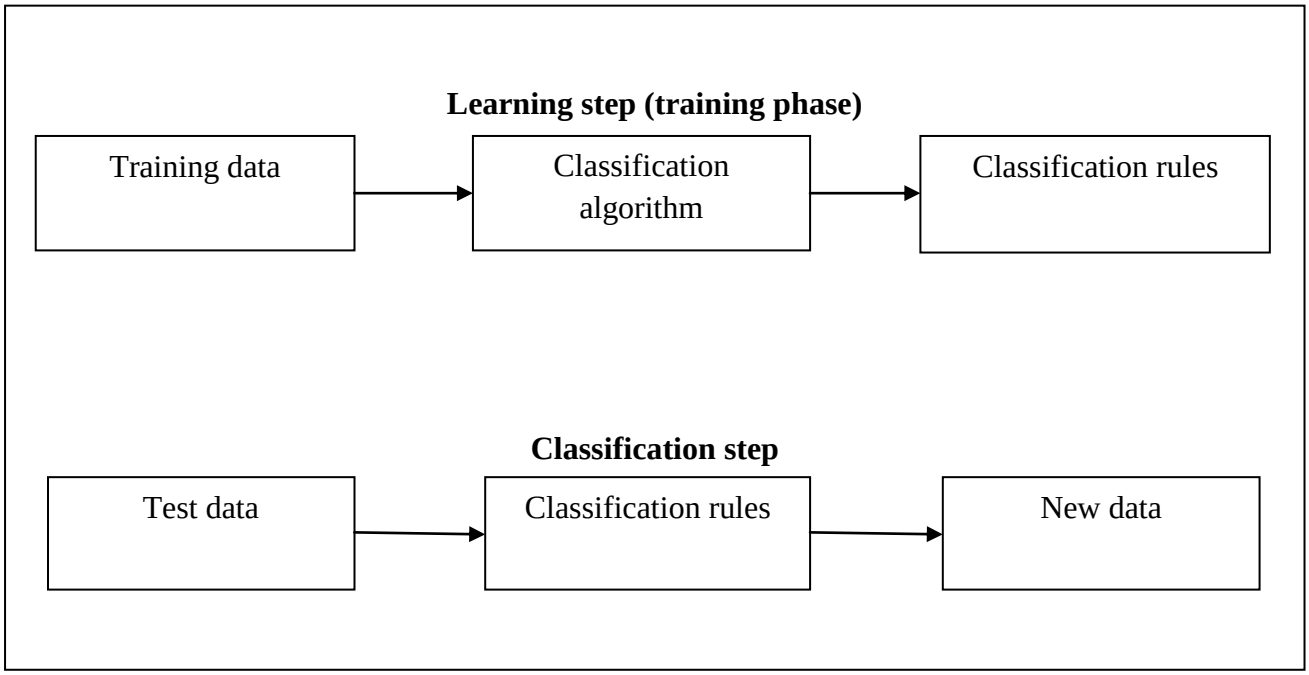

Fig 3.1: Phases of classification

Clustering can discover and identify patterns and trends without any supervision or previously known information [26]. Clustering algorithms are divided into two groups: hierarchical and partitioning. Partition clustering algorithms find all clusters simultaneously

without forming a hierarchical structure. Hierarchical clustering algorithms find clusters recursively in top-down or bottom-up approach.

K-means is the oldest and most popular partitioning method. It is a faster clustering technique. It works well in large datasets. But still, it has various shortcomings as:

a) K-means clustering does not work well when clustering dataset contains noise or clutter.
b) It tends to get stuck in local optima.
c) There is no proper description on how to assume number of clusters.
d) Accuracy of obtained clusters is not so great.
e) Several problems may arise due to bad initial centres.

3.2 Proposed solution

Due to the aforementioned drawbacks, I propose a new framework of clustering the data through a meta-heuristic technique called harris-hawk optimization technique. Meta-heuristic algorithms are more popular than classical algorithms due to the following reasons:

a) Have diverse solution strategies.
b) It can easily adapt itself in accordance to problem domain.
c) It can search solution space with different initial points.
d) Much better accuracy is obtained.

The searching steps have two phases: exploration and exploitation. In the exploration phase, the algorithm should deeply explore various regions. Thereafter, exploitation stage is performed after exploration phase. It intensifies the searching process in a local region. The new framework will lead to optimized cluster centres.

3.3 Proposed algorithm

In this section, HHO combined with K-Means algorithm is proposed. The proposed algorithm is as follows:

Step 1: Input the number of clusters K.

Step 2: Randomly access K clusters from the dataset. Matrix C will be formed of 3*4 size.

Step 3: Find the Euclidean distance between K-cluster centroid and the data objects using the formula:

$$D = \sqrt{\sum(x(i) - c(j))^2} \quad (1)$$

Step 4: Find the minimum distance and assign data objects to clusters.

Step 5: Find out the best accuracy among the three clusters and assign it as the best location of the rabbit.

Step 6: Calculate escaping energy and jump strength of the rabbit:

$$E_0 = 2rand\ () - 1 \quad (2)$$

$$J = 2(1 - rand()) \quad (3)$$

$$E = 2 * E_0 \left(1 - \frac{t}{T}\right) \quad (4)$$

Step 7:if(|E|>=1) then // exploration phase

update rabbit location using the formula:

$$X(t+1) = \begin{cases} X_{rand}\ (t) - r_1 |X_{rand}\ (t) - 2r_2 X(t)|\ , & q \geq 0.5 \\ \left(X_{rabbit}\ (t) - X_m\ \&(t)\right) - r_3\ \left(LB + r_4(UB - LB)\right), & q < 0.5 \end{cases} \quad (5)$$

if (|E|<1) then // exploitation phase

if (r>=0.5 and |E| >=0.5) then // soft beseige

update location using formula:

$$X(t+1) = X_{rabbit}\ (t) - X(t) - E|J * X_{rabbit}\ (t) - X(t)| \quad (6)$$

else if (r>=0.5 and |E|<0.5) //hard besiege

update location using formula:

$$X(t+1) = X_{rab\,bit}\,(t) - E|X_{rabbit}\,(t) - X(t)| \quad (7)$$

else if (r<0.5 and |E|>=0.5) then // soft besiege with progressive dives

update location using formula:

$$X(t+1) = \begin{cases} Y, & if\ F(Y) < F(X(t)) \\ Z, & if\ F(Z) < F(X(t)) \end{cases} \quad (8)$$

$$Y = X_{rabbit}\,(t) - E|J * X_{rabbit}\,(t) - X(t)| \quad (8.1)$$

$$Z = Y + S * LF(D) \quad (8.2)$$

$$LF = 0.01 * \frac{u*\sigma}{|v|_{\beta}^{\frac{1}{\beta}}} \quad (8.3)$$

$$\sigma = \sqrt{\frac{\gamma(1+\beta)*sin\frac{\pi\beta}{2}}{\gamma\left(\frac{1+\beta}{2}\right)*\beta*2^{\left(\frac{\beta-1}{2}\right)}}} \quad (8.4)$$

else if (r<0.5 and |E|<0.5) then //hard besiege with progressive dives

update location using formula:

$$X(t+1) = \begin{cases} Y, & if\ F(Y) < F(X(t)) \\ Z, & if\ F(Z) < F(X(t)) \end{cases} \quad (9)$$

$$Y = X_{rabbit}\,(t) - E|J * X_{rabbit}\,(t) - X_m(t)| \quad (9.1)$$

$$Z = Y + S * LF(D) \quad (9.2)$$

Step 8: Set the boundary constraints of new locations.

Step 9: Thus,we get new rabbit locations as optimized cluster centres.

3.4 Methodology

Harris-hawk optimization algorithm is a population based optimization technique. It consists of two phases namely diversification and intensification.

3.4.1 Exploration point

Exploration point proposes that the hawks wait, observe and monitor the site to detect the rabbits (prey). ‘q’ is considered to be an equal chance of perching. We can update rabbit locations using equation (5).

When escaping energy |E|>=1, exploration phase is performed and when |E|<1, exploitation phase is performed.

3.4.2 Exploitation point

The hawks perform the surprise pounce (seven kills) by attacking the rabbits. Four strategies are anticipated to form the attacking stage:

a) Soft encircle

Soft encircle means that the rabbit tries to escape by random misleading jumps but it cannot. The hawks encircle it softly to make the rabbit more exhausted and then perform surprise pounce. For surprise pounce to happen, the escaping energy ‘E’ should be greater than or equal to 0.5 and chance of successful escape of a rabbit ‘r’ should be greater than or equal to 0.5. Soft encircle is performed using equation (6).

b) Hard encircle

Hard encircle means that the rabbit is extremely exhausted and has low escaping energy. The hawks do not encircle the rabbit to perform the surprise pounce. For hard encircle to perform, the escaping energy ‘E’ should be less than 0.5 and ‘r’ should be greater than or equal to 0.5. Hard encircle is performed using equation (7).

c) Soft encircle with progressive dives

For soft encircle with progressive dives to occur, ‘E’ should be greater than or equal to 0.5 and ‘r’ should be less than 0.5. For progressive dives, concept of levy flight is used. The levy flight is used to mimic the real zig-zag deceptive motion of rabbits. Soft encircle with progressive dives is performed using equation (8).

d) Hard encircle with progressive dives

For hard encircle with progressive dives to occur, ‘E’ should be less than 0.5 and ‘r’ should be less than 0.5. Hard encircle with progressive dives is performed using equation (9).

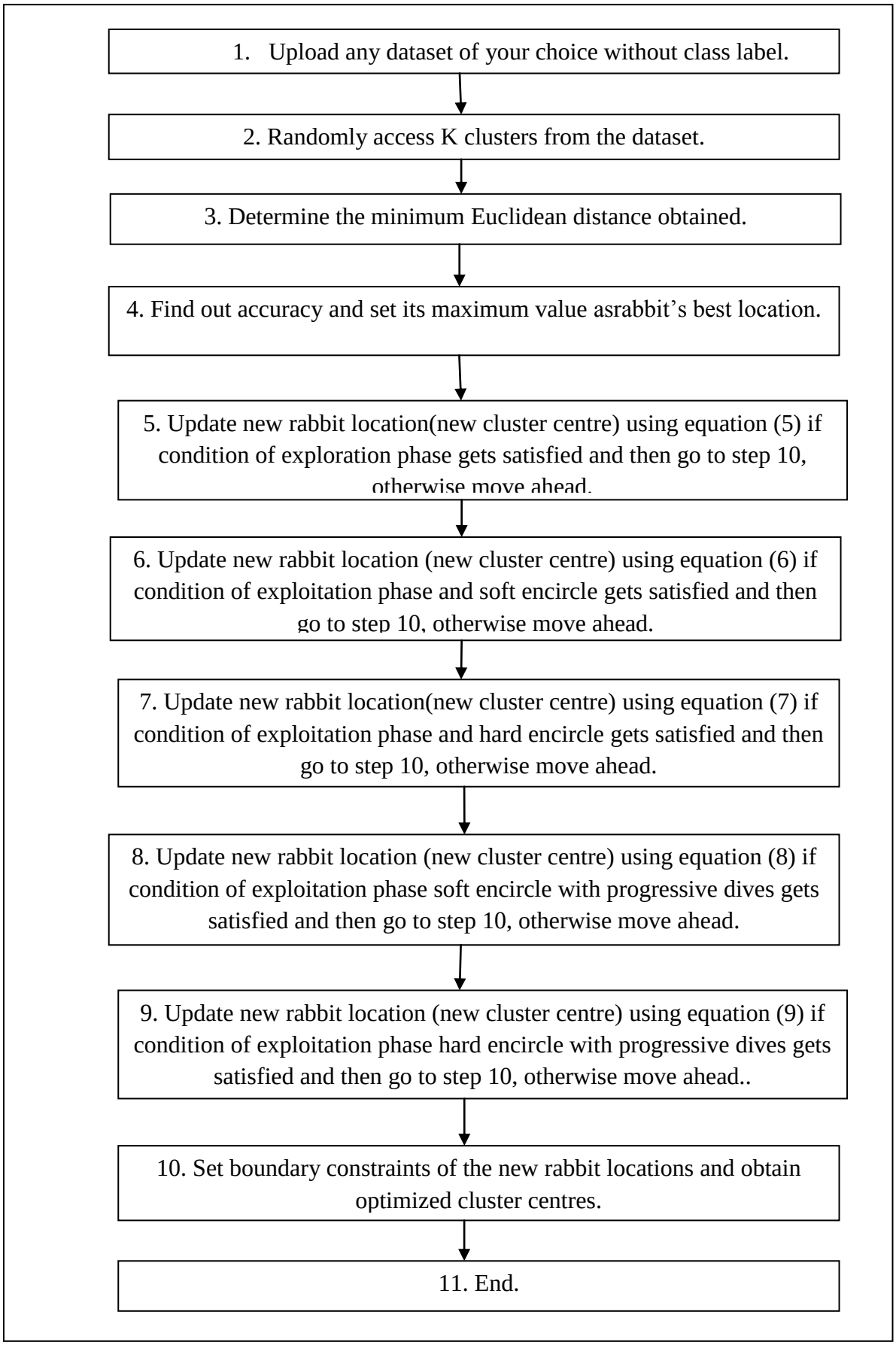

Fig 3.2: Flowchart of the proposed algorithm.

CHAPTER 4

IMPLEMENTATION AND EXPERIMENT

4.1 Software requirement

The experiment is performed on MATLAB 2018 using 8GB RAM, Windows 10 and core i3 processor. The proposed algorithm tends to optimize cluster centres by minimizing the objective function. Accuracy and intra-cluster distance for the proposed algorithm is calculated. To test the performance of this algorithm, we use five real datasets from UCI machine learning.

4.2 Experimental results

4.2.1 Dataset used

Five real-time datasets are used in this work such as Iris, Wine, Glass, Zoo and Contraceptive Method Choice (CMC). They are downloaded from UCI Repository. Following are the description of the datasets:

Table 4.1: Dataset description

Dataset	**Number of clusters**	**Number of rows**	**Number of columns**
Iris	3	150	4
Wine	3	178	13
Zoo	7	101	16
Glass	6	214	9
CMC	3	1473	9

4.2.2 Intra-cluster distance

Intra-cluster distance is one of the performance parameters of the clustering algorithm. It can be defined as sum of distances between instances within a cluster to the centre points of cluster. If sum of intra-cluster distance is low, it is considered to be a good quality cluster.

Table 4.2: Comparison between performance of various datasets through K-means and proposed algorithm on the basis of intra-cluster distance

Datasets	**K-Means**	**Proposed method (HHO)**
Iris	6.1122	**4.9088**
Wine	666.9707	**642.4258**
Zoo	4.0655	**3.8676**
Glass	63.3292	**11.5261**
CMC	27.5624	**22.8542**

It is observed that value of intra-cluster distance is minimum for all datasets through proposed algorithm. Zoo dataset has minimum intra-cluster distance, followed by iris dataset. The maximum value of intra-cluster distance is observed in wine dataset. Thus, high quality clusters are observed through use of proposed algorithm.

4.2.3 Accuracy

Accuracy is the second parameter of the clustering algorithm. It is defined as number of test tuples that are correctly classified by the classifier.

$$A = \frac{TP+TN}{P+N} \qquad (1)$$

Accuracy of proposed algorithm is tested through five real-time datasets. This is also compared with accuracy obtained through K-Means method.

Table 4.3: Comparison between performance of various datasets through K-Means and proposed method on the basis of accuracy

Datasets	K-Means	Proposed method (HHO)
Iris	30.6667	**100**
Wine	11.8056	**23.94**
Zoo	20.6446	**50**
Glass	16.667	**35.2941**
CMC	29.7456	**67.9641**

It is clearly seen that accuracy of iris dataset through proposed algorithm is maximum followed by CMC Dataset. Wine dataset has least accuracy. Overall, accuracy values obtained through proposed method is much more useful.

The following figures show the clusters of different datasets through the proposed algorithm.

Fig 4.1: 3-D view of Iris dataset clusters

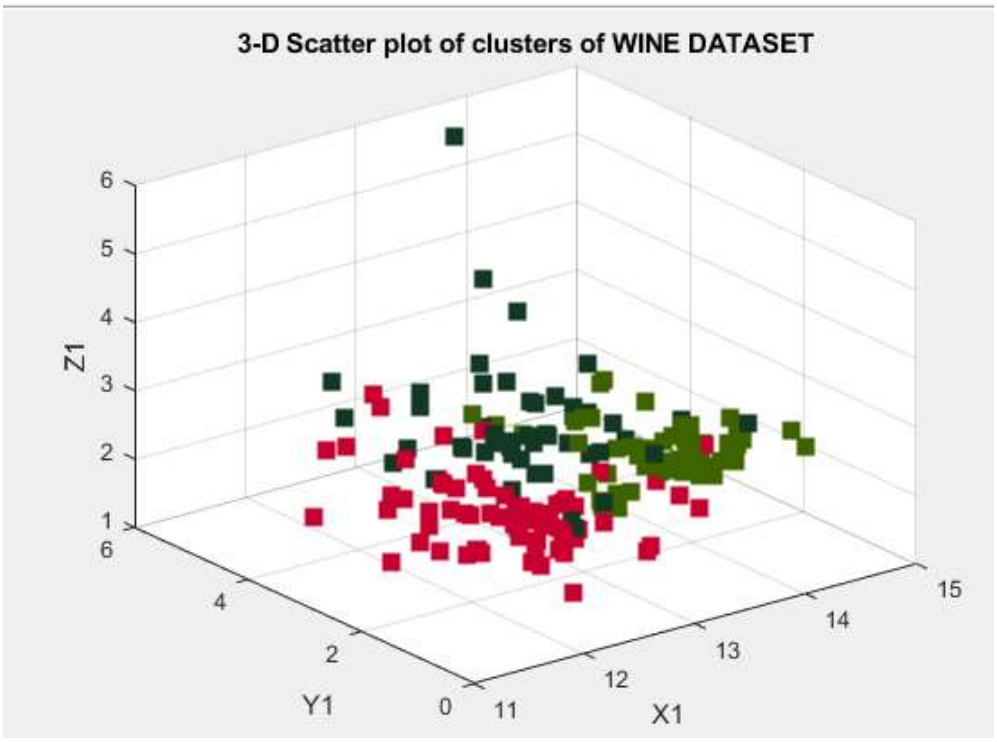

Fig 4.2: 3-Dview of Wine dataset clusters

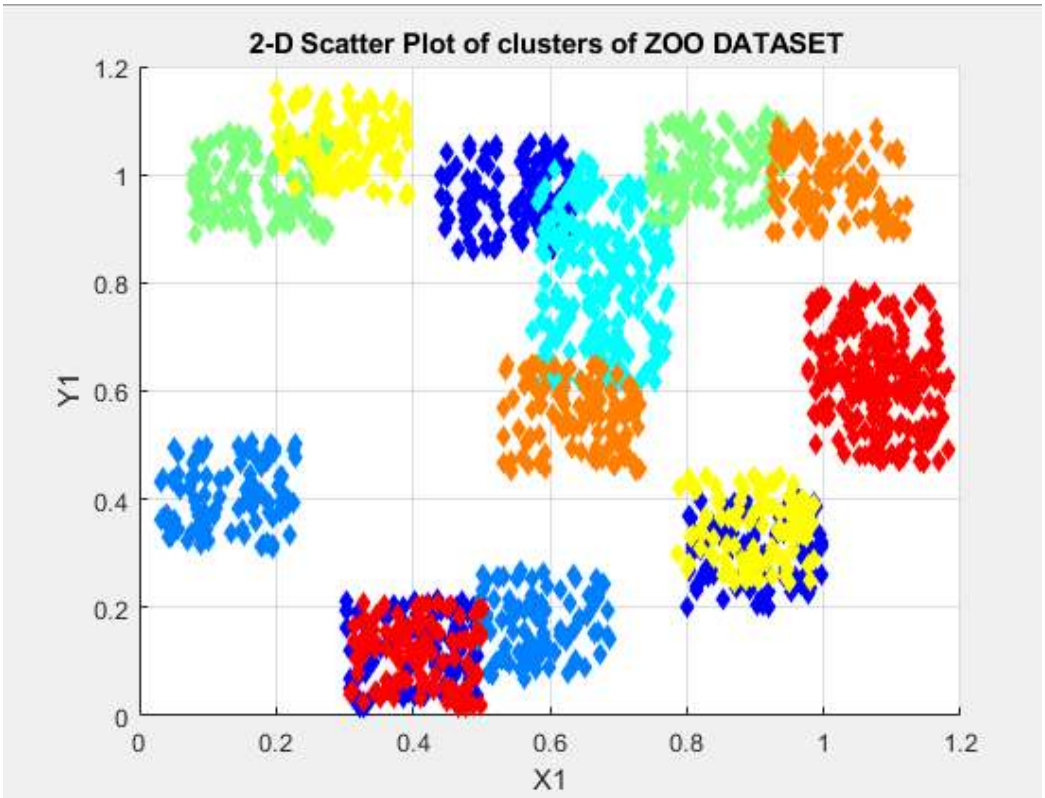

Fig 4.3: 2-D view of Zoo dataset clusters

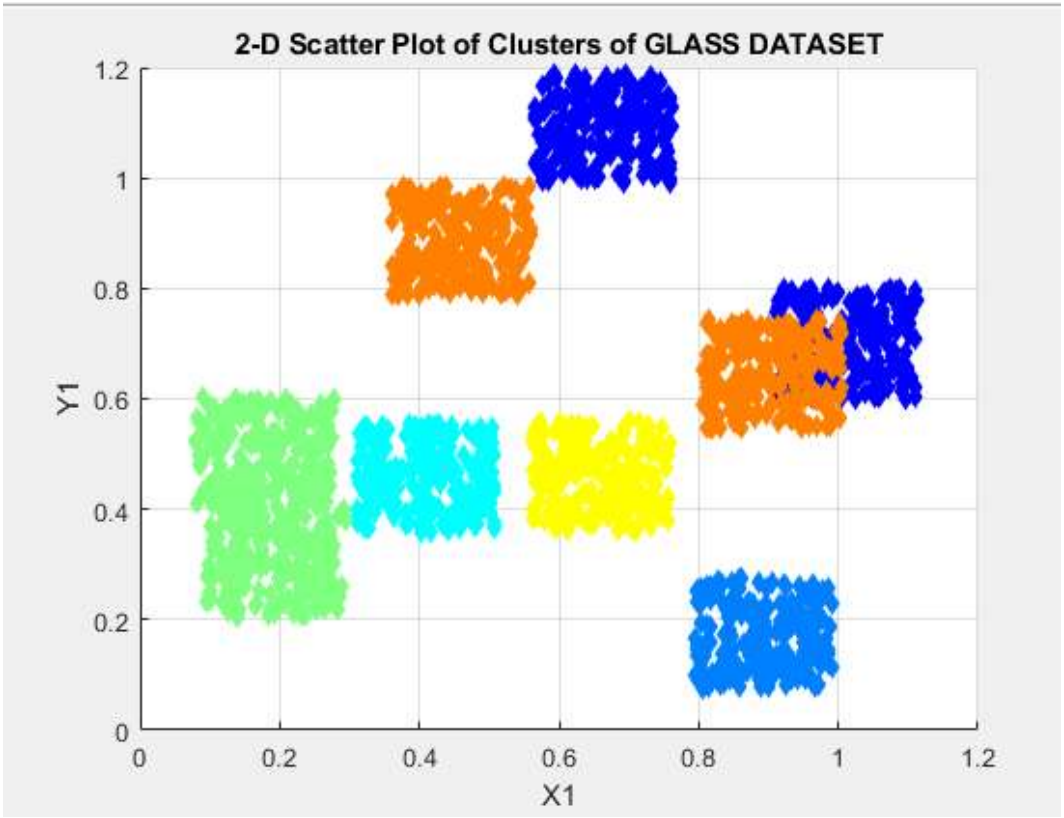

Fig 4.4:2-D view of Glass dataset clusters

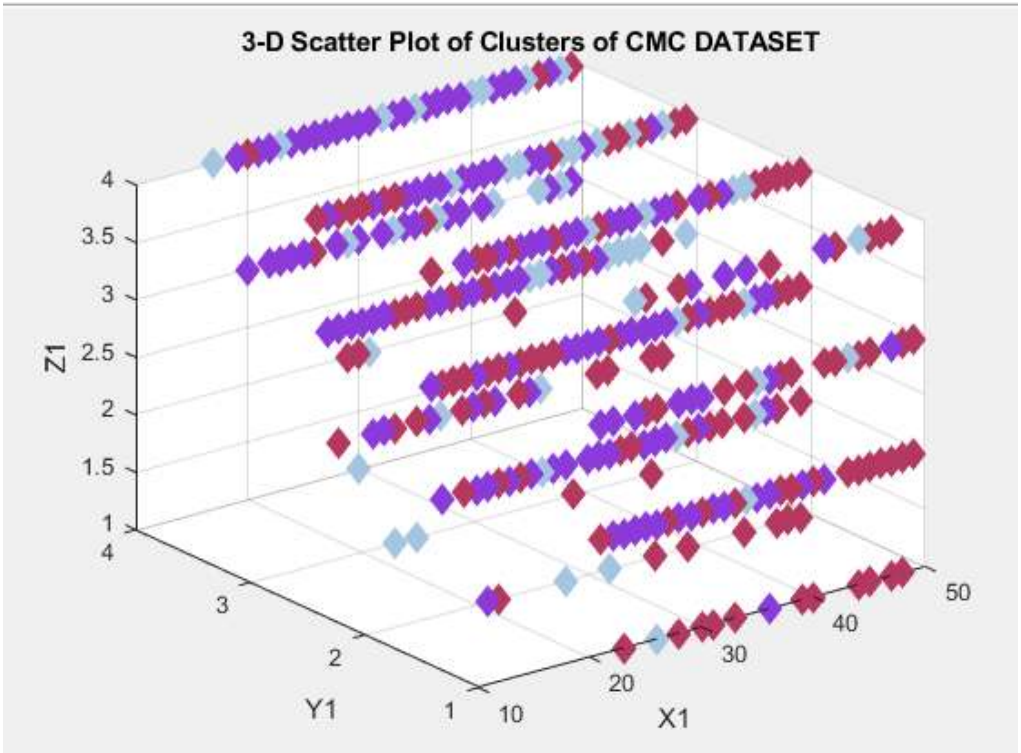

Fig 4.5:3-D view of CMC dataset clusters

4.3 Comparison with other algorithms

The following figures draw a comparison between centroids obtained by different datasets using proposed algorithm and K-Means algorithm. Clearly, centroids obtained by proposed algorithm is much better and optimized as compared to centroids obtained by K-means.

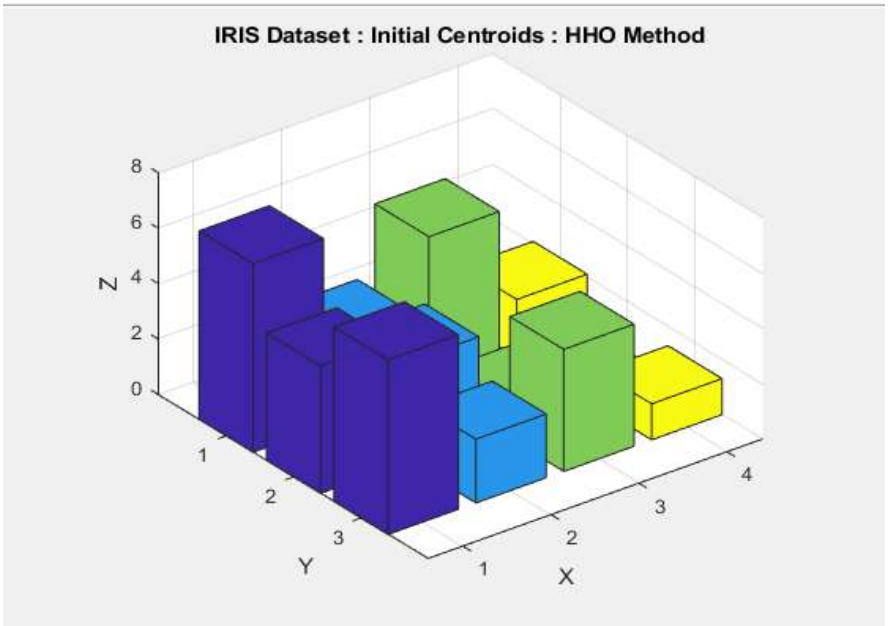

Fig 4.6: Bar graph showing initial centroids of Iris Dataset using proposed method

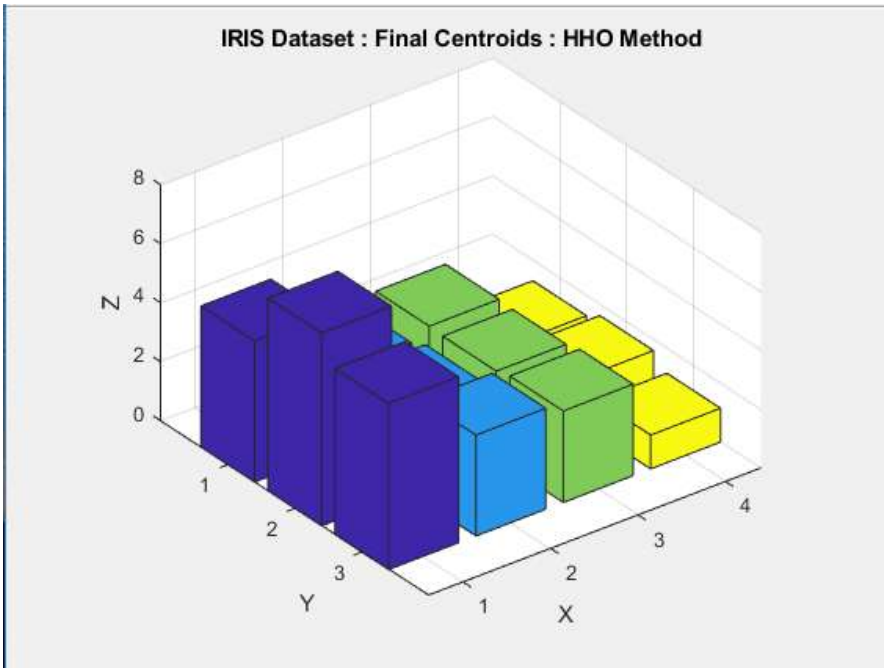

Fig 4.7: Bar graph showing final centroids of Iris Dataset using proposed method

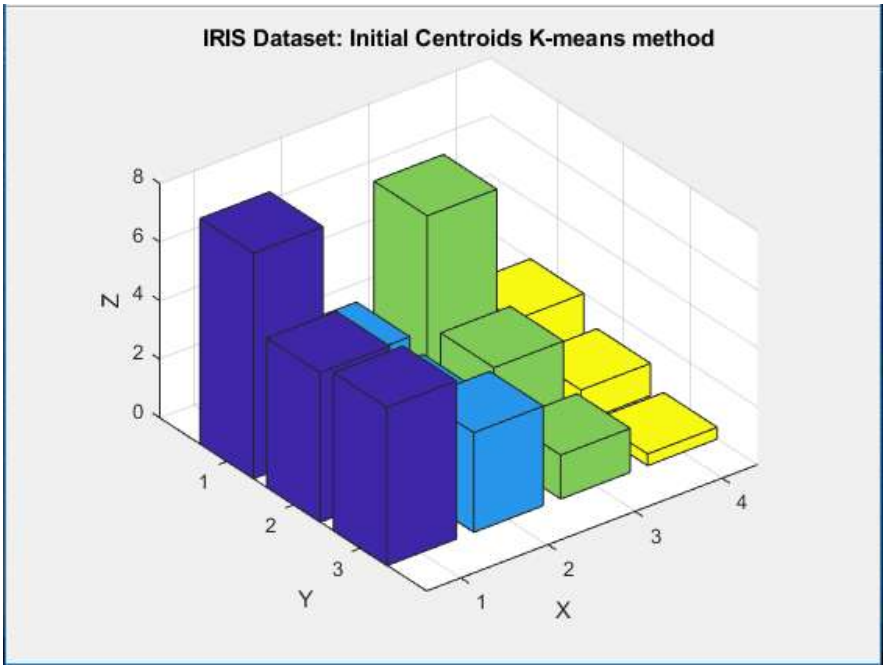

Fig 4.8: Bar graph showing initial centroids of Iris Dataset using K-Means method

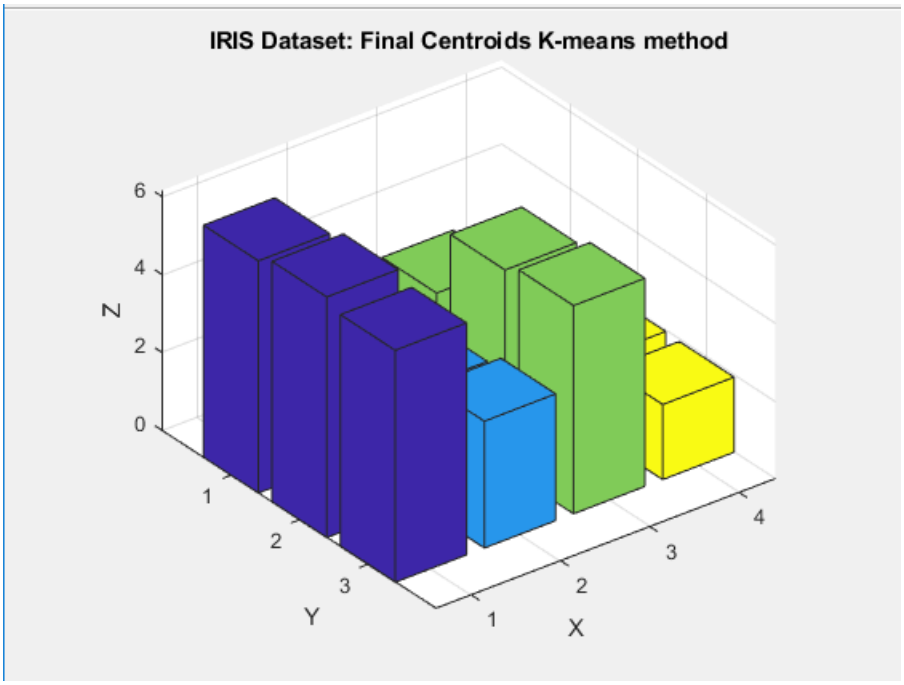

Fig 4.9: Bar graph showing final centroids of Iris Dataset using K-Means method

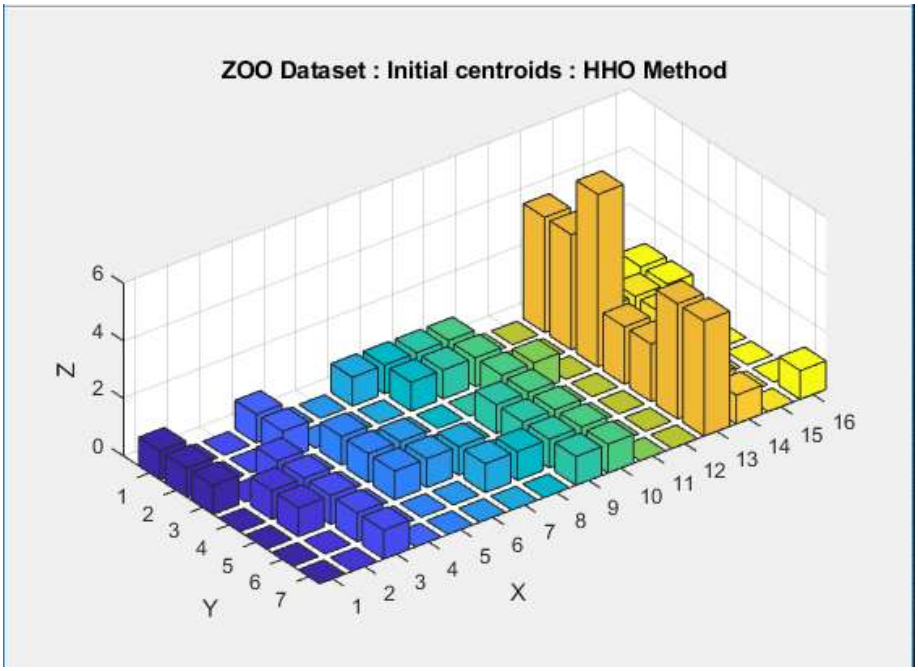

Fig 4.10: Bar graph showing initial centroids of Zoo Dataset using proposed method

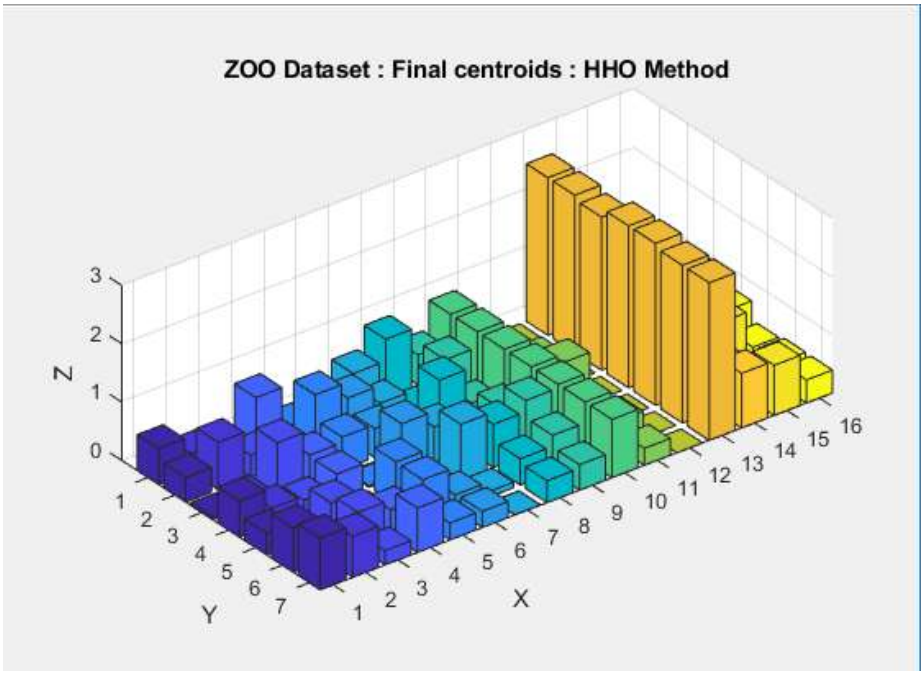

Fig 4.11: Bar graph showing final centroids of Zoo Dataset using proposed method

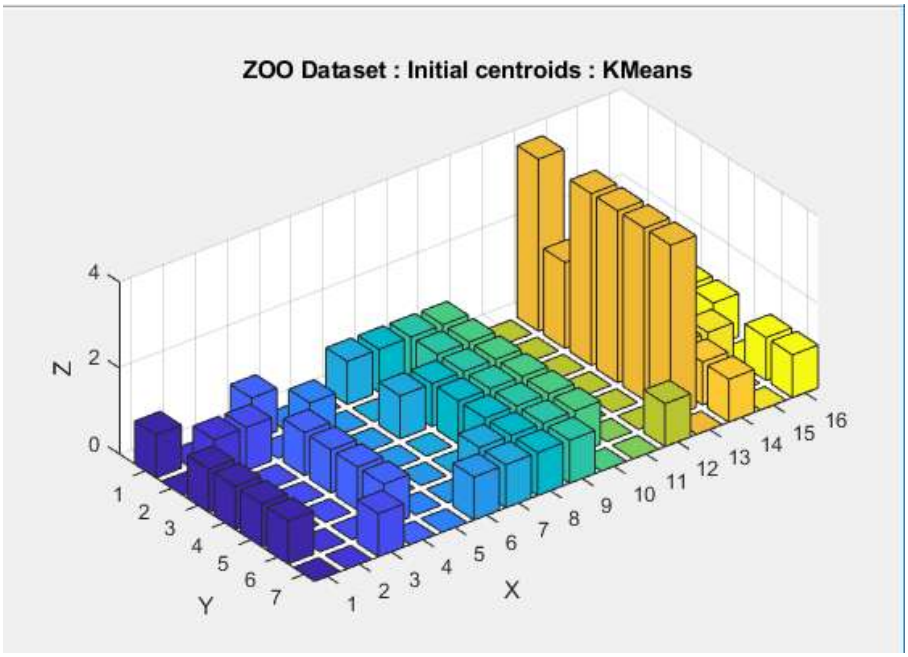

Fig 4.12: Bar Graph showing initial centroids of Zoo Dataset using K-Means method

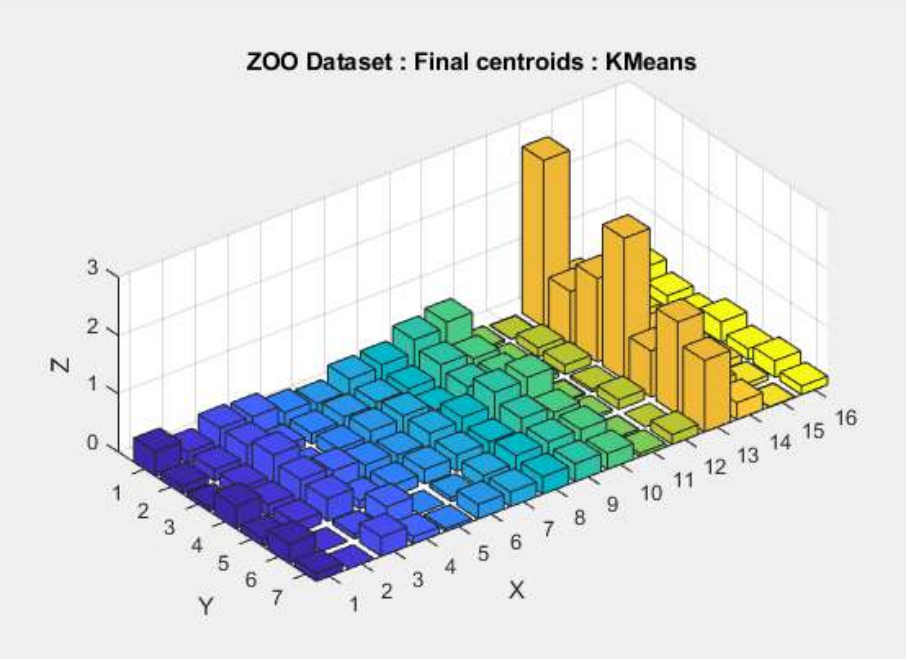

Fig 4.13: Bar graph showing final centroids of Zoo Dataset using K-Means method

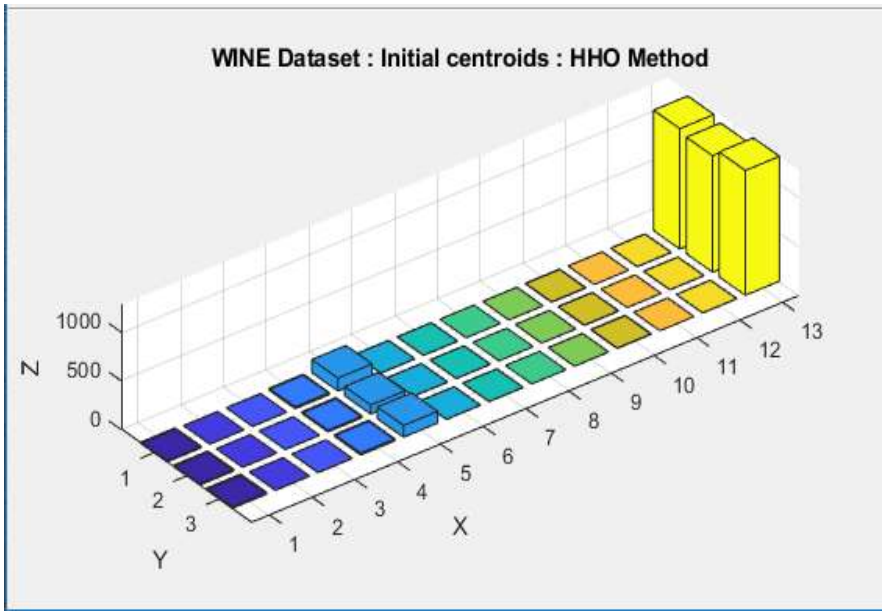

Fig 4.14: Bar graph showing initial centroids of Wine dataset using proposed method

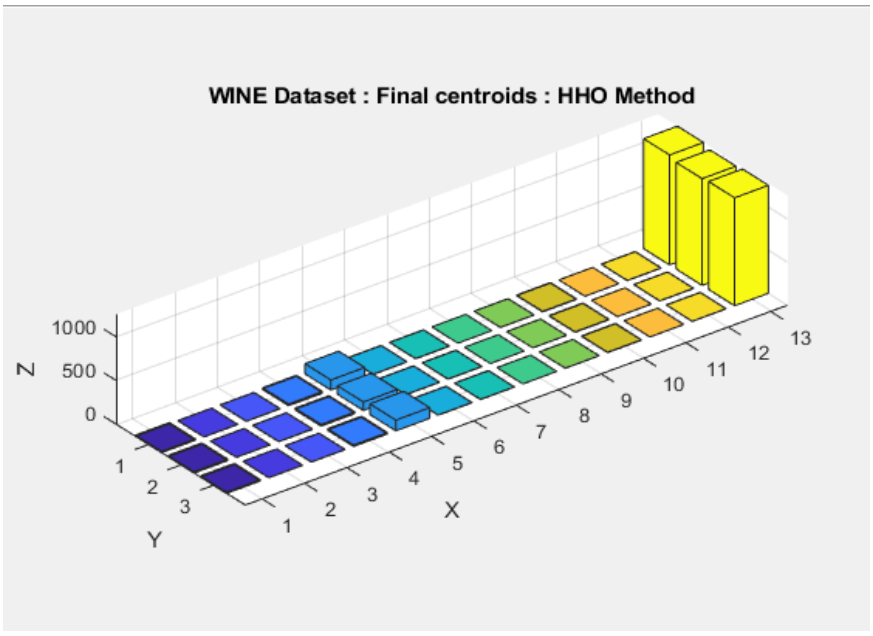

Fig 4.15: Bar graph showing final centroid of Wine dataset using proposed method

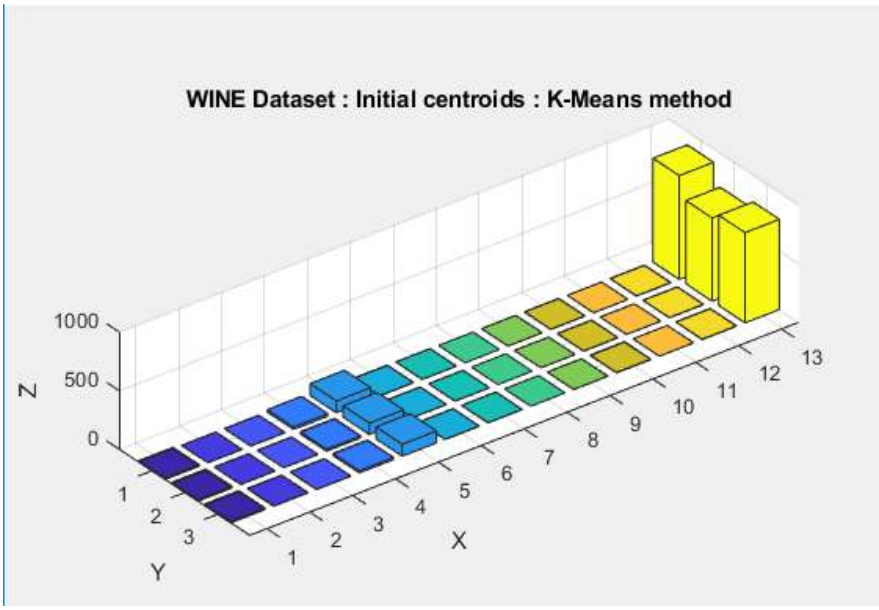

Fig 4.16: Bar graph showing initial centroids of Wine Dataset using K-Means method

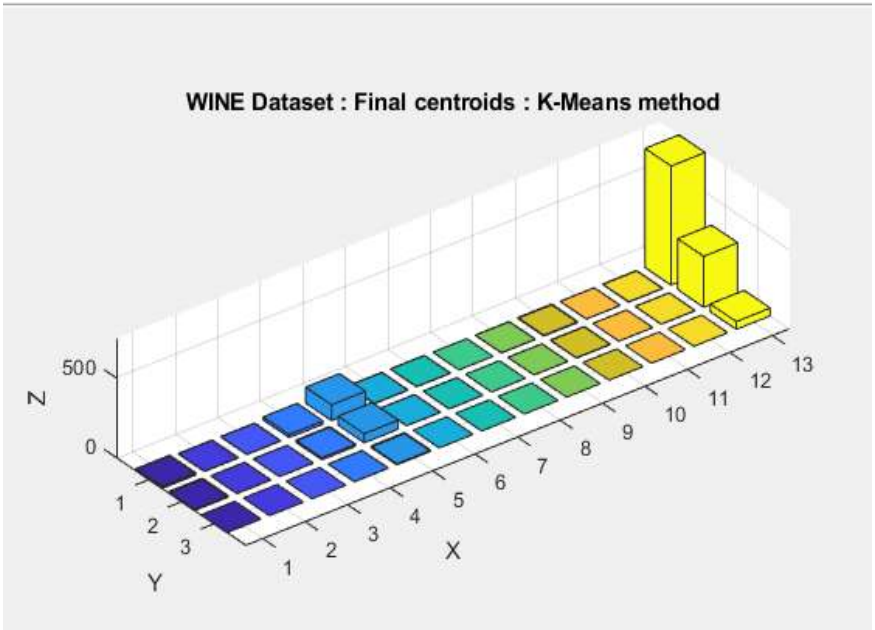

Fig 4.17: Bar graph showing final centroids of Wine Dataset using K-Means method

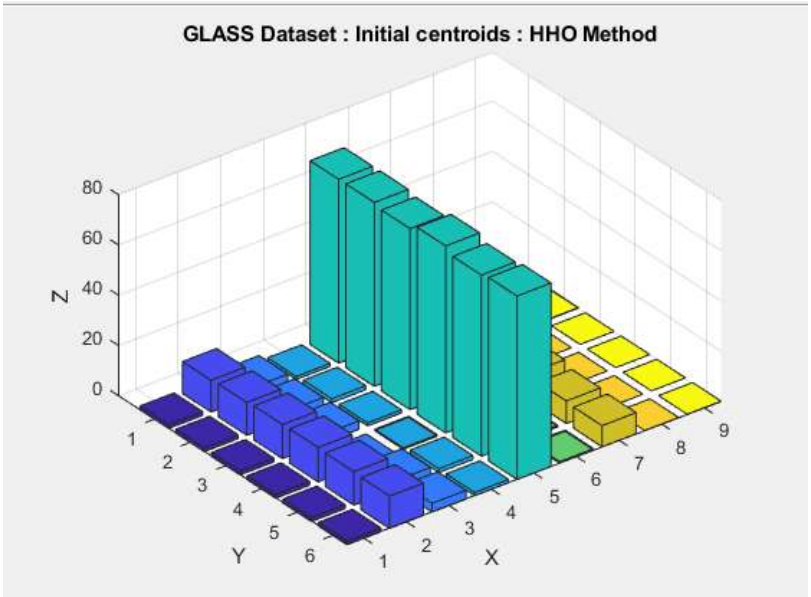

Fig 4.18: Bar graph showing initial centroids of Glass dataset using proposed method

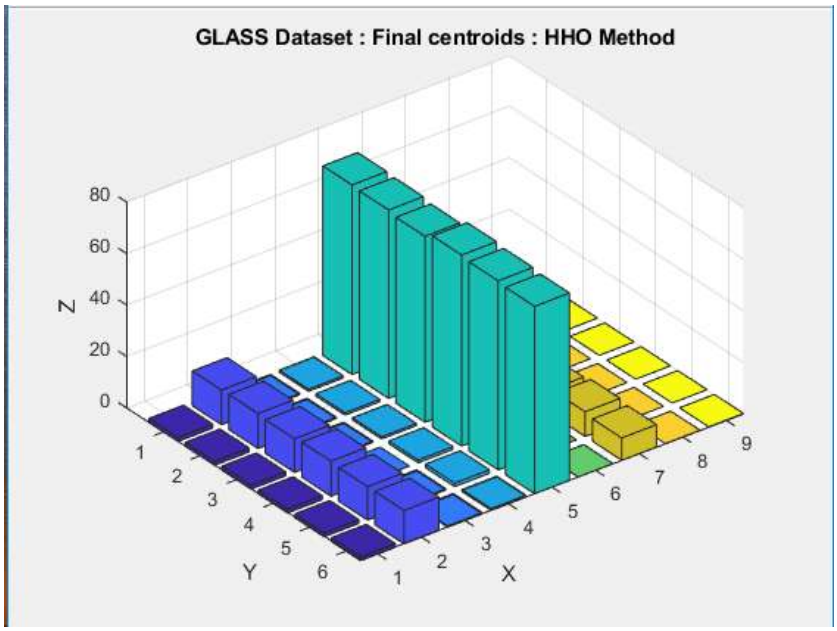

Fig 4.19: Bar graph showing final centroids of Glass Dataset using proposed method

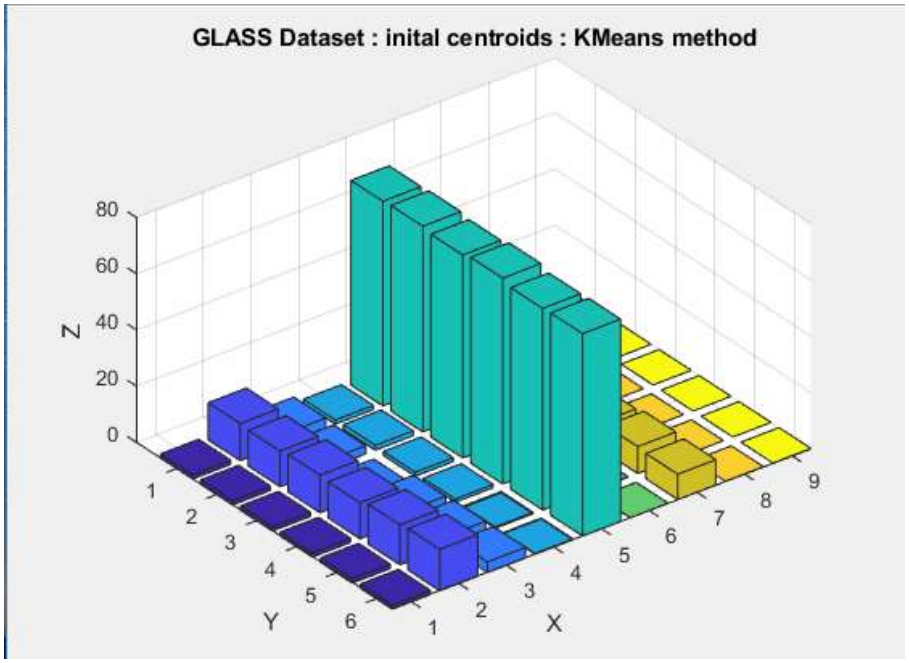

Fig 4.20: Bar graph showing initial centroids of Glass Dataset using K-Means method

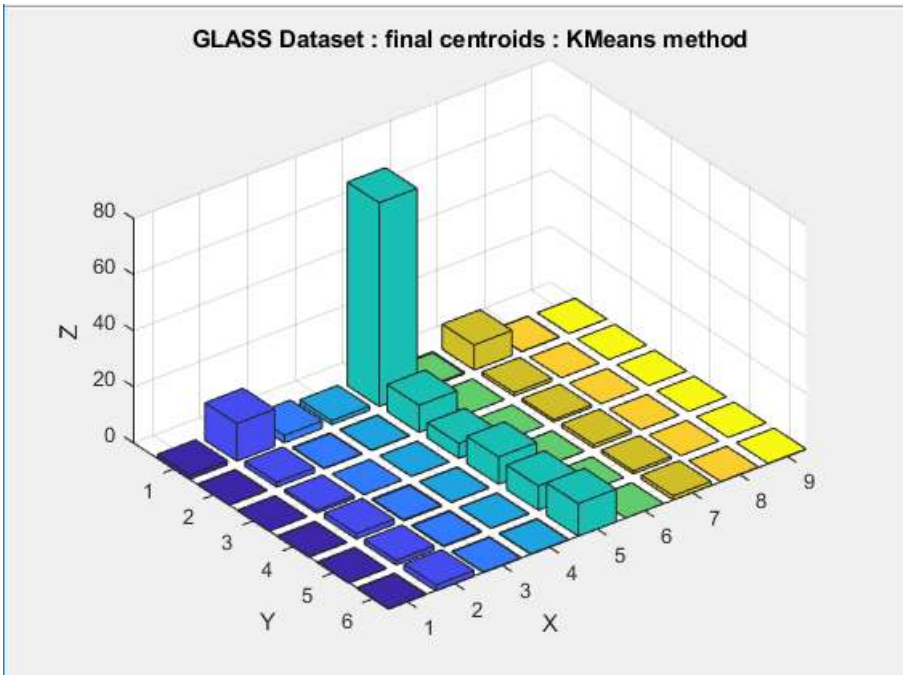

Fig 4.21: Bar graph showing final centroids of Glass Dataset using K-Means method

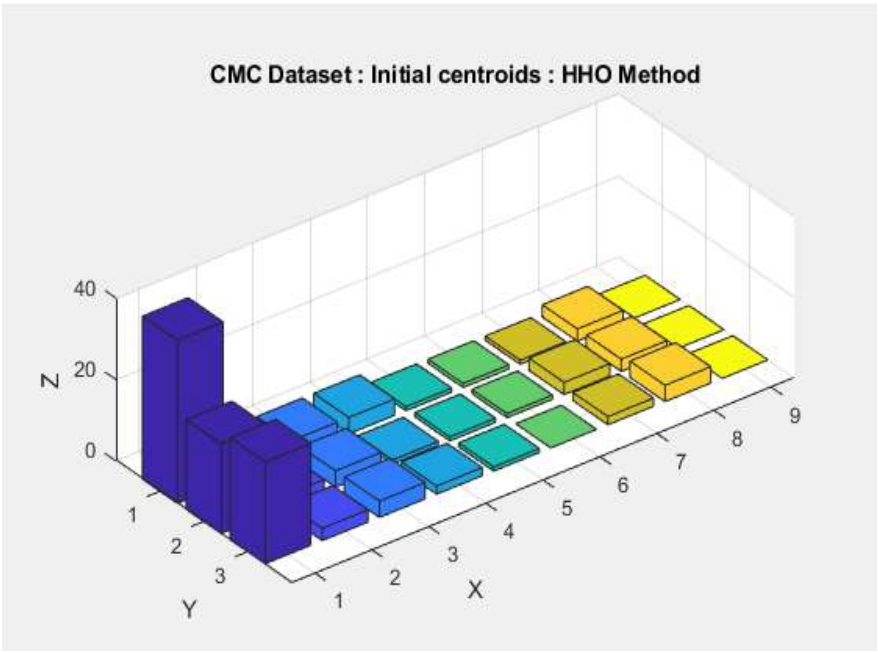

Fig 4.22: Bar graph showing initial centroids of CMC Dataset using proposed method

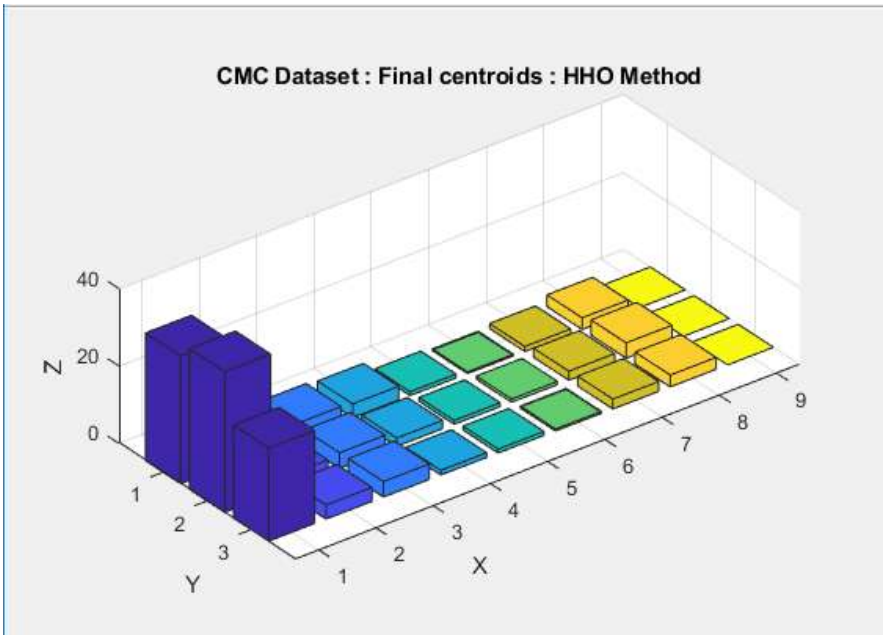

Fig 4.23: Bar graph showing final centroids of CMC Dataset using proposed method

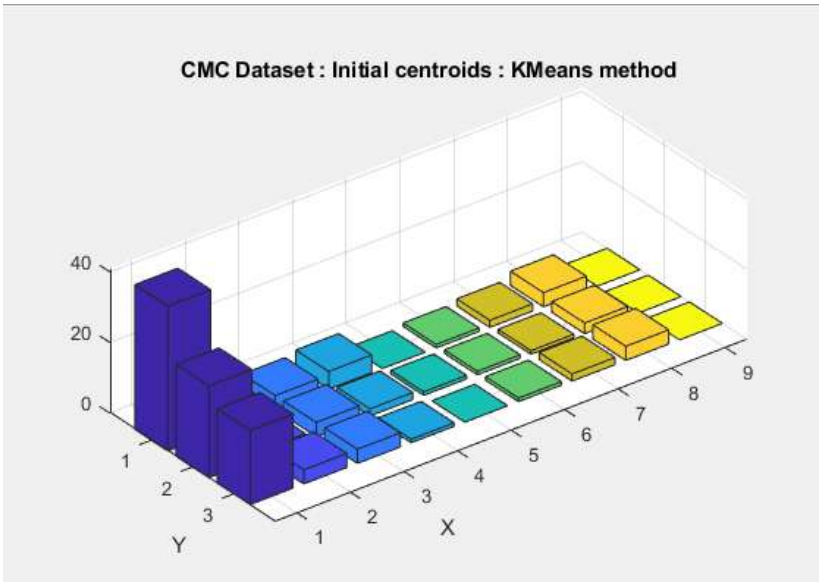

Fig 4.24: Bar graph showing initial centroids of CMC Dataset using K-Means method

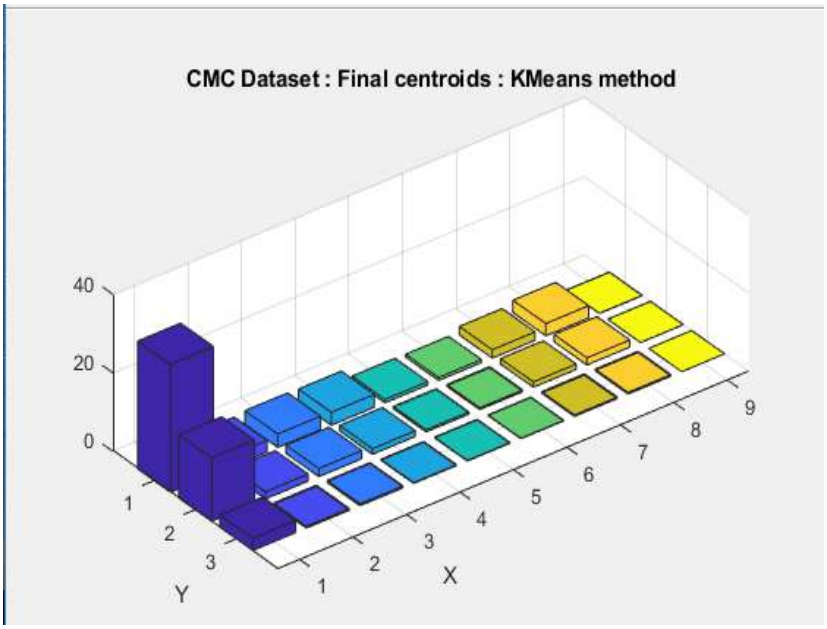

Fig 4.25: Bar graph showing final centroids of CMC Dataset using K-Means method

The following figures draw a comparison between inter- cluster distance of different datasets using K-Means and proposed algorithm through stem plot. It is observed that inter cluster distance obtained through proposed method is much better than the former.

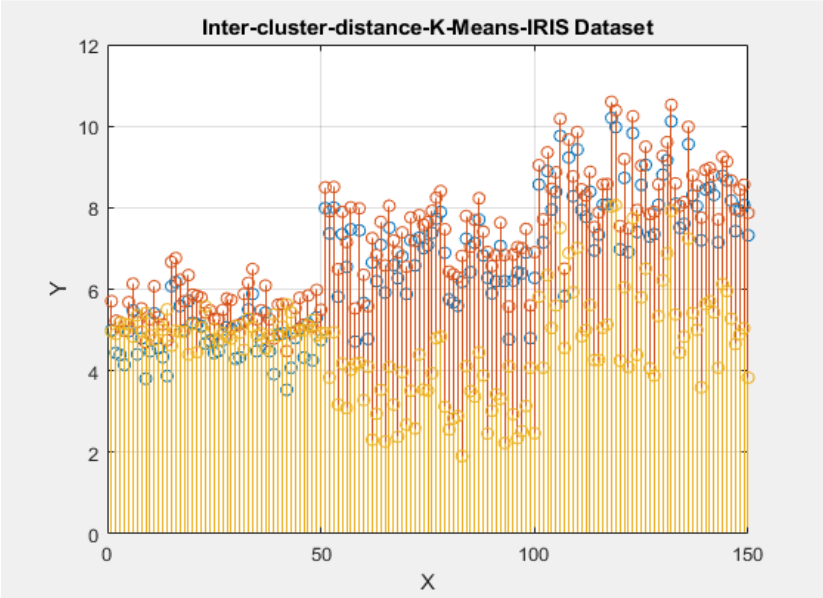

Fig 4.26: Stem plot showing inter-cluster-distance of Iris Dataset using K-Means method

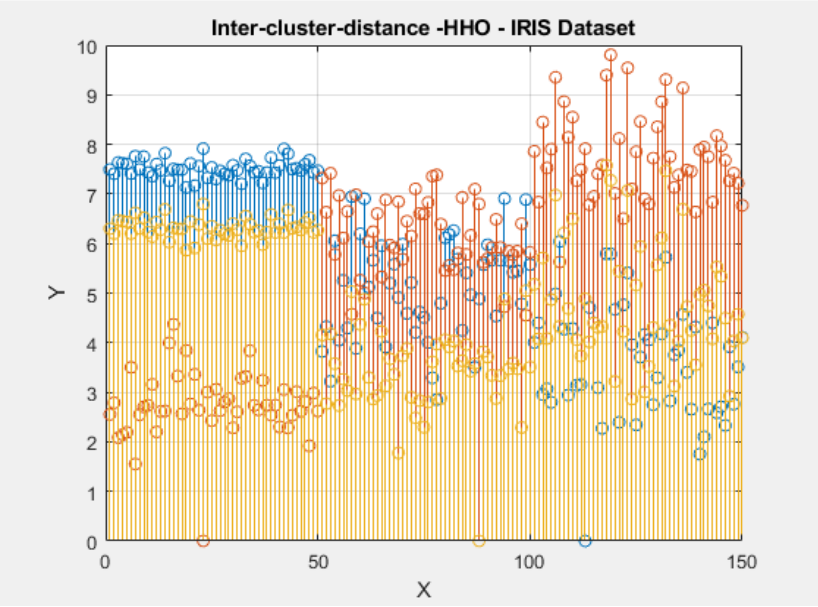

Fig 4.27: Stem plot showing inter-cluster-distance of Iris Dataset using proposed method

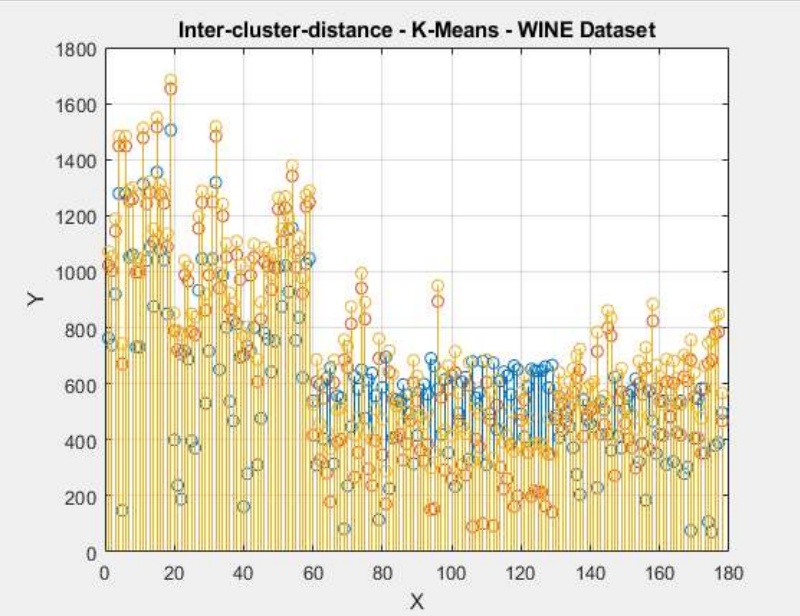

Fig 4.28: Stem plot showing inter-cluster-distance of Wine Dataset using K-Means method

Fig 4.29: Stem plot showing inter-cluster distance of Wine Dataset using proposed method

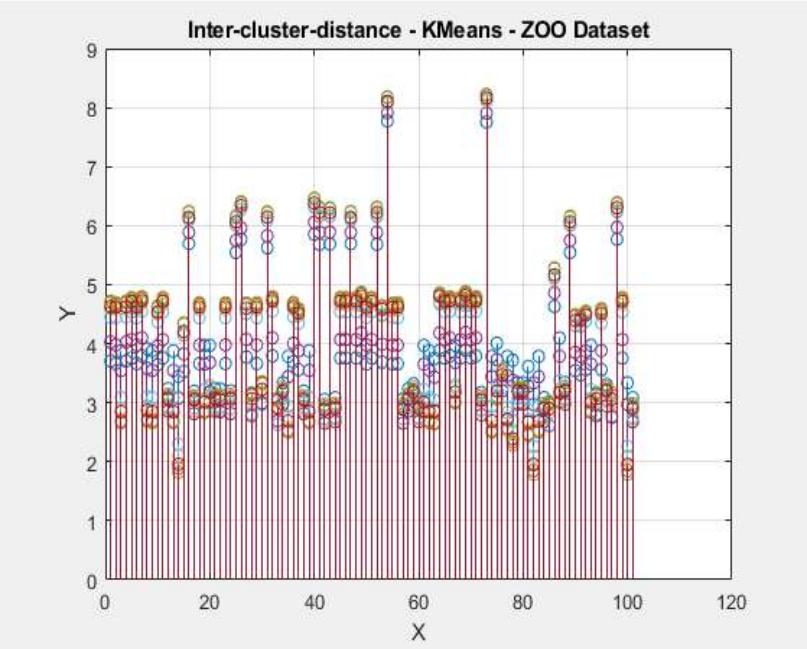

Fig 4.30: Stem plot showing inter-cluster-distance of Zoo dataset using K-Means method

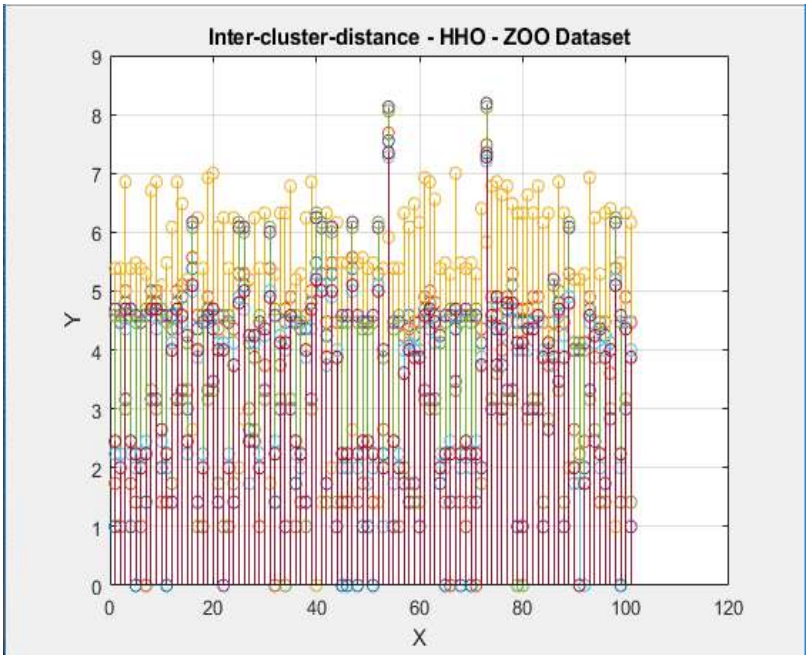

Fig 4.31: Stem plot showing inter-cluster-distance of Zoo Dataset using proposed method

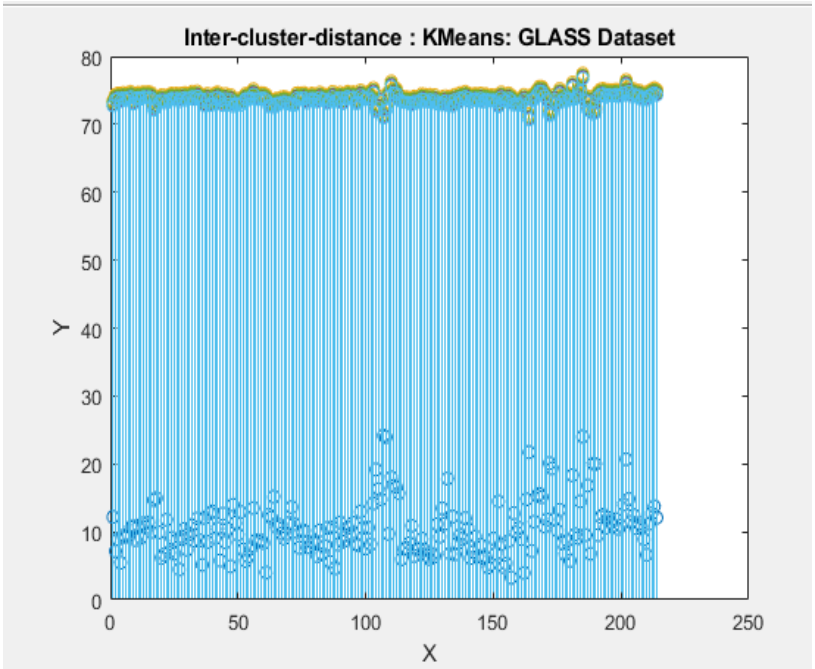

Fig 4.32: Stem plot showing inter-cluster-distance of Glass Datasetusing K-Means method

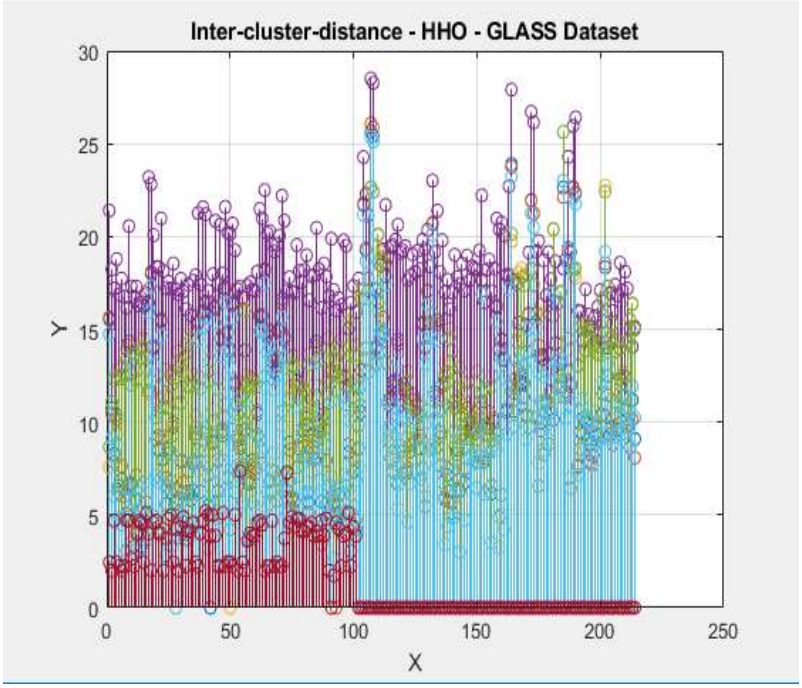

Fig 4.33: Stem plot showing inter-cluster-distance of Glass Dataset using proposed method

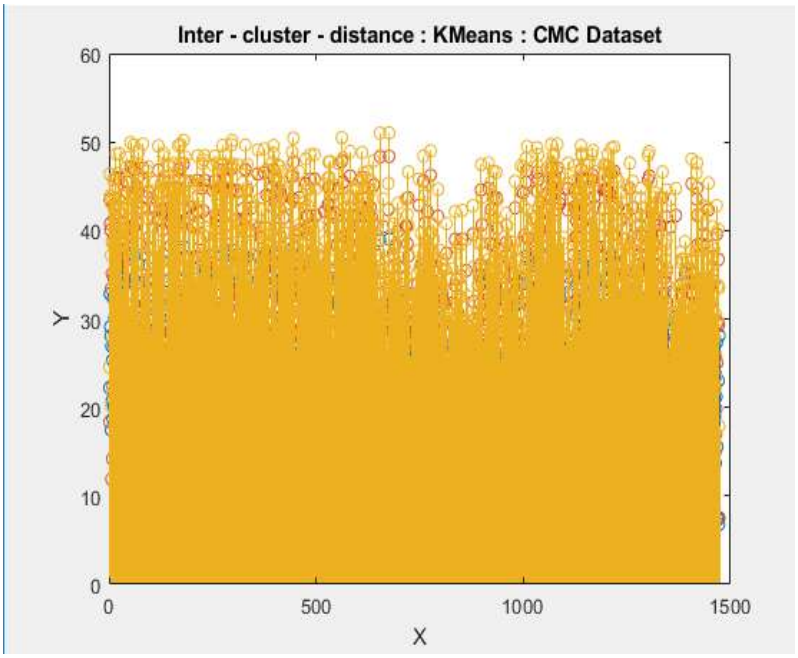

Fig 4.34: Stem plot showing inter-cluster distance of CMC Dataset using K-Means method

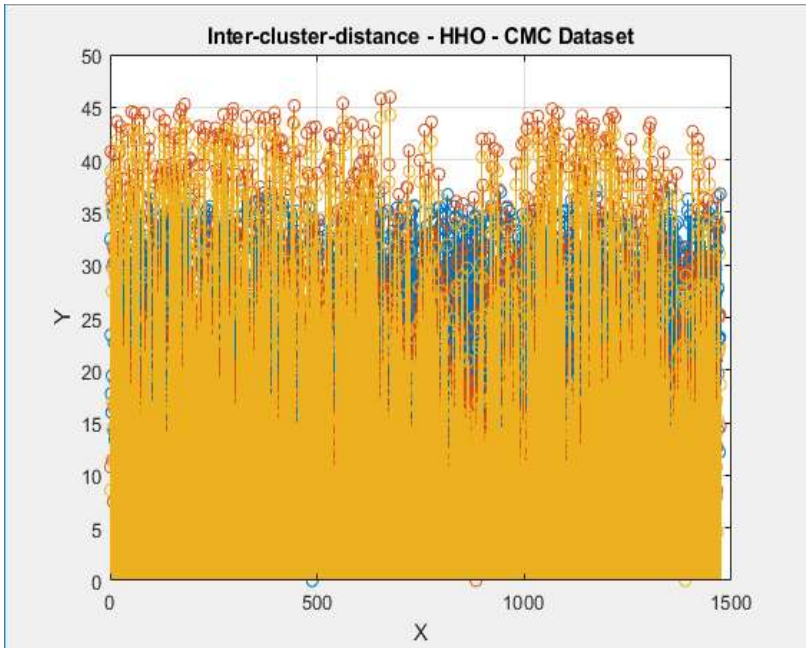

Fig 4.35: Stem plot showing inter-cluster distance of CMC Dataset using proposed method

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 General

In this work, harris-hawk meta-heuristic optimization algorithm is presented for solving partition clustering problems. The proposed algorithm is inspired from cooperative behaviour and surprise pounce chasing style of hawks. In the K-Means algorithm, clusters are updated because of "mean" method. The proposed algorithm aims to optimize cluster centres through "exploration and exploitation technique" followed by harris-hawks. Hawks represent the number of clusters needed. Locations of rabbit represent the cluster centre. Initial centroids are the initial rabbit locations. Final centroids are represented by new rabbit locations. Best location of rabbit is obtained through maximum value of accuracy. The proposed algorithm is evaluated on two parameters – accuracy and intra-cluster distance.

5.2 Future Works

In future works, the proposed algorithm can be deployed to work for multi-objective clustering. It can be used to solve more practical engineering problems with superior performance. The binary and multi-objective versions of harris - hawk optimization can also be used in clustering.

References

[1].Jain, Nikita, and Vishal Srivastava. "Data mining techniques: a survey paper." *IJRET: International Journal of Research in Engineering and Technology* 2.11 (2013): 2319-1163.

[2].Heidari, Ali Asghar, et al. "Harris hawks optimization: Algorithm and applications." *Future generation computer systems* 97(2019): 849-872.

[3].Baalamurugan, K.M., and S. Vijay Bhanu. "An efficient clustering scheme for cloud computing problems using metaheuristic algorithms", *Cluster Computing* 22.5 (2019):12917-12927

[4].Moh'd Alia, Osama, et al. "Data clustering using harmony search algorithm", *International Conference on Swarm, Evolutionary and Memetic Computing.* Springer, Berlin, Heidelberg, 2011

[5]. Zhang, Changsheng, Dantong Ouyang, and Jiaxu Ning. "An artificial bee colony approach for clustering." *Expert systems with applications* 37.7 (2010): 4761-4767.

[6]. Pal,Raju, and Mukesh Saraswat. "Data clustering using enhanced biogeography-based optimization." *2017 Tenth International Conference on Contemporary Computing (IC3)* IEEE, 2017.

[7]. Kumar, Vijay, Jitender Kumar Chhabra, and Dinesh Kumar. "Data clustering using differential search algorithm." *Pertanika J. Sci. &Technol*24.2 (2016): 295-306.

[8]. Hatamlou, Abdolreza. "Black hole: A new heuristic optimization approach for data clustering." *Information sciences* 222 (2013): 175-184.

[9]. Lukasik, Szymon, et al. "Data clustering with grasshopper optimization algorithm." *2017 Federated Conference on Computer Science and Information Systems (FedCSIS).* IEEE, 2017.

[10]. Zhou, Yongquan, et al. "Automatic data clustering using nature-inspired symbiotic organism search algorithm." *Knowledge-Based Systems* 163(2019): 546:557.

[11]. Tsai, Chun-Wei, et al. "A high-performance parallel coral reef optimization for data clustering." *Soft computing* 23.19 (2019): 9327-9340.

[12]. Marinakis, Yannis, et al. “A hybrid stochastic genetic-GRASP algorithm for clustering analysis.” *Operational Research* 8.1(2008): 33-46.

[13]. Aljarah, Ibrahim, et al. “Clustering analysis using a novel locality-informed grey wolf-inspired clustering approach.” *Knowledge and Information Systems* 62.2 (2020): 507-539

[14]. Tang, Rui, et al. “Integrating nature-inspired optimization algorithms to K-means clustering.” *Seventh International Conference on Digital Informatin Management (ICDIM, 2012),* IEEE, 2012

[15]. Jarboui, Bassem, et al. “Combinatorial particle swarm optimization (CPSO) for partitional clustering problem.” *Applied Mathematics and Computation* 192.2(2007): 337-345.

[16]. Lakshmi, K., N. KarthikeyaniVisalakshi, and S. Shanthi. “Data clustering using K-means based on crow search algorithm.” *Sadhna*43.11(2018): 190.

[17]. Yapici, H., & Cetinkaya, N. (2019). A new meta-heuristic optimizer: Pathfinder algorithm. *Applied Soft Computing, 78,* 545-568.

[18]. Zhang, Qingyang, et al. “Collective decision optimization method.” *Neurocomputing* 221 (2017): 123-137.

[19]. Senthilnath, J., et al. “FPA clust: evaluation of the flower pollination algorithm for data clustering.” *Evolutionary Intelligence* (2019): 1-11.

[20]. Kaveh, A., and M. Khayatazad. “A new meta-heuristic method: ray optimization.” *Computers & structures* 112(2012): 283-294.

[21]. Wang, Rui, et al. “Flower pollination algorithm with bee pollinator for cluster analysis.” *Information Processing Letters* 116.1(2016): 1-14.

[22]. Zhou, Yongquan et al. “A simplex method-based social spider optimization algorithm for clustering analysis.” *Engineering Applications of Artificial Intelligence* 64(2017): 67-82.

[23]. Abdel-Basset, Mohamed, and Laila A. Shawky. “Flower pollination algorithm: a comprehensive review.” *Artificial Intelligence Review* 52.4(2019): 2533-2557.

[24]. Yang, X., Luo, Q., Zhnag, J., Wu, X., &Zhou, Y. (2017,August). Moth swarm algorithm for clustering analysis. In *International Conference on Intelligent Computing* (pp. 503-514). Springer, Cham.

[25]. Alyasseri, Zaid Abdi Alkareem, et al. "Variants of the flower pollination algorithm: a review." *Nature-InspiredAlgorithms and Applied Optimization.* Springer, Cham, 2018. 91-118

[26]. Hatamlou,Abdolreza, Salwani Abdullah, and Hossein Nezamabadi-Pour. "A combined approach for clustering based on K-means and gravitational search algorithms." *Swarm and Evolutionary Computation* 6(2012): 47-52.

[27]. Boushaki, Saida Ishak, Nadjet Kamel, and Omar Bendjeghaba. "A new quantum chaotic search algorithm for data clustering." *Expert Systems with Applications* 96(2018): 358-372.

[28]. Mageshkumar, C., S. Karthik, and V. P. Arunachalam. "Hybrid metaheuristic algorithm for improving the efficiency of data clustering."*Cluster Computing* 22.1 (2019): 435-442

Pavika-Plag

ORIGINALITY REPORT

8%	0%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	C. Mageshkumar, S. Karthik, V. P. Arunachalam. "Hybrid metaheuristic algorithm for improving the efficiency of data clustering", Cluster Computing, 2018 Publication	2%
2	Submitted to Birla Institute of Technology Student Paper	1%
3	Yamina Mohamed Ben Ali. "Unsupervised Clustering Based an Adaptive Particle Swarm Optimization Algorithm", Neural Processing Letters, 2015 Publication	1%

4 Rui Tang, Simon Fong, Xin-She Yang, Suash Deb. "Integrating nature-inspired optimization algorithms to K-means clustering", Seventh International Conference on Digital Information Management (ICDIM 2012), 2012
Publication
1%

5 FarzanehZabihi, Babak Nasiri. "A Novel History-driven Artificial Bee Colony Algorithm for Data Clustering", Applied Soft Computing, 2018
Publication
1%

6 "Intelligent Computing Methodologies", Springer Science and Business Media LLC, 2017
Publication
1%

7 Submitted to University of South Florida
Student Paper
<1%

8 Submitted to University of Malaya
Student Paper
<1%

9	Submitted to Thapar University, Patiala Student Paper	<1%
10	Chun-Wei Tsai, Huei-Jyun Song, Ming-Chao Chiang. "A hyper-heuristic clustering algorithm", 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012 Publication	<1%
11	Submitted to Higher Education Commission Pakistan Student Paper	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 15/07/2020

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: Pavika Bhardwaj Department: CSE/IT Enrolment No 182201

Contact No. _____ E-mail. _____

Name of the Supervisor: Dr. Pardeep Kumar/ Dr. Yugal Kumar

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): _____

Designa new Clustering algorithm for partition Clustering problem

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 8%.....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Pardeep Kumar.

(Signature of Guide/Supervisor) 15/07/2020 Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String	8%	Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature

Librarian

.....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com