Water Wave Optimization Algorithm for Medical Datasets

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

in

Computer Science and Engineering/Information Technology

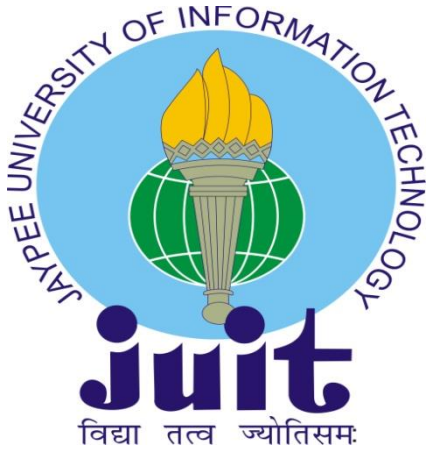
By

Lokinder Singh Mittal (151261)

Under the supervision of

Dr. Yugal Kumar

to



Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Water Wave Optimization Algorithm for Medical Datasets"** in fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2018 to May 2019 under the supervision of **Dr.Yugal Kumar**,**Assistant Professor (Senior Grade)** The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Lokinder Singh Mittal (151261)

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

Dr. Yugal Kumar
Assistant Professor (Senior Grade)
Department of Computer Science & Engineering and Information Technology

Dated:

Acknowledgement

I wish to express my profound and sincere gratitude to Dr. Yugal Kumar, Assistant Professor(Senior Grade), Department of Computer Science and Information Technology, Jaypee University of Information Technology, who guided us into the intricacies of this project with matchless magnanimity. He constantly co-operated and helped with the research work. He also evinced keen interest and invaluable support in the field of Machine Learning for progress of our project work.

Date: **Lokinder Singh Mittal**

TABLE OF CONTENTS

List of Figures

List of Graphs

List of Tables

Abstract

Clustering is a tool for data mining used to extract the hidden information of various structures and “clusters” found in large data sets. In the fields of science and engineering, it is observed that the trend has shifted toward the use of nature-inspired computing techniques. The report presents the new meta-heuristic, that is, the Water Wave Optimization (WWO) technique for solving various global optimization problems. Vibrating Particle System is population based meta-heuristic algorithm based on the damped free vibration of single degree of freedom system. We have evaluated the proposed algorithm on a set of 5 benchmark datasets based on “health care” taken from the UCI Machine Learning Repository. The computational results show that WWO outshines the other state-of-the-art algorithms in terms of calculations and accuracy measures.

CHAPTER-1

INTRODUCTION

1.1 Introduction

The clustering process is the process of identifying associated data groups in a given input data set. Entities in a group under consideration are relatively more alike to those in that group than those in other groups. The main objective of clustering is to divide the input data points or the population into numerous groups in such a way that the similarity index of the data points belonging to a particular group is relatively high as compared to other data points belonging to other groups or clusters. The prime objective of clustering is to partition groups having similar attributes and then map each and every input data point to one of the partitions or clusters. In general, clustering can be divided into two categories, hard and soft clustering. In the former, each and every tuple is either completely assigned to a cluster or not, whereas in the latter, instead of assigning each and every data point to a separate partition or cluster, the likelihood that the data point is in one of those clusters is assigned.

Broadly speaking within the domain of machine learning, the tasks can be classified into two main categories: supervised and unsupervised. Moreover, in the case of supervised learning, we have an idea or information about exactly what output the algorithm or model should give for our sample points. Therefore, the primary goal of supervised learning is to learn a function that best assumes and foresees the relationship between the input and output recorded in the data along with a given sample of data sets and a desired output. Whereas unsupervised learning does not have a defined label output, so the objective of unsupervised learning is to infer the natural structure present within a given set of data points. Supervised learning is mainly done in the classification context, in instances where we want to generate a relationship between the input and output labels and finally generates a mapping onto them or in the scenario where our intention is to build a mapping between the input and output that is continuous. Commonly used algorithms for supervised learning include naive bays, logistical regression, vector support machines, artificial neural networks and random forests. In both regression as

well as classification, the ultimate goal is to find specific structures or relationships in the input data that allow us to efficiently produce correct output data. While doing supervised learning, the chief points that have to be considered are model complexity, and the bias-variance trade off which are both correlated.

Unsupervised learning on the other hand is the type of learning where we only have the input data and have no corresponding output variable for that particular input data .The main objective of unsupervised learning is to model the distribution in the data so as to have more idea about the input data. In case of unsupervised learning the algorithms are left to their own devises to find out the interesting structure that is present in the input data.

In cluster analysis or clustering, two types of distances are calculated. One is "inter-cluster distance", that is, the distance between two different clusters, or, it can be calculated by calculating the distance between the two centroids. Another one is "intra-cluster distance", that is, the distance between the objects or data points in the same cluster.

There are a number of clustering algorithms that are known today. Each of the clustering algorithm or methodology follows a unique set of rules in order to define the type of similarity that exists among the input data points. A few of them are:

i. DENSITY MODELS:

   In the case of density models, the data space is searched to identify areas of varying data point density within the data space. This model essentially isolates the different density regions and then assigns or designates the data input points in the same cluster in these regions. Examples of density based clustering include Density Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points to Identify the Clustering Structure (OPTICS).

ii. DISTRIBUTION MODELS

   The distribution clustering models are fundamentally based on the idea that all data points in a particular cluster are actually the same. Over fitting is an issue that these models face very often. Expectation-maximization algorithm is a popular example of the distribution model. This algorithm makes use of multivariate normal distributions.

iii. CONNECTIVITY MODELS

The connectivity models are essentially based on the idea that the input data points which are closer to each other in the data space show higher degree of similarity to each other when contrasted against those that are at a significant distance in the data space. The connectivity models can follow two distinct methodologies. The first approach is based on classifying all data points into separate, distinct clusters. The classification of data points is followed up by combining them as the distance between the points under consideration decreases. Whereas in the second approach initially all the data points are classified as a single cluster post where these are then partitioned as and when the distance increases. Both these models have a major drawback that they lack scalability which makes them unsuitable for handling bigger datasets. Examples of this type of clustering include Hierarchical Clustering and its applied variations.

iv. CENTROID MODELS

These are essentially the iterative clustering algorithms in which the similarity concept is determined by the proximity of the cluster centre to the data points. The K-Means algorithm for clustering is a well liked algorithm in the centroid model category. The prerequisite of these models is that the number of clusters that are necessary at the end have to be mentioned beforehand .Therefore it is imperative that we have former information of the dataset.

K Means Clustering:

The procedure follows an easy method to categorize a given data set through a certain number of clusters which is fixed beforehand. The key scheme is to define k centroids, one for each of the cluster. The better choice is to place each of the centroid as far away from each away as possible. After identifying the cluster centres the next step includes taking each point belonging to a given input data set and associating it to the nearest centroid. When all the points are covered, the first step is completed. After the completion of this step k new centroids are recalculated as barycentre of the clusters resulting from the last step. After we

have calculated the k new centroids, a new binding has to be done between the same input data set points and the nearest new centroid. As a result of this loop formation we may notice that the k centroids change their location step by step until no more changes are done

In our project the main focus is on the implementation of the Water wave Optimization (WWO) technique for solving various global optimization problems and then subsequently comparing the performance of the Vibrating Particle System clustering algorithm against the performance of Water Wave Optimization clustering algorithm for the same data sets..What the Vibrating Particle System algorithm essentially does is that it stimulates the free vibration of single degree of freedom systems with viscous damping .Vibrating Particle System is a new meta-heuristic algorithm based on the free vibration of single degree of viscous damping freedom systems. In the case of the WWO algorithm, the candidates for the solution are regarded as particles that gradually approach their balance positions .Balance positions are attained with the present population and historically the best position to balance diversification and intensification properly. To evaluate the performance of the proposed method is applied on various health care data sets to obtain optimized cluster centres.

This algorithm is straightforward and simple to apply that can be applied on various engineering optimization problems that are present in the real-world. We have evaluated the proposed algorithm on a set of 5 benchmark datasets all of which are based on “health care”. The computational results show that WWO outshines the other state-of-the-art algorithms in terms of calculations and accuracy measures.

1.2 Problem Statement

In this section of project report, the need of algorithm is illustrated.
A number of clustering algorithms have been implemented in the past for various optimization problems but almost all of them have faced similar problems some of which are listed as:

I. Lack of balance between exploration and exploitation processes.
   Exploration and exploitation have often been regarded as the two keystone terms around which organizational adaptation research revolves but unfortunately the previously implemented algorithms are unable to strike a harmonious balance between the exploration and exploitation processes.

II. Lack of diversity and local optima
III. Slow convergence problem in the last generations of the iterative process

The aim is to find the best cluster centroids (cluster representative) which have the minimum or least intra-cluster distances between them. Also, the main focus is on to increase the accuracy of the dataset. The other objective is of this project is to introduce the chaotic maps into the proposed algorithm to resolve randomness and to implement a local search method to enhance the quality of the solution and the problems of local optima. Our project aims at overcoming these shortcomings by implementation of the Water wave Optimization for solving various global optimization problem. Experiments on different datasets based on health care show that WWO is a competitive clustering algorithm with the present meta-heuristic algorithm. We have successfully applied the VPS algorithm on different health care based optimization problems, the results of which exhibit the applicability and effectiveness of WWO to real world problems.

Also the project aims at comparing the performance of the Vibrating Particle System clustering algorithm against the performance of Water Wave Optimization clustering algorithm for the same data sets.

1.3 Objectives

The project aims to implement the Water wave Optimization on various healthcare data sets in order to minimize the intra cluster distance and obtain optimized cluster centres. Also the project aims at comparing the performance of the Vibrating Particle System clustering algorithm against the performance of Water Wave Optimization clustering algorithm for the same data sets.

The sole objective of this project is to investigate the performance of the Water wave Optimization (WWO) algorithm on health care datasets and compare the performance of the VPS and WWO algorithm by implementing them on the same data sets as mentioned below.

i. Dermatology
ii. Heart

iii. WDBC
iv. Thyroid
v. Bupa
vi. BCW

All these datasets have been taken from the UCI Machine Learning Repository. All these datasets are taken individually and are monitored or examined using the VPS algorithm and various functions that are being used.

Thus, our aim is to build clusters and find the data which has higher accuracy. The data points are selected from these dataset and different classes are made. These classes are made based on the section that has been provided in the dataset taken from "UCI Machine Learning Repository". After the identification of the classes is done the clusters are then made according to the classes that have been previously identified.

1.4 Methodologies

The ultimate objective of clustering is to identify similarities among the data point and then group these similar data points together. Over time a number of algorithms have been developed to implement this clustering. In our project we have made use of one of the most popular and widely used algorithm in machine learning which is the K means clustering. Our project involves making use of unsupervised learning under the category of clustering along with the classification of the data. The ultimate objective of our project is to find the besluster centroids which have the least or minimum intra-cluster distances among them. Also, the primary focus is on increasing the overall accuracy of the dataset in hand. Experiments and computations on the different datasets based on health care clearly indicate that VPS is indeed a competitive clustering algorithm with the present meta-heuristic algorithm.

The training of model is done by working on the algorithm using the modified vibrating particle system optimization algorithm and then running some functions which is called time to time in order to find the fitness and accuracy of the input data points. The point which is found to have higher accuracy percentage among all the points is then declared as the fittest of all.

Our dataset is based on health care. Social insurance examination depends on information and informational indexes specifically. Human services informational collections incorporate immense measure of medicinal information, different estimations, monetary

information, factual information, socio-economics of particular populaces, and protection information, to give some examples, assembled from different social insurance information sources. Because of the assorted variety of medicinal services information sources information institutionalization is a key column for productive and important utilization of the data and joint effort of social insurance experts, care suppliers, back up plans, and government organizations.

1.4.1 K Means Clustering

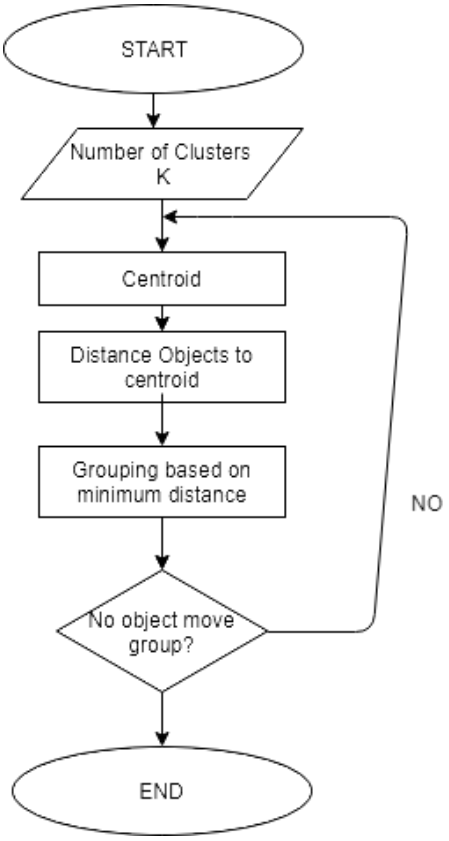


Fig 1.1 Steps in K-Means Clustering

Steps involved in “K Means clustering” algorithm are

1. The first step starts with K as the input which is the total number of clusters you want to find. These K centroids are then placed in random locations in the space.
2. Now by making use of the euclidean distance between each of the data points and centroids every data tuple is assigned to the cluster which is the nearest to it.
3. The third step includes recalculating the cluster centres by finding the mean of data points assigned to that particular cluster centre.

4. Steps 2 and 3 are repeated till a point when no further changes occur.

The k means clustering method precisely produces different clusters. The main objective of K-Means clustering is to reduce total intra-cluster variance, or, the squared error function to a Minimum value:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

$J: Objective\ Function$
$K: Number\ of\ Cluster$
$N: Number\ of\ cases$

1.5 Organization

Chapter 1: Includes a brief introduction to the project. A basic idea of what we are doing and what we are trying to accomplish with this .We provide with the problem statement and the objective of our project with the inclusion of the basic concepts required to accomplish the goal.

Chapter 2: Includes literature survey. We have studied various papers and journal from reputed sources on K –means clustering and meta-heuristic algorithm approach.

Chapter 3: Includes details on system development. This section provides with the overall procedure and steps which are involved in the project. We have provided with a flow chart of the algorithm with explanation for better understanding of the concepts involved. This chapter also includes the algorithm that we are implementing with the explanation for the same and the equations which are used. This chapter depicts the test plan used .We have described the various datasets that we are going to implement on our algorithm.

Chapter 4: This chapter provides with the results and performance analysis. We have applied the algorithm on a given dataset and provide with the result values and graph for the same.

Chapter 5:Conclusion for the project is stated .We conclude the project with the observations and conclusions and represent the future work for further development and completion of the project.

CHAPTER-2

LITERATURE SURVEY

In this section of project report, all the information collected from research papers and websites are illustrated.

2.1 Title: "A new nature-inspired meta-heuristic - Water Wave Optimization"

Author: Yu-Jun Zheng

Year of Publications: March 2015

Publishing Details: Elsevier Journal

[1]Clustering is a tool for data mining used to extract the hidden information of various structures and clusters found in large data sets. In the fields of science and engineering, it is observed that the trend has shifted toward the use of nature-inspired computing techniques. The paper presents the new meta-heuristic, that is, the water wave optimization (WWO) technique for solving various optimization problems which are at global scale. The phenomena, such as breaking and propagation, refraction, of water waves are used to search in a HD solution space. WWO is inspired by shallow water-wave theory; and it is the simple and the easy most way to implement algorithm that can be applied on various engineering optimization problems present in the real-world. The computational results show that WWO outshines the other state-of-the art algorithms in terms of calculations and the accuracy measures.

2.2 Title: "A new meta-heuristic algorithm: Vibrating particle system"

Author: A.Kaveh and M.Ilchi Ghazaan

Year of Publications: September 2016

Publishing Details: Scientia Iranica

[2]The vibrating particle system is a new meta-heuristic algorithm based on the introduction of free vibrations with viscous damping of single-degree freedom systems. The candidates for the solution gradually approach their balancing positions and are

regarded as particles. So as to achieve the right amount of balance between diversification and intensification, the current population finds equilibrium positions and historically the best position. The proposed method is used to optimize four skeletal structures, including trusses and frames, in order to evaluate their performance. The proposed method shows its ability to solve limited problems by finding superior optimal designs for three of the four problems examined. The results obtained by VPS are competitive with other methods of optimization and also offer fast converging functions.

2.3 Title: "Fuzzy magnetic optimization clustering algorithm with its applications to health care."

Author: Neetu Kushwaha , Millie Pant

Year of Publications: July 2018

Publishing Details: Springer

[3]Clustering helps in finding hidden structures and "clusters" found in large datasets and is an important tool for data mining and knowledge discovery. Because of its capabilities of clustering the datasets that are uncertain or vague Fuzzy C-means (FCM) is considered as a popular data clustering method. In case of poor initialization the algorithm gets trapped into a local minima and the performance of FCM is usually affected. To overcome this issue, a new clustering algorithm known as fuzzy magnetic optimization clustering is proposed which embeds the concept of fuzzy clustering into magnetic optimization algorithm. Efficiency of Fuzzy –MCO is compared with other three fuzzy clustering algorithms. Its efficiency is calculated on the basis of four different performance parameters: F1, accuracy, purity and RI. The experimental results provide us with a consistent performance of the Fuzzy-MCO algorithm on most of the datasets.

2.4 Title: "An Efficient k-Means Clustering Algorithm: Analysis and implementation."

Author: Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D.Piatko, Ruth Silverman,Angela Y. Wu.

Year of Publications: July 2002

[4]The criteria determining a good cluster relies on the application and there are different ways to find clusters that are subject to different criteria, both systematic as well as ad hoc. The Lloyd algorithm is a popular K-mean clustering heuristic. An simple efficient implementation of Lloyd's k-means clustering Algorithm is proposed, also known as a filtering algorithm. This algorithm requires a kd-tree as the only major data structure required. The efficiency of the filtering algorithms has been recognized by making use of two distinct methods. In the first way as the separation between the cluster centres increases it shows that the performance of the algorithm significantly improves. Secondly, a number of experimental studies usually involving both synthetically formulated data and real data sets from colour quantization, data compression and image segmentation applications. The kd-tree used does not require an update because it is calculated for data points instead of centres. The kd-tree used does not require an update because it is calculated for data points instead of centres. This approach helps to achieve efficiency, as the data points do not vary throughout the calculation and data structure recalculation is not necessary. The method proposed represents the efficiency of the algorithm both theoretically and empirically through data sensitive analysis and experiments on real data sets and synthetic data sets.

2.5 Title: "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm"

Author: K. A. Abdul Nazeer, M. P. Sebastian

Year of Publications: July 2009

[5]With the emergence of modern techniques, there has been an accumulation of large quantities of data from different fields of study have been observed. To extract information from huge data sets conventional database methods have resulted to be inadequate. Therefore, it becomes practically impossible to extract information from a large dataset. The k-means clustering algorithm is a widely used algorithm for many practical applications, one of the most important data analysis methods being cluster analysis. Since it is computationally complex and the initial centroids determine the quality of the resulting clusters, many techniques have been proposed and implemented to improve the performance of the k-means clustering algorithm have been proposed. Different methods are proposed to make a more effective and efficient clustering algorithm with reduced complexity. It also looks out for increasing the accuracy of the algorithm. Based on previous research and results, the computational complexity of the standard k-means algorithm is high, which means that data points must be reassigned several times during each iteration of the loop. The proposed work puts forward an improved version of the k-means algorithm that ensures that the entire clustering process takes O(n^2)time without sacrificing the accuracy of the clusters. The proposed method combines together a methodical method for determining the initial centroids and a competent way for assigning data tuples to clusters.

2.6 Title: "Performance based analysis between k-Means and Fuzzy C – Means clustering algorithms for connection oriented telecommunication data"

Author: T.Velurugan

Year of Publications: February 2014

Publishing Details: Elsevier Journal

[6]The process of discovering significant new correlation, patterns and trends by transferring large amounts of data by making use of pattern recognition techniques/technologies and statistical and mathematical techniques is what is called data mining .Perhaps one of the most important data analysis techniques commonly used in data mining is the cluster analysis. Two partition based clustering algorithms known as k-Means clustering and Fuzzy C –Means are analysed .These algorithms are implemented

in order to analyze its performance based on their computational time. The analysis of the computational complexity of both the algorithms is performed and the results are compared with each other .After the comparison it is obtained that the more accurate and easy to understand from both the algorithms is the Fuzzy C-means in comparison to the k-Means but has a computational time greater than that of k-means algorithm. Through the work proposed it is known that the k-Means algorithm has an advantage of favourable execution time and has a drawback of knowing how many cluster centres are searched for in advance. The data points are distributed evenly in K-Means as compared to FCM which has some variations in the distribution.

2.7 Title: "Improved K-means Algorithm Based on Density Canopy"

Author: Geng Zhang , Chengchang Zhang , Huayu Zhang

Year of Publications: January 2018

Publishing Details: Knowledge based System Journal –Elsevier

[7]An enhanced k-means algorithm developed using the concepts of Canopy density is put forward to enhance the accuracy and stability of the k-means algorithm and to determine the apt number of clusters and the best initial seed. An enhanced k-means algorithm developed using Canopy density is proposed to improve the accuracy and stability of the k-means algorithm and to determine the appropriate number of clusters and the best initial seeds. The method that has been put forward is used as the pre-processing methodologies of k-means and its output is used as the cluster number and the initial clustering centres of k-means algorithm. The density canopy algorithm results in finding the value K of the datasets and initial clustering centres which are used as input parameters of the k-means algorithm. After testing of the new algorithm on various UCI data sets and simulated data sets with different proportions of noise samples the results show that the new improved algorithm is better than the traditional k-means algorithms in terms of clustering results and is insensitive to noisy data .It is stated that the accuracy is improved significantly.

2.8 Title: "A survey on nature inspired meta-heuristic algorithms for partitioned clustering"

Author: Satyasai Jagannath Nanda, Ganapati Panda

Year of Publications: November 2013

Publishing Details: Elsevier Journal

[8]After the beginning of the partition clustering concept perhaps with the K-means algorithm, many classical partition clustering algorithms came into being in 1957. In 1990, a fresh age in the domain of cluster analysis was established with the application of meta-heuristics inspired by nature. A review of all key meta-heuristic algorithms used until now for partitional clustering is shown here. In addition, she discusses the formulation of different meta-heuristics as a clustering problem and application areas.The entire search area with the population involved is explored by meta-heuristic algorithms inspired by nature and ensures optimal partitioning. A single objective algorithm provides an optimal solution, while flexibility is provided by multi-objective algorithms to choose the required solution from a variety of optimum solutions.

2.9 Title: "A Hybrid Meta-heuristic and Kernel Intuitionistic Fuzzy c-means Algorithm for Cluster Analysis"

Author: R.J. Kuo, T.C. Lin, F.E. Zulvia, C.Y. Tsai

Year of Publications: February 2018

Publishing Details: Elsevier Journal

[9]Cluster analysis is an exceptionally helpful information mining approach. Albeit many grouping calculations have been proposed, it is extremely hard to discover a bunching strategy which is reasonable for a wide range of datasets. This examination proposes a transformative based bunching calculation which joins a metaheuristic with a bit intuitionistic fluffy c-implies (KIFCM) calculation. The KIFCM calculation enhances the fluffy c-implies (FCM) calculation by utilizing an intuitionistic fluffy set and a portion work. As indicated by past investigations, the KIFCM calculation is a promising calculation. Notwithstanding, despite everything it has a shortcoming because of its high

affectability to beginning centroids. Along these lines, this examination defeats this issue by utilizing a metaheuristic calculation to enhance the KIFCM result. The metaheuristic can give better introductory centroids to the KIFCM calculation. This examination applies three metaheuristics, molecule swarm streamlining (PSO), hereditary calculation (GA) and counterfeit honey bee state (ABC) calculations.

2.10 Title: "Clustering performance comparison of new generation meta-heuristic algorithms."

Author: Lale Ozbakõr, Fatma Turna

Year of Publications: May 2017

Publishing Details: Knowledge-Based Systems-Elsevier

[10]Two meta-heuristic algorithms of new generation are discussed here. Benchmark standard test functions have been used to demonstrate the performance of these algorithms.These algorithms are used to solve the problem of clustering. The so-called Ions Motion Optimization is one of the two algorithms and is based on ion motions in nature.

After comparing the results obtained both the algorithms are competitive solution approaches for clustering problems .Both IMO and WSA algorithms seem to be promising new generation meta-heuristics algorithms.

2.11 Title: "Refining Initial Points for K-Means Clustering"

Author: Paul S. Bradley, Usama M. Fayyad

Year of Publications: May 2012

Publishing Details: Proceedings of the 15th International Conference on Machine Learning

[11]Viable clustering methods use an iterative method (e.g. K-Means, EM) that combines one of several nearby minima. It is clear that these iterative strategies are particularly relevant to initiating conditions. We present a method for determining a refined starting

condition from a given starting condition that depends on an effective system for evaluating the circulation methods. The refined starting condition allows the iterative calculation to at least unite with a higher neighbourhood. The strategy is relevant to a wide range of group calculations for discrete and uninterrupted information.

2.12 Title: "Cluster centre initialization algorithm for K-mean Clustering"

Author: Shehroz S. Khan, Amir Ahmad

Year of Publications: July 2004

Publishing Details: Elsevier Journal

[12]Execution of iterative bunching calculations which combines to various nearby minima depend exceptionally on starting group focuses. For the most part introductory group focuses are chosen arbitrarily. In this paper we propose a calculation to register starting bunch places for K-implies grouping. This calculation depends on two perceptions that a portion of the examples are fundamentally the same as one another and that is the reason they have same bunch participation independent to the decision of starting group focuses. Likewise, an individual trait may give some data about introductory group focus. The underlying bunch focuses registered utilizing this approach are observed to be near the coveted group focuses, for iterative grouping calculations. This strategy is relevant to grouping calculations for persistent information. We exhibit the use of proposed calculation to K-implies grouping calculation. The exploratory outcomes indicate enhanced and predictable arrangements utilizing the proposed calculation.

2.13 Title: "A nature-inspired meta heuristic algorithm: Lion Optimization Algorithm (LOA)"

Author: Fariborz Jolai & Maziar Yazdani

Year of Publications: December 2015

Publishing Details: IEEE Explore

[13]Amid the previous decade, taking care of complex streamlining issues with meta-heuristic calculations has gotten significant consideration among specialists and scientists. Henceforth, numerous meta-heuristic calculations have been created in the course of the most recent years. A large number of these calculations are enlivened by different marvels of nature. In this paper, another populace based calculation, the Lion Optimization Algorithm (LOA), is presented. Uncommon way of life of lions and their participation attributes has been the fundamental inspiration for improvement of this streamlining calculation. Some benchmark issues are chosen from the writing, and the arrangement of the proposed calculation has been contrasted and those of some notable and most up to date meta-heuristics for these issues. The outcomes affirm the superior of the proposed calculation in contrast with alternate calculations utilized in this paper.

2.14 Title: "A new meta-heuristic algorithm: Ant colony optimization"

Author: G. Di Caro & M. Dorigo

Year of Publications: July 1999

Publishing Details: IEEE Explore

[14]As of late, various calculations motivated by the scavenging conduct of subterranean insect states have been connected to the arrangement of troublesome discrete advancement issues. We put these calculations in a typical system by characterizing the Ant Colony Optimization (ACO) meta-heuristic. A few paradigmatic instances of utilizations of this novel meta-heuristic are given, and in addition a short diagram of existing applications

2.15 Title: "Optimal Cluster Analysis: Metaheuristic with Fuzzy c-means" Algorithm for Hybrid Chemical Reaction

Author: Ajith Abraham, Bighnaraj Naik , Janmenjoy Nayak & Himansu Sekhar Behera

Year of Publications: February 2017

Publishing Details: Elsevier Journal

[15]Hybridization of at least two calculations has dependably been an unmistakable fascination of research because of the nature of enhancement in seeking ability. Taking the positive bits of knowledge of both the calculations, the created half breed calculation attempts to limit the generous constraints. Bunching is an unsupervised learning technique, which bunches the information as indicated by their comparative or unique properties. Fluffy c-implies (FCM) is one of the prevalently utilized grouping calculations and performs better when contrasted with other bunching methods, for example, k-implies. In any case, FCM has certain impediments, for example, untimely catching at neighbourhood minima and high affectability to the group focus instatement. Thinking about these issues, this examination proposes a novel half breed approach of FCM with an as of late created compound based meta-heuristic for getting ideal group focuses. The execution of the proposed methodology is looked at as far as bunch wellness esteems, between group separate and intra-group remove with other developmental and swarm enhancement based methodologies. A thorough experimentation is reproduced and test result uncovers that the proposed crossover approach is performing better when contrasted with different methodologies.

2.16 Title: "Clustering performance comparison of new generation meta-heuristic algorithms"

Author: Fatma Turna & Lale Ozbakir

Year of Publications: May 2017

Publishing Details: Elsevier Journal

[16]This article tended to two new age meta-heuristic calculations that are acquainted with the writing as of late. These calculations, demonstrated their execution by benchmark standard test capacities, are actualized to take care of grouping issues. One of these calculations called Ions Motion Optimization and it is set up from the movements of

particles in nature. The other calculation is Weighted Superposition Attraction and it is predicated on two essential principals, which are "pulled in developments of specialists" and "superposition". Both of the calculations are connected to various benchmark informational collections comprised of ceaseless, downright and blended factors, and their exhibitions are contrasted with Particle Swarm Optimization also, Artificial Bee Colony calculations. To kill the infeasible arrangements, Deb's standard is incorporated into the calculations. The examination results demonstrated that both of the calculations, Ions Motion Optimization and Weighted Superposition Attraction, are focused arrangement approaches for grouping issues.

2.17 Title: "A Literature Survey: Meta-heuristic Algorithms in Car Engine Design"

Author: Hongming Xu, Xin Yao ,Mohammad-H & Tayarani-N

Year of Publications: September 2014

Publishing Details: IEEE Explore

[17]Meta-heuristic calculations are frequently enlivened by characteristic wonders, including the development of species in Darwinian common choice hypothesis, subterranean insect practices in science, rush practices of a few winged animals, tempering in metallurgy, and so forth. Because of their incredible potential in taking care of hard streamlining issues, meta-heuristic calculations have discovered their routes into car motor plan. There are diverse advancement issues emerging in various territories of vehicle motor administration including alignment, control framework, blame finding and demonstrating. The survey covers an extensive variety of research, including the use of meta-heuristic calculations in motor adjustment, advancing motor control frameworks, motor blame conclusion, enhancing diverse parts of motors and demonstrating.

2.18 Title: "Rough set based meta-heuristic clustering approach for the social e-learning systems"

Author: Aboul Alla Hassanien and S. Selva Kumar & Ahmad Taher Azar and H. Hannah Inbarani

Year of Publications: April 2015

[18]A basic test of Web 2 is the manner in which that an unfathomable proportion of data has been instigated over a short time. Labels are by and large used to burrow and organize the Web 2.0 assets. Grouping the label data is uncommonly inauspicious since the label space is noteworthy in a couple of social labelling locales. Label bunching is the strategy for gathering the relative labels into gatherings. The label grouping is really useful for seeking and masterminding the Web 2.0 assets besides fundamental for the accomplishment of social labelling frameworks. We proposed a half and half resilience unpleasant set-based molecule swarm streamlining for bunching labels. At that organize, the proposed system is stood out from benchmark bunching calculation k-implies with molecule swarm improvement (PSO)- based gathering strategy. The exploratory examination speaks to the character of the recommended approach. The label bunching issue is a genuine essential issue and has pulled in much consideration of numerous specialists. This paper has proposed another crossover calculation for tackling the group initialization issue which depends on the blend of TRS and meta-heuristic bunching calculations.

2.19 Title: "An insight into the Classification with imbalanced data: the Empirical Results and current trends on using data intrinsic characteristics"

Author: Francisco Herrera ,Alberto Fernandez, Victoria Lopez Vasile Palade & Salvador Garcia

Year of Publications: July 2013

Publishing Details: Elsevier Journal

[19]Preparing classifiers with datasets which endure of imbalanced class conveyances is an essential issue in information mining. This issue happens when the quantity of precedents speaking to the class of intrigue is much lower than the ones of alternate classes. Its essence in some genuine applications has brought along a development of consideration from specialists. We in the blink of an eye survey the numerous issues in machine learning and uses of this issue, by presenting the qualities of the imbalanced dataset situation in characterization, showing the particular measurements for assessing

execution in class imbalanced learning and listing the proposed arrangements. Specifically, we will depict pre handling, cost delicate learning and outfit strategies, doing a test concentrate to differentiate these methodologies in an intra and between family examination. We will do a careful discourse on the fundamental issues identified with utilizing information natural qualities in this arrangement issue. This will enhance the current models as for: the nearness of little disjuncts, the absence of thickness in the preparation information, the covering between classes, the distinguishing proof of uproarious information, the hugeness of the marginal examples, and the dataset move between the preparation and the test appropriations. At last, we acquaint a few methodologies and suggestions with location these issues related to imbalanced information, and we will demonstrate some trial models on the conduct of the learning calculations on information with such inborn attributes.

2.20 Title: "The entropy weighting K Mean algorithm for the subspace clustering of HD Sparse Data"

Author: Joshua Zhexue Huang, Liping Jing and Michael K. Ng

Year of Publications: June 2007

Publishing Details: IEEE Explore

[20]Clusters of objects are often found or exist in subspace rather than in entire space when we are working in high-dimensional data. A situation may occur in which we need to identify clusters of similar kind of objects where the similarity is defined according to the subset of the attributes like, for instance, we have a text document and the clusters of the document are made on the basis of different subsets of keywords and terms; which can cause a data sparsity problem. This paper presented a modified version of K-Means algorithm by adding a new equation which computes the weights of all dimension in each cluster. After the experiments, this paper concluded that the new K-Means algorithm is able to compute better results than other subspace clustering algorithms by simultaneously minimizing the within cluster dispersion an maximizing the negative weight entropy in the process of clustering. This paper introduces another k-implies type calculation for grouping high-dimensional protests in sub-spaces. In high-dimensional information, bunches of items regularly exist in subspaces as opposed to in the whole

space. For instance, in content grouping, bunches of reports of various subjects are classified by various subsets of terms or catchphrases. The catchphrases for one bunch may not happen in the reports of different groups. This is an information scarcity issue looked in grouping high-dimensional information. In the new calculation, we broaden the k-implies grouping procedure to figure a weight for each measurement in each bunch and utilize the weight esteems to recognize the subsets of critical measurements that arrange distinctive groups. This is accomplished by incorporating the weight entropy in the target work that is limited in the k-implies grouping process. An extra advance is added to the k-implies grouping procedure to consequently register the weights of all measurements in each bunch. The examinations on both engineered and genuine information have demonstrated that the new calculation can create preferable grouping results over other subspace bunching calculations. The new calculation is likewise versatile to substantial informational indexes.

2.21 Title: "A hybrid data clustering approach based on improved cat swarm optimization and *K*-harmonic mean algorithm"

Author: kumar, Sahoo & Yugal

Year of Publications: June 2015

Publishing Details: IOS Press

Kumar and Sahoo[21], proposed a half and half methodology for taking care of information bunching issue. The proposed methodology is involving Improved feline swarm streamlining and K-Harmonic methods calculations. In this work two engineered and five genuine informational collections are considered to register the test and assess the execution of the proposed calculation. It is seen that the proposed calculation can keep the KHM calculation from nearby optima and improves the union speed of the Cat swarm streamlining calculation. the multiobjective molecule swarm enhancement calculation for partitional grouping. The target of the proposed work is to gives all around isolated, associated and smaller bunches. In this work twenty-seven informational collections are considered to figure the execution of the proposed work. It is assessed from the exploratory outcomes that the proposed calculation is powerful, productive and give increasingly ideal groups when contrasted with different calculations

2.22 Title: “Data Clustering Using Variants of Rapid Centroid Estimation”

Author: Yuwonu, Ngyugen, Su, Moulton

Year of Publications: September 2013

Publishing Details: IEEE

Yuwono et al. [22], presented a quick centroid estimation calculation for information clustering.The point of the proposed calculation is to diminish the computational multifaceted nature and rearranges update guidelines of PSC. In this work seven informational indexes are considered to register the trial and assess execution. It is seen that the proposed calculation is a lot quicker and precise and upgrades the grouping quality when contrasted with different calculations. The proposed calculation is to streamline the huge scale information bunching. In this work four manufactured and one genuine informational collections are considered to register the analysis and assess the execution of the proposed calculation. It is assessed that the proposed calculation can manage expansive scale information and nature of bunching is likewise kept up at that dimension.

CHAPTER-3

SYSTEM DEVELOPMENT

3.1 Flowcharts

In this section of project report, all the basic steps to design an algorithm is illustrated.

First, as shown in figure-3.1, we start by initializing the initial cluster centers. Now, we make a call to 'classfit' function. This function will return the class variables and fitness of data points (named as fitness1) according to cluster centre. Now, the call to next function, that is, 'accusum' function is made which will return the accuracy of the cluster centers and intra-cluster distance. Every cluster centre is now propagated to new position according to the 'propagation equation'. Now, by using the 'classfit' function, fitness for these new cluster centers will be retrieved (named fitness2). For every fitness2 that is less than fitness1, refract the particle according to the 'refraction equation' and assign it to the corresponding cluster centre. If fitness2 is greater than fitness1, then we will check whether the fitness2 is greater than Best_fit (best fitness that has been achieved till now) or not, and, if it is greater than the Best_fit, then break the Best Particle according to the 'breaking equation'. No matter what is it value, assign the corresponding new cluster centre to old cluster centre. This will go in loop till all the waves don't cover. At last, plot the graph between wave (x-axis) and Intra-cluster Distance (y-axis).

The figure-3.2 shows the flowchart of the 'classfit' function. This function is called by main function. Whenever it is called, first, it identifies the objective function (which is just the square root of sum of distance between data points and cluster centres). Then the objective function is sorted and column number sorting is also done. Now, all the distinct class values are stored in array named class. Now from column number sorting, individual correctness is calculated for distinct class variables and that correctness is named as fitness of function or cluster centre. Now it will return array class and fitness to main function.

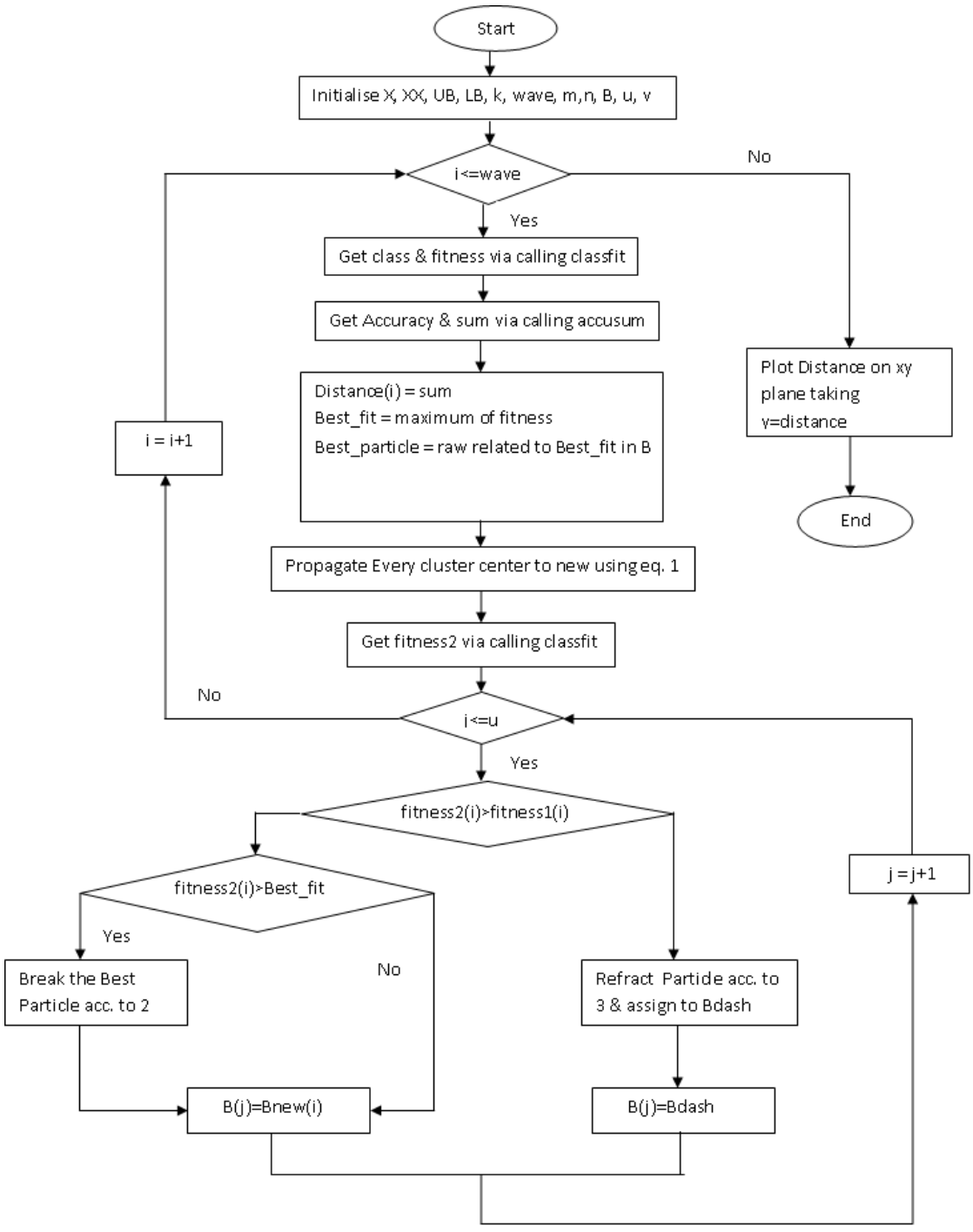


Figure-3.1. The flowchart of the Water wave optimization algorithm

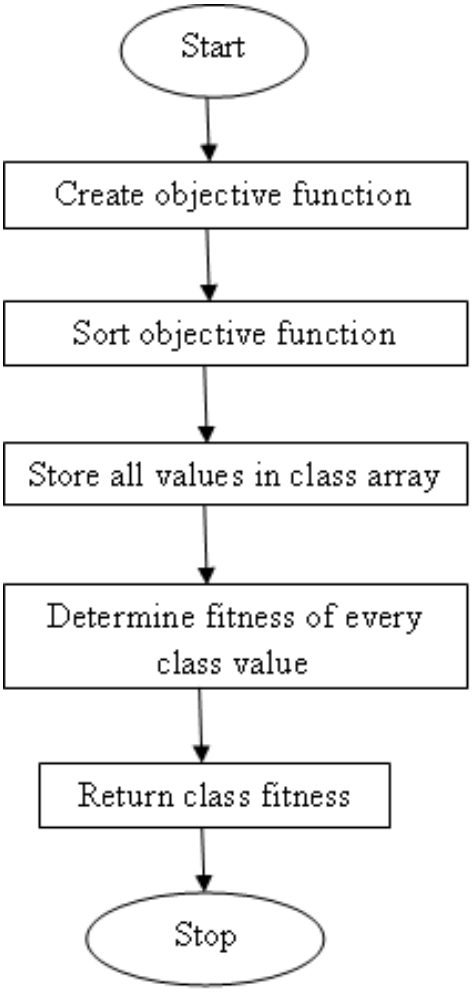


Figure-3.2. The flowchart of the 'classfit' function

The figure-3.3 shows the flowchart of the 'accsum' function. Whenever this function is called by main function, first it identifies the 'accuracy' of the algorithm by calculating the correct and incorrect assigned class. Then for every individual cluster, it finds the inter-cluster distance between data points in that cluster and cluster centre of that cluster. Now sum of all the individual inter-cluster distance of different cluster is done and assigned in sum. Now it will return the Accuracy and Sum to main function.

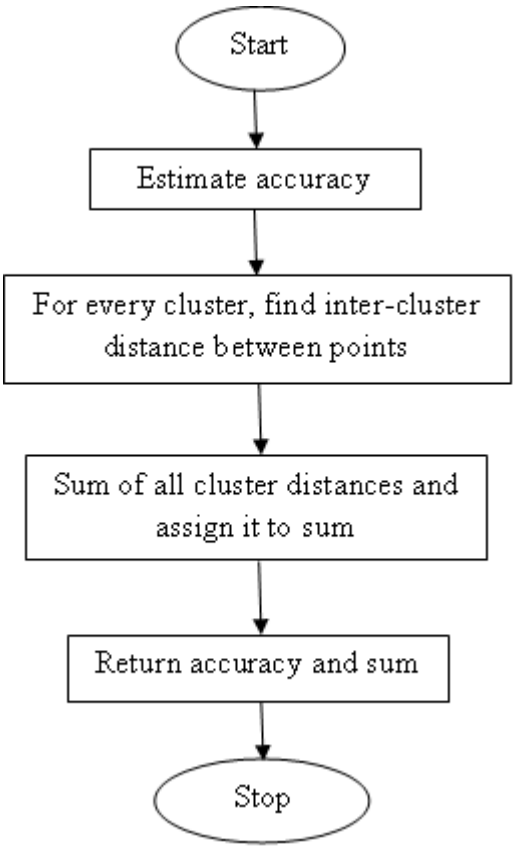


Figure-3.3. The flowchart of the ‘accsum’ function

Figure--3.4, shows the flow chart for VPS algorithm for Heath Care Data

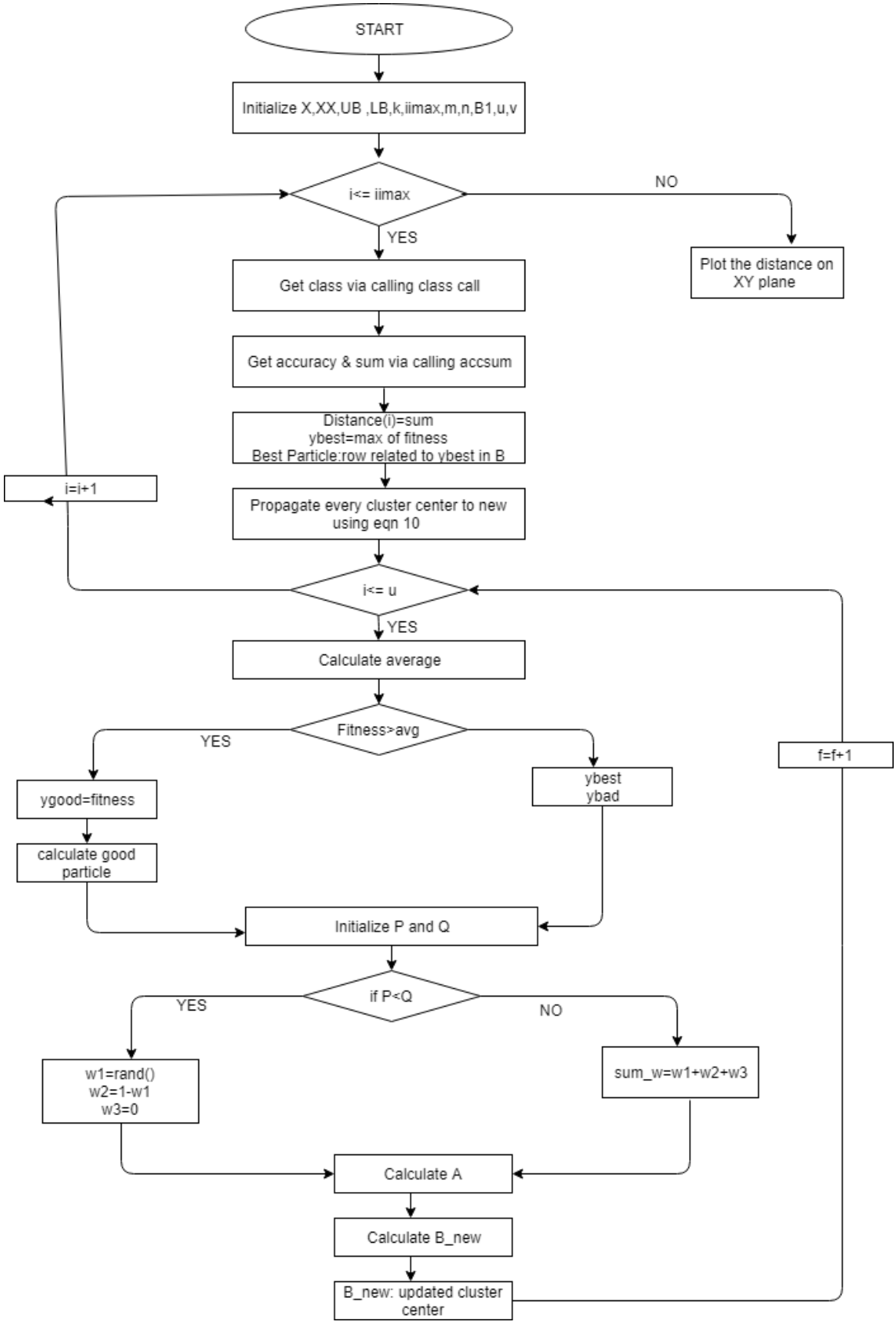


Figure-3.4. The flowchart of the Vibrating Particle System algorithm

3.2 Tools Used

MATLAB

MATLAB stands for laboratory matrix. It is a high - performance language that is mainly used for technical computing purposes. MATLAB includes viewing, computing and programming. MATLAB provides an interactive and user - friendly environment in which problems and solutions to these challenges are expressed in the form of familiar mathematical notation. MATLAB is applied across fields ,typical applications of MATLAB include:

- Math and computation
- Development of algorithm
- Simulation & prototyping
- Data analysis, exploration& visualization
- Scientific and engineering graphics

It's an interactive system. The basic data element in MATLAB is nothing but an array in which dimensions are not necessary This property allows us to solve a large number of technical computing problems within a part of the time it would take to write a program in a scalar, non - interactive language such as Fortran .With the input of many users, MATLAB has developed over the years. It is the classic instructional instrument for mathematics courses, engineering in university surroundings .In the industry, MATLAB is the instrument of choice for research, advancement and analysis of high productivity.

It comprises a set of application-specific solutions called toolboxes. Toolboxes are extremely useful for MATLAB users as it enables them to learn and relate specific technology. Toolboxes are nothing but complete collections of MATLAB functionalities that are extended and incorporated into the MATLAB environment in order to resolve specific class of problems.

3.3 Algorithm

3.3.1 Water Wave Optimization (WWO)

The algorithm used is Water Wave Optimization(WWO) which uses the basic three operations of wave, that is, propagation, refraction, and breaking; that are used in the algorithm, which are explained below.

i. Propagation:

[1]Once the waves get generated, each wave should be propagated exactly once. A new wave by name of **x′** is created by the propagation operator. This is done by shifting our original wave **x** by dimension **d**.

$$x'(d) = x(d) + rand(-1,1) \cdot \lambda L(d) \qquad (1)$$

where rand(-1,1) is a uniformly distributed random number within the range [-1,1], and L(d) represents the length of the d^{th} dimension of the search space.

After the propagation operation is applied on wave, the fitness of the wave **x′** is calculated**.** If the result shows **f (x′) > f(x)**, then **x** is replaced by **x′** in the population.

ii. Refraction:

[1]Using the shallow water wave theory, it is mentioned that water waves travels fastest in deep medium. Thus, the water waves slow down as they pass from deep water into shallow water.

Thus, after refraction, the position of wave gets changed due to change in the medium and the speed, and a simple way to calculate the new position is**:**

$$x'(d)=N([x^*(d) + x(d)] / 2 , [|x^*(d) - x(d)|] / 2) \qquad (2)$$

where **x*** is the best solution founded so far, and **N(μ,σ)** is a Gaussian random number with mean μ and standard deviation σ.

Thus we can say that the new position **x′** is the random number which is centreed halfway between the recently calculated best position and the initial original position, and the calculated standard deviation is equal to the absolute value of their difference.

iii. Breaking:

[1]As the waves reach the shore, they break into a train of solitary waves. While using the WWO algorithm, the wave operation, i.e. breaking is applied only on a wave x that finds a new best position (i.e., **x** becomes the new **x***). It conduct a local search around **x*** to find the waves that replicate **x*** after the wave breaking. Mathematically, a solitary wave **x′** is generated at each dimension **d** as**:**

$$x'(d)=x(d)+N(0,1)\cdot\beta L(d) \qquad (3)$$

where β is the coefficient of breaking. Now the solitary waves are compared with the **x***. If the solitary waves are better than the **x***, **x*** gets replaced by the wave which is the fittest one among the train of solitary waves.

The Pseudo code for the Water Wave Optimization Clustering algorithm is provided below:

Algorithm 1.
1. Randomly initialize a population *P* of *n* waves (solutions);
2. while stop criterion is not satisfied do
3. for each x ∈ P do
4. Propagate x to a new x′ based on Eq. (1);
5. if f(x′)>f(x)then
6. if f(x′)>f(x*)then

7.	Use Eq. (3) to break x′;
8.	x* is updated with x′;
9.	x is replaced with x′;
10.	else
11.	Use Eq. (2) to refract x to a new x′;
12.	return x*

Fig 3.1 WWO Algorithm (Pseudo-code)

3.3.2 Vibrating Particle System (VPS)

[2]The algorithm that has been used to compare with WWO is Vibrating Particle System (VPS) that is a new population - based meta-heuristic algorithm based on a system's damped free vibration from single degree of freedom.

$$D = \left(\frac{iter}{iter_{max}} \right) - \infty$$

In the above equation iteration is the current iteration number that is being used ,$iter_{max}$ is the maximum number of iterations that are used and α is a constant used .

[2]The equation which is used for updating the positions is given as follows:

$$x_i^j = w_{1.}\left[D.A.rand1 + HB^j\right] + w_{2.}\left[D.A.rand2 + GP^j\right] + w_{3.}\left[D.A.rand3 + BP^j\right]$$

[2]To calculate x_i^j the equation given below is used:

$$A = \left[w_{1.}\left(HB^j - x^j{}_i\right)\right] + \left[w_{2.}\left(GP^j - x^j{}_i\right)\right] + \left[w_{3.}\left(BP^j - x^j{}_i\right)\right]$$

$$w_1 + w_2 + w_3 = 1$$

The Pseudo code is provided below:

	Procedure Vibrating Particle System(VPS)
1.	Initialize algorithm parameters
	Initial positions are created randomly
2.	The initial value of the objective function is evaluated and HB is stored
3.	While maximum iterations are not fulfilled
	For each particle
	The GP and BP are chosen
	If P < rand
	$w_3=0$ and $w_2=1-w_1$
	End if
	For each component
	Now location is obtained by Eq 10
	End for
4.	Violated components are regenerated by harmony search based handling approach
	End for
	The value of the objective function is calculated and HB is updated
5.	End while
	End procedure

Fig 3.2 VPS Algorithm (Pseudo-code)

3.4 Test Plan

In this chapter we discuss about the different datasets we are using on which we implement our algorithm to find the optimized cluster centres .Data sets used are stated below:

1. BCW

2. WDBC

3. Heart

4. BUPA

5. Diabetes

6. Thyroid

3.4.1 Data Sets

Describing the datasets in detail:

1. BCW

Information about the dataset:

The samples recorded in the dataset are received periodically as Dr.Wolberg reports his clinical cases. The +dataset was provided in the year 1992 on July 15th. Below we provide the chronological order in which the data samples were received.

Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)
Total: 699 points

Data Set Characteristics	Multivariate
Attribute Characteristics	Integer
Associated Tasks	Classification
Number of Instances	699
Number of Attributes	10
Missing Values	Yes
Area	Life
Date Donated	1992-07-15
Number of Web Hits	423279
Attributes	Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion ,Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, Class (2 for benign,4 for malignant)

Fig 3.3 Information of BCW Dataset

2. WDBC

Information about the dataset:

By the use of digitized image of a fine needle aspirate (FNA) of breast mass features are computed .The characteristics of the cell nuclei are described which are present in the image.

Data Set Characteristics	Multivariate
Attribute Characteristics	Real
Associated Tasks	Classification
Number of Instances	569
Number of Attributes	30
Missing Values	No
Area	Life
Date Donated	1995-11-01
Number of Web Hits	809194
Attributes	Id Number ,Diagnosis(0-malignant,1-benign)Ten real valued features are computed for each cell nucleus such as radius ,texture, perimeter, area, symmetry etc.

Fig 3.4 Information of WDBC Dataset

3. HEART

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Real
Associated Tasks	Classification
Number of Instances	270
Number of Attributes	13
Missing Values	No
Area	Life
Date Donated	N/A
Number of Web Hits	162363
Attributes	Age, Sex, Chest pain type, Resting blood Pressure, serum cholesterol in mg/dl , fasting blood sugar > 120 mg/dl , resting electrocardiographic results (values 0,1,2) , maximum heart rate achieved , exercise induced angina , old peak = ST depression induced by exercise relative to rest ,. the slope of the peak exercise ST segment , number of major vessels (0-3) colored by flourosopy , thal : 3 = normal; 6 = fixed defect; 7 = reversible defect

Fig 3.5 Information of Heart Dataset

4. BUPA

Information about the dataset:

The samples in the dataset provided constitutes each record of a single male individual. The initial 5 variables correspond to the test results given by blood samples and thought to be responsive to disorders of the liver that might be a consequence of consumption of alcohol in excessive amounts.

Data Set Characteristics	Multivariate
Attribute Characteristics	Integer, Categorical, Real
Associated Tasks	N/A
Number of Instances	345

Number of Attributes	7
Missing Values	No
Area	Life
Date Donated	1990-05-15
Number of Web Hits	136334
Attributes	mcv mean corpuscular volume, alkphos alkaline phosphotase,sgpt alanine aminotransferase,sgot aspartate aminotransferase,gammagt gamma-glutamyl transpeptidase,drinks number of half-pint equivalents of alcoholic beverages drunk per day,selector field created by the BUPA researchers to split the data into train sets

Fig 3.6 Information of Bupa Dataset

5. DIABETES

Data Set Characteristics	Multivariate
Attribute Characteristics	Integer
Associated Tasks	Classification
Number of Instances	768
Number of Attributes	8
`Missing Values	No
Area	Life
Date Donated	N/A
Number of Web Hits	373254
Attributes	Number of times pregnant, Plasma glucose concentration, Blood Pressure, Triceps skin fold thickness, Serum insulin, Body Mass Index, Diabetes pedigree function, Age

Fig 3.7 Information of Diabetes Dataset

6. THYROID

Information about the dataset:A total of 10 different datasets were provided by Gravan Institute out of which one of them is used here.The given dataset is provided by Stefan Aeberhard.

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Real
Associated Tasks	N/A
Number of Instances	215
Number of Attributes	5
Missing Values	No
Area	Life
Date Donated	1987-01-01
Number of Web Hits	165156

Fig 3.8 Information of Thyroid Dataset

3.4.2 Metrics

1. Accuracy Matrix

   Accuracy matrix is just one row matrix with columns equal to number of iterations .This matrix show about how much our cluster centres that are predicted are right along with how much the class variables assigned by algorithm are correctly assigned by comparing with the class file that we already have .Generally ,the matrix shows increasing trend because in every new iteration ,the cluster centres changes then accordingly class variables assigned changes and are placed at nearly right positions where they should go which leads to increase in accuracy.

2. Cluster Centre Matrix

   A cluster matrix shows the centres of the data that it will achieve after the possible number of iterations. Cluster centre is a point or value in matrix that is assumed to

be centre of the data points of similar class. In above data set we will have three rows in cluster matrix as we have 3 options for the class. In every iteration new cluster centres are obtained using propagation equation .Then old cluster centres are updated to new cluster centres accordingly.

3. Distance Matrix:

Distance matrix is also one row matrix with column equal to number of iterations. This matrix shows the intra cluster distance that is how far away are data points and the cluster centres .As we have calculated the cluster centres, now for every data point distance between cluster centre and that point is calculated for every cluster centre using 'root mean square' method.Then sum of all the values is assigned in Distance Matrix,This matrix shows trends opposite to accuracy matrix that is shows decreasing trends.

3.4.3 Test Setup

This testing is not like other where we will give testing data and check if this shows the correct answer or no. Hence after the algorithm is designed ,we plot the graph between intra cluster distance and the number of iterations .If somewhat it shows the trend of decreasing that means the algorithm designed is working for data set.To check further ,we can get more datasets to see if their graph is also what they need to be.If they are ,then we can say that the algorithm designed is working properly.

CHAPTER-4

RESULTS AND PERFORMANCE ANALYSIS

After implementing the algorithm used on the various healthcare datasets the following results have been observed. To check the working and correctness of the Vibrating Particle System Algorithm we have implemented our algorithm on six different healthcare datasets and we have obtained the graphs between intra cluster distance and number of iterations used. Further we have compared the performance of the VPS clustering algorithm with that of the Water Wave Optimization clustering algorithm by implementing the WWO algorithm on the same six healthcare datasets so as to draw a clear cut comparison between the performances of these two algorithms. Accuracy obtained for each algorithm has been mentioned along with the graph to clearly indicate which of the two algorithms performs better for each of the healthcare dataset.

Given below are the results obtained for both the algorithms (viz. VPS and WWO) when they were implemented on the different healthcare datasets

1. Thyroid

   When we ran the algorithm on the data set of Thyroid Disease dataset the following results were obtained

   Fig 4.1 The accuracy matrix

		85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
1	574	70.2326	71.1628	71.1628	71.1628	71.1628	72.0930	73.0233	73.0233	73.0233	73.9535	74.4186	76.2791	77.6744	78.1395	78.6047	80.4667

   Fig 4.2 The distance matrix obtained is:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	855.3628	822.2917	779.1816	716.4102	716.3734	710.1892	690.6768	678.3515	672.7942	670.9419	653.1444	642.8594	640.4628	638.9441	638.2429	635.1869

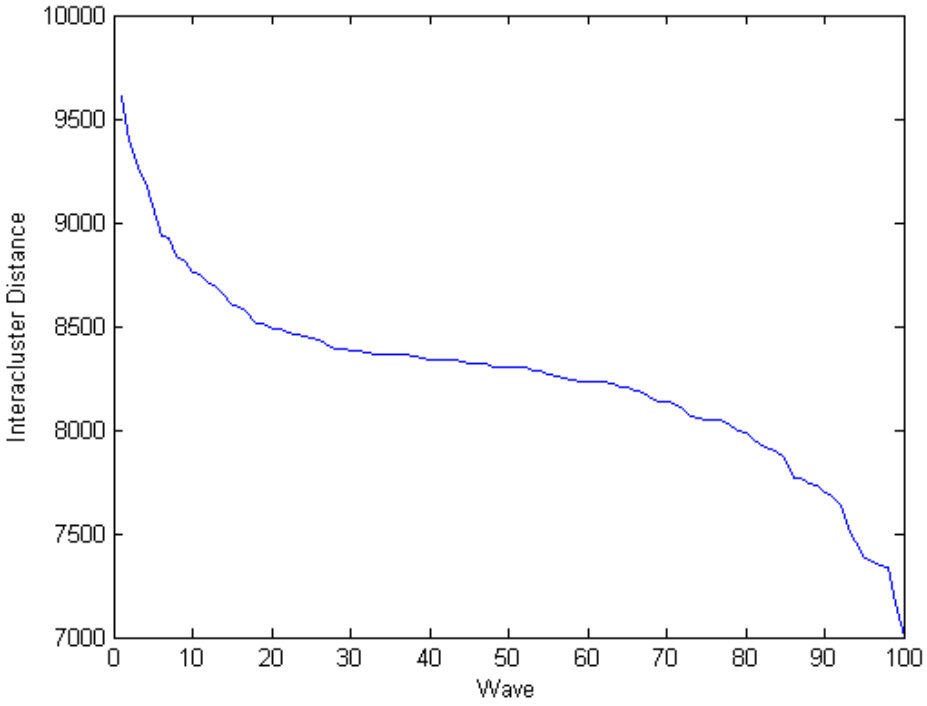


Fig 4.1 Graph obtained for WWO (Thyroid dataset)

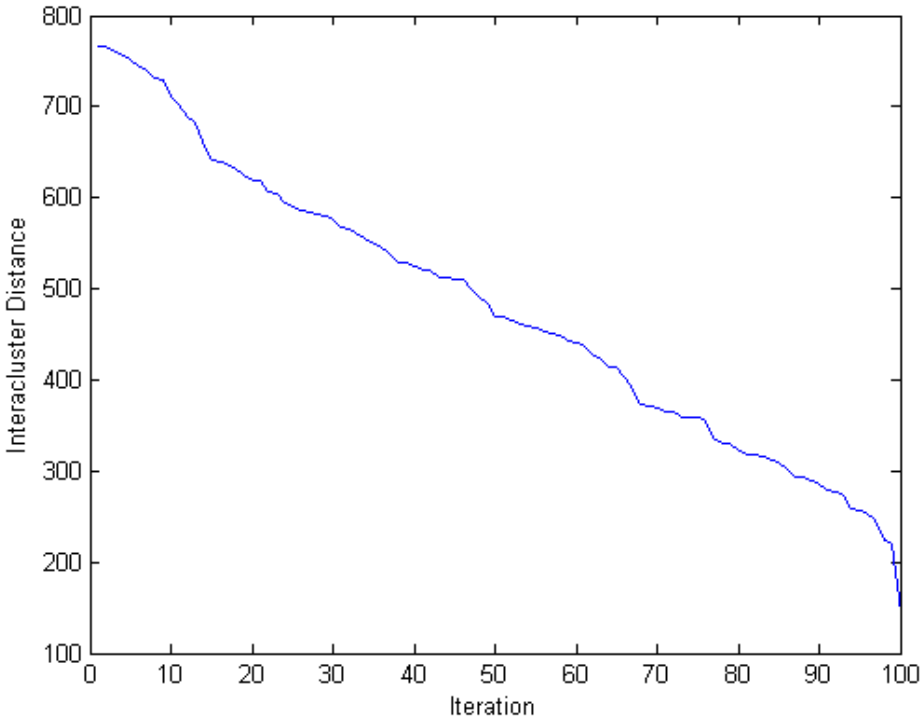


Fig 4.2 Graph obtained for VPS (Thyroid dataset)

After implementing the both the algorithms on Thyroid disease dataset for a maximum of 100 iterations we obtain the following results.

The accuracy obtained is as follows:

For VPS: 80.46

For WWO:82.79

We can see clearly in graph, in Figure-4.1 and Figure-4.2 clearly that the inter cluster distance is decreasing with an increase in the number of iterations and we can note from the graph that first there is a sudden decrease in the intra cluster distance due to the fact that initially we chose random cluster centre so the distance is considerably high but then gradually the cluster centres are coming to correct position so the inter cluster distance is subsequently decreasing but after certain number of iterations the decrease in the distance also slows down because there is only a little shift in cluster centre which lead to low decreasing in distance. Also as discussed earlier , we can see that with every iteration the value of accuracy is also increasing somehow thereby showing the increasing trend. Further we can clearly note the decreasing trend of values of intra-cluster distance between them.

that the inter cluster distance is decreasing with an increase in the number of iterations/waves and we can also see that the graph first shows a sudden decrease due to the fact that at first we chose random cluster centre so distance is high but then cluster centre are coming to correct position so is decreasing but after certain iteration decreasing process also came to low as there is only a little shift in cluster centre which lead to low decreasing in distance. Also as we have discussed, we can see that in every iteration value of accuracy is nearly increasing somehow showing the increasing trend. Further we can clearly note the decreasing trend of values of intra-cluster distance between them as we have discussed so far.

Also as we can clearly note from the accuracy obtained in the case of the Vibrating Particle System Clustering algorithm came out to be 80.46% while the accuracy obtained in the case of Water Wave Optimization clustering algorithm is 82.79% which clearly indicates that the performance of the WWO algorithm is slightly better as compared to the performance of the VPS algorithm for this particular dataset though the difference in accuracy is not quite significant as it stands at only 2.33%.

2. BCW

When we ran the algorithm on the data set of BCW Disease dataset the following results were obtained

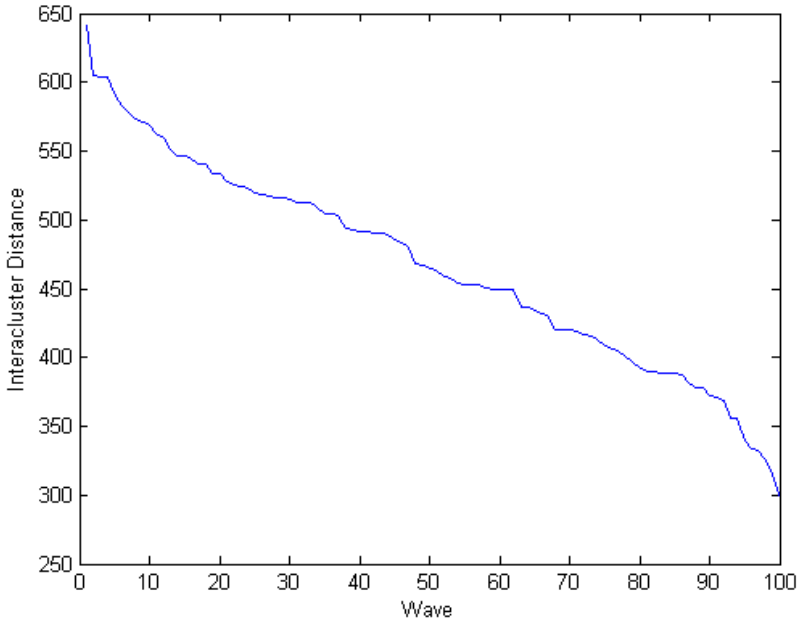


Fig 4.3 Graph obtained for WWO (BCW dataset)

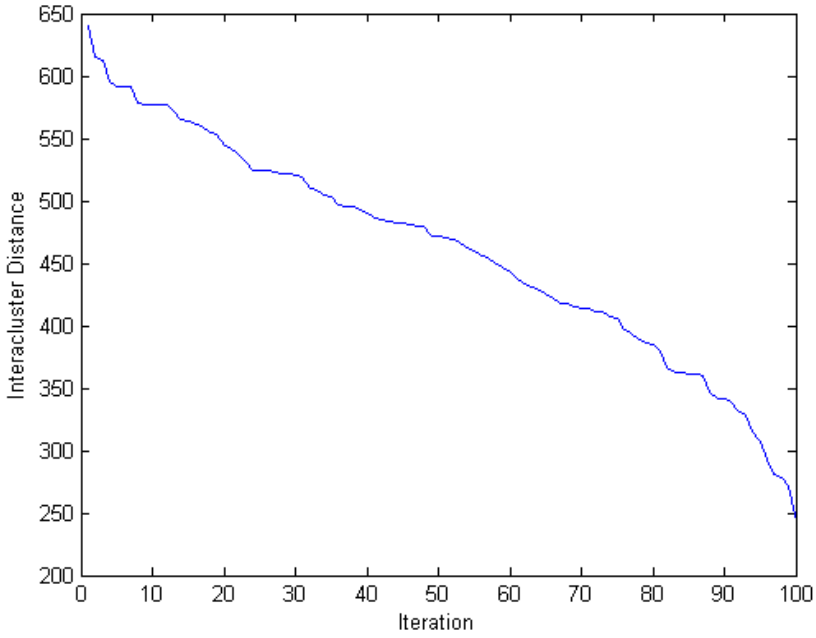


Fig 4.4 Graph obtained for VPS (BCW dataset)

After implementing the both the algorithms on the BCW disease dataset for a maximum of 100 iterations we obtain the following results.

The accuracy obtained is as follows:

For VPS: 92.56

For WWO: 81.13

We can see clearly in graph, shown in figure-4.3 and figure-4.4, that the inter cluster distance is decreasing with an increase in the number of iterations. Also as we can clearly note from the accuracy obtained in the case of the Vibrating Particle System Clustering algorithm came out to be 92.56% while the accuracy obtained in the case of Water Wave Optimization clustering algorithm is 81.13% which clearly indicates that the performance of the VPS algorithm is comparatively much better as compared to the performance of the WWO algorithm for this particular dataset.\

3. WDBC

When we ran the algorithm on the data set of WDBC Disease dataset the following results were obtained

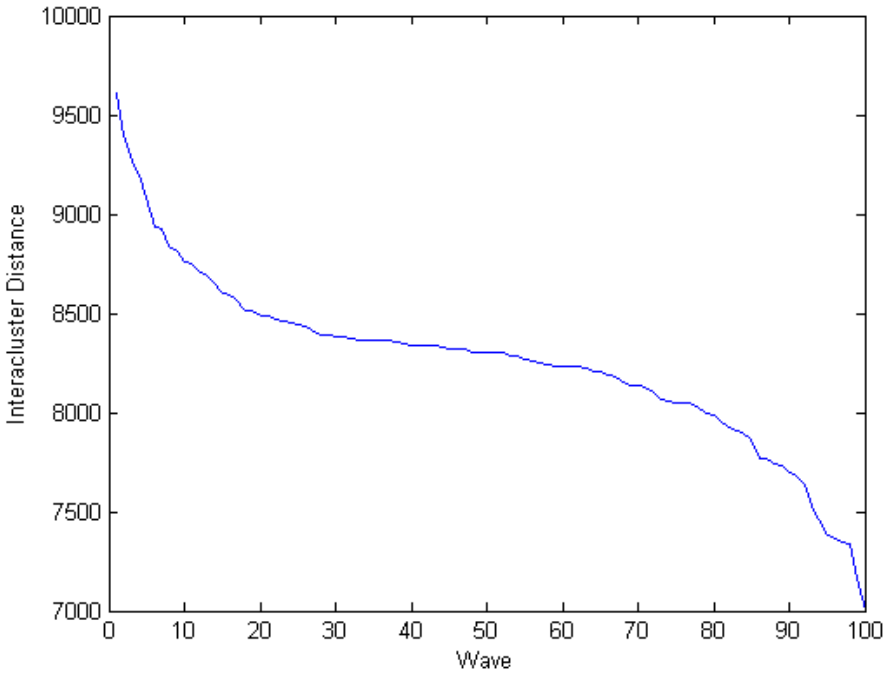


Fig 4.5 Graph obtained for WWO (WDBC dataset)

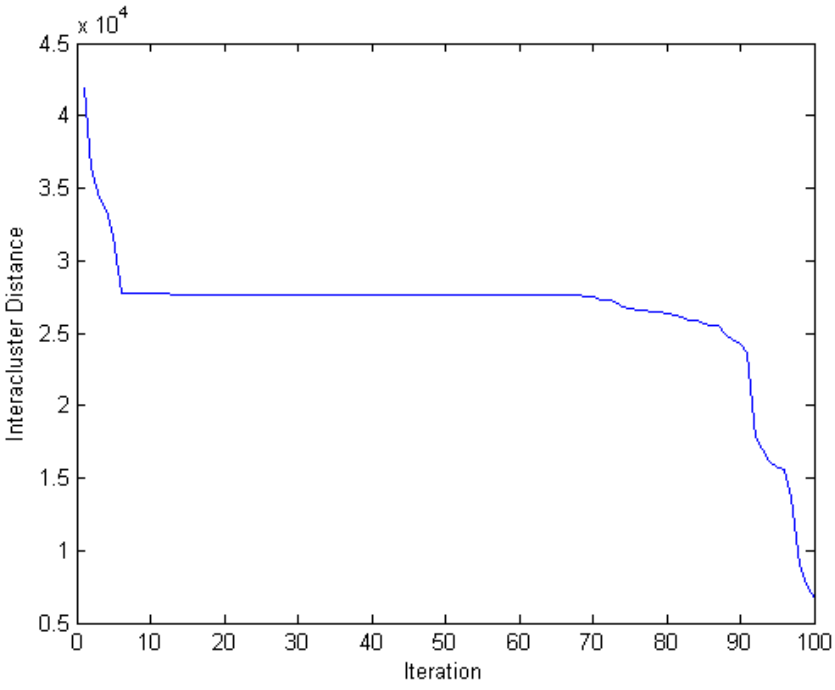


Fig 4.6 Graph obtained for VPS (WDBC dataset)

After implementing the both the algorithms on the BCW disease dataset for a maximum of 100 iterations we obtain the following results.

The accuracy obtained is as follows:

For VPS: 79.26

For WWO: 81.19

We can see in graph, shown in figure-4.5 and figure-4.6, that the inter cluster distance is decreasing with an increase in the number of iterations. Also as we can clearly note from the accuracy obtained in the case of the Vibrating Particle System Clustering algorithm came out to be 79.26% while the accuracy obtained in the case of Water Wave Optimization clustering algorithm is 81.19% which clearly indicates that the performance of the WWO algorithm is slightly better as compared to the performance of the VPS algorithm for this particular dataset though the difference in accuracy is not quite significant as it stands at only 1.93%.

4. Heart

When we ran the algorithm on the data set of Heart Disease dataset the following results were obtained:

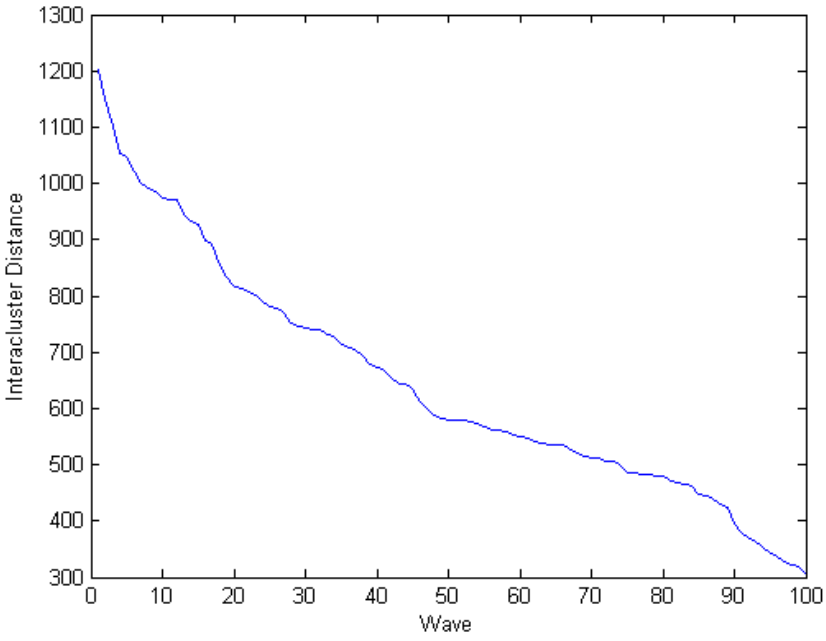


Fig 4.7 Graph obtained for WWO (Heart dataset)

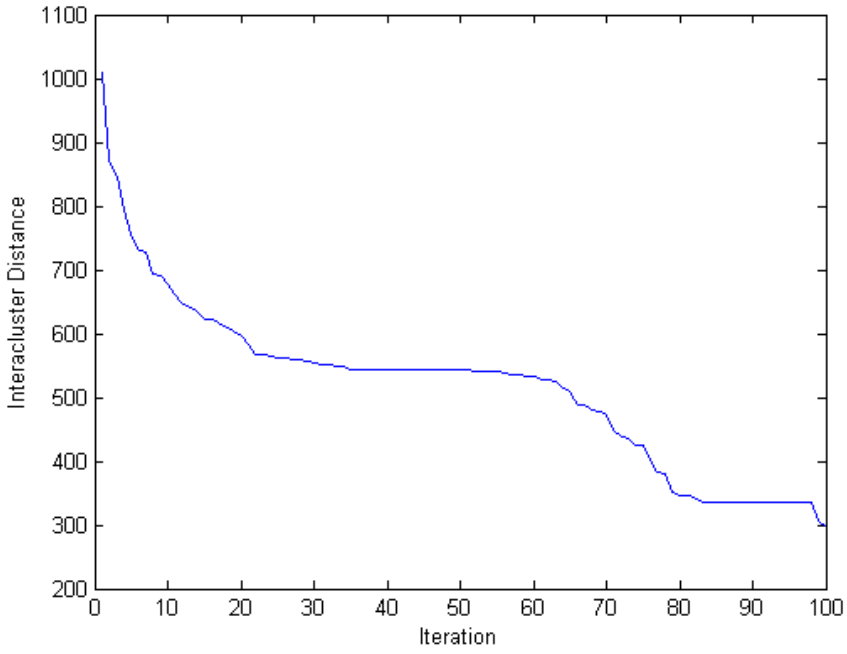


Fig 4.8 Graph obtained for VPS (Heart dataset)

After implementing the both the algorithms on the Heart disease dataset for a maximum of 100 iterations we obtain the following results.

The accuracy obtained is as follows:

For VPS: 62.59

For WWO: 59.63

We can clearly note from the preceding graphs the decreasing trend of values of intra-cluster distance as the number of iterations increases. Also as we can clearly see from the accuracy obtained in the case of the Vibrating Particle System Clustering algorithm came out to be 62.59% while the accuracy obtained in the case of Water Wave Optimization clustering algorithm is 59.63% which clearly indicates that the performance of the VPS algorithm is slightly better as compared to the performance of the WWO algorithm for this particular dataset.

5. Bupa

When we ran the algorithm on the data set of Bupa Disease dataset the following results were obtained:

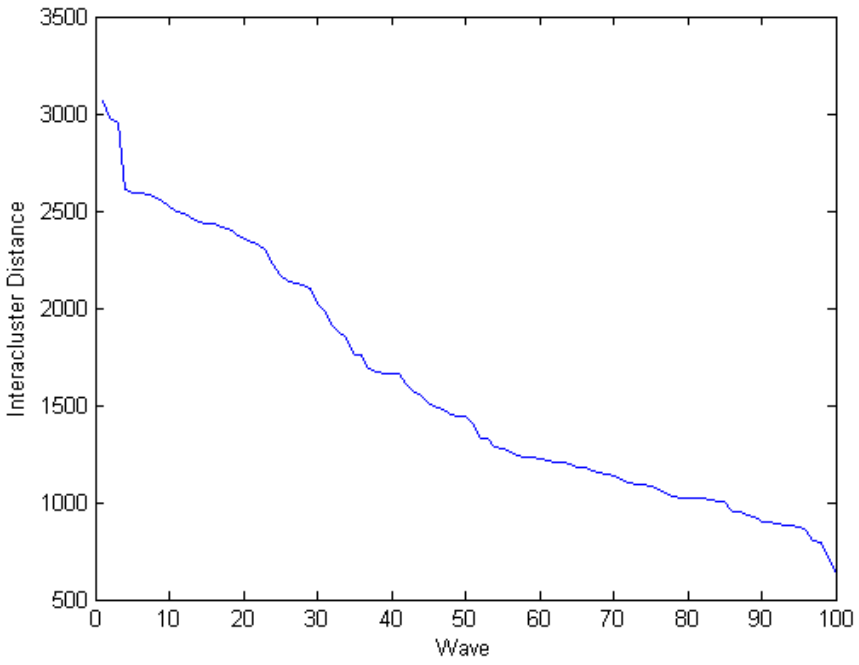


Fig 4.9 Graph obtained for WWO (Bupa dataset)

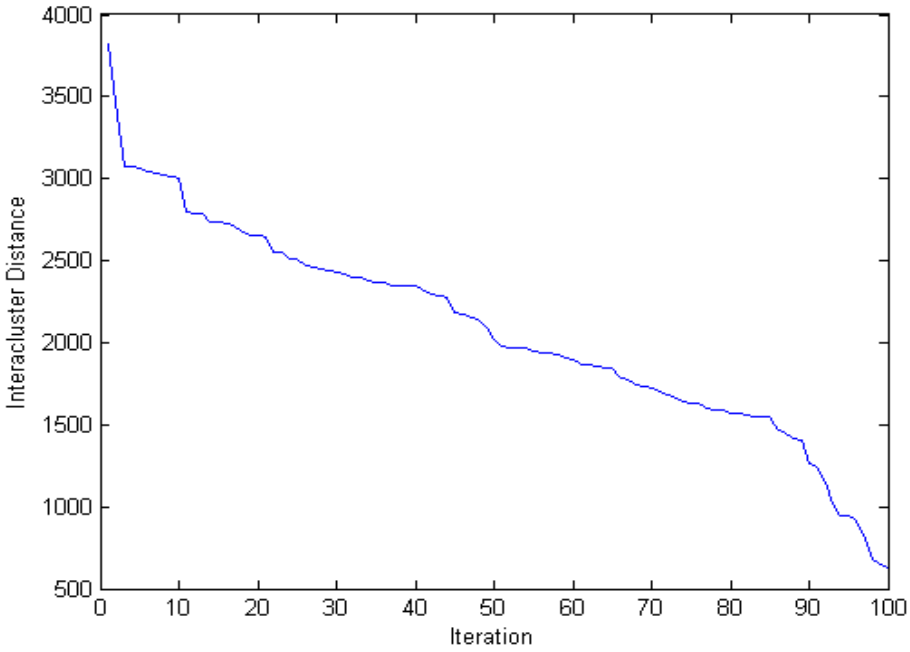


Fig 4.10 Graph obtained for VPS (Bupa dataset)

After implementing the both the algorithms on the Heart disease dataset for a maximum of 100 iterations we obtain the following results.

The accuracy obtained is as follows:

For VPS: 60.57

For WWO: 68.69

As clearly indicated in the in figure-4.9 and figure-4.10, the inter cluster distance is decreasing with an increase in the number of iterations. Also as we can clearly note from the accuracy obtained in the case of the Vibrating Particle System Clustering algorithm came out to be 60.57% while the accuracy obtained in the case of Water Wave Optimization clustering algorithm is 68.69% which clearly indicates that the performance of the WWO algorithm is comparatively much better than the performance of the VPS algorithm for this particular dataset with a difference in accuracy standing at 8.12%.

6. Diabetes

When we ran the algorithm on the data set of Diabetes Disease dataset the following results were obtained:

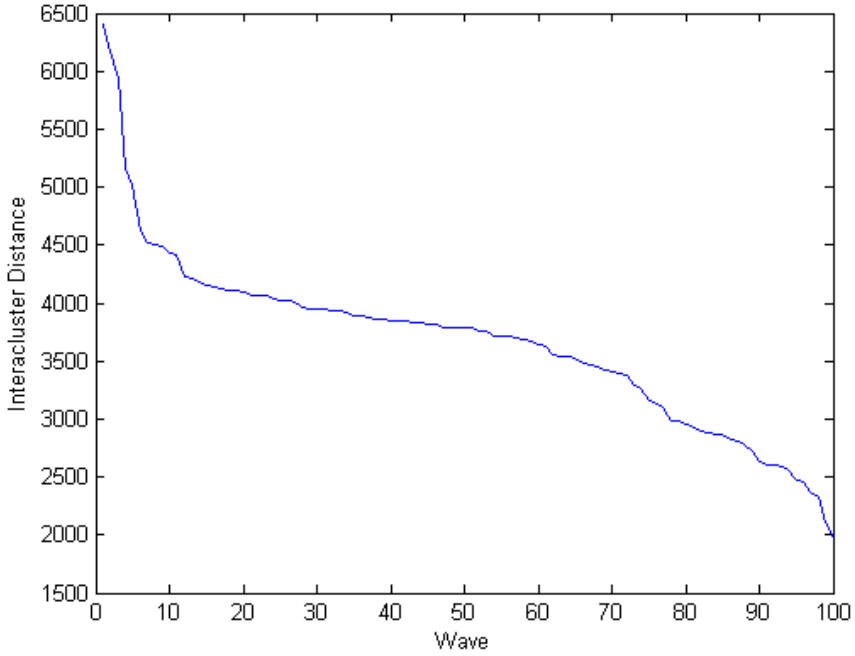


Fig 4.11 Graph obtained for WWO (Diabetes dataset)

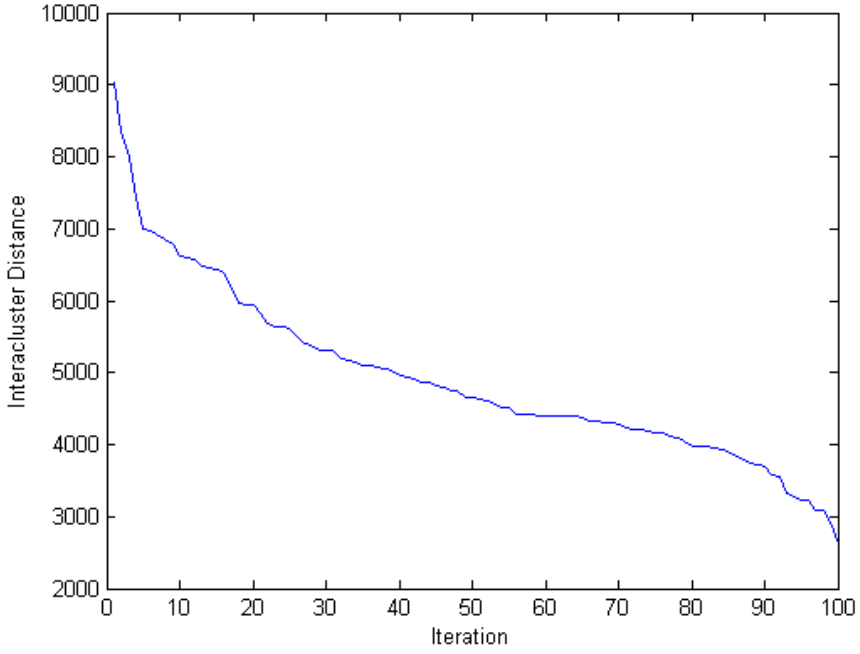


Fig 4.12 Graph obtained for VPS (Diabetes dataset)

After implementing the both the algorithms on the Diabetes disease dataset for a maximum of 100 iterations we obtain the following results.

The accuracy obtained is as follows:

For VPS: 72.91

For WWO: 74.35

Figure-4.11 and figure-4.12 clearly indicate that the inter cluster distance is decreasing with an increase in the number of iterations and we can note from the graph that first there is a sudden decrease in the intra cluster distance due to the fact that initially we chose random cluster centre so the distance is considerably high but then gradually the cluster centres are coming to correct position so the inter cluster distance is subsequently decreasing but after certain number of iterations the decrease in the distance also slows down because there is only a little shift in cluster centre which lead to low decreasing in distance. Also as discussed earlier , we can see that with every iteration the value of accuracy is also increasing somehow thereby showing the increasing trend. Further we can clearly note the decreasing trend of values of intra-cluster distance between them.

Also as we can clearly note from the accuracy obtained in the case of the Vibrating Particle System Clustering algorithm came out to be 72.91% while the accuracy obtained in the case of Water Wave Optimization clustering algorithm is 74.35% which clearly indicates that the performance of the WWO algorithm is slightly better as compared to the performance of the VPS algorithm for this particular dataset though the difference in accuracy is not quite significant as it comes out to be only 1.44%.

CHAPTER-5

CONCLUSION

After implementing both of the algorithms, Water Wave Optimization (WWO) and Vibrating Particle System (VPS) on the six datasets that we had described earlier, we have seen that for some of the datasets VPS is better and for some WWO is better. To predict which is better from both of the algorithms, we have to see how much accurate result the algorithm is providing than the other algorithm.

VPS algorithm provides better accuracy for only two of the datasets and those are Heart and BCW. But if we look at WWO algorithm, it provides better result for other four datasets of WDBC, BUPA, Diabetes and Thyroid.

From the point that we described above we can say that WWO algorithm performs better for the Healthcare Datasets than VPS algorithm.

Future works that can be done using these algorithms are, these can be used for multi-partitioning clustering, prediction of live datasets. One of the main future work in the area of healthcare can be as with these can be used to predict the presence of disease in early stage and if that became possible than curing of the disease will also become easy.

REFERENCES

[1] Yu-Jun Zheng, “A new nature-inspired meta-heuristic - Water Wave Optimization”, Elsevier Journal, March 2015

[2]A.Kaveh and M.IlchiGhazaan,”A new meta-heuristic algorithm: Vibrating particle system”, Scientia Iranica, September 2016.

[3]Neetu Kushwaha , Millie Pant,”Fuzzy magnetic optimization clustering algorithm with its applications to health care.”, Springer, July 2018

[4]Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D.Piatko, Ruth Silverman,Angela Y. Wu. , “An Efficient k-Means Clustering Algorithm: Analysis and implementation.”, July 2002

[5] K. A. Abdul Nazeer, M. P. Sebastian, “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm”, July 2009

[6]T.Velurugan,”Performance based analysis between k-Means and Fuzzy C –Means clustering algorithms for connection oriented telecommunication data”, Elsevier Journal, February 2014.

[7] GengZhang ,Chengchang Zhang , Huayu Zhang,” Improved K-means Algorithm Based on Density Canopy”,Knowledge based System Journal –Elsevier, January 2018.

[8]SatyasaiJagannathNanda ,GanapatiPanda,”A survey on nature inspired metaheuristic algorithms for partitional clustering”,Elsevier Journal,November 2013.

[9]R.J. Kuo, T.C. Lin, F.E. Zulvia, C.Y. Tsai,” A Hybrid Meta-heuristic and Kernel Intuitionistic Fuzzyc-means Algorithm for Cluster Analysis

”,Elsevier Journal,February 2018.

[10]LaleOzbakõr , Fatma Turna,”Clustering performance comparison of new generation meta-heuristic algorithms.”,Elsevier Journal,May 2017.

[11]Paul S. Bradley, Usama M. Fayyad,”Refining Initial Points for K-Means Clustering”,Proceedings of the 15th International Conference on Machine Learning,May 2012.

[12] Shehroz S. Khan, Amir Ahmad, “Cluster centre initialization algorithm for K-mean Clustering”, Elsevier Journal, July 20014

[13] Fariborz Jolai & Maziar Yazdani, “A nature-inspired meta heuristic algorithm: Lion Optimization Algorithm (LOA)”, IEEE, December 2015

[14] G. Di Caro & M. Dorigo, “A new meta-heuristic algorithm: Ant colony optimization”, IEEE, July 1999

[15] Ajith Abraham, Bighnaraj Naik , Janmenjoy Nayak & Himansu Sekhar Behera, “Optimal Cluster Analysis: Metaheuristic with Fuzzy c-means”, Elsevier Journal,, February 2017

[16] Fatma Turna & Lale Ozbakir, “Clustering performance comparison of new generation meta-heuristic algorithms”, Elsevier Journal, May2017

[17] Hongming Xu, Xin Yao ,Mohammad-H & Tayarani-N, “A Literature Survey: Meta-heuristic Algorithms in Car Engine Design”, IEEE, September 2014

[18] Aboul Alla Hassanien and S. Selva Kumar & Ahmad Taher Azar and H. Hannah Inbarani, “Rough set based meta-heuristic clustering approach for the social e-learning systems”, April 2015

[19] Francisco Herrera ,Alberto Fernandez, Victoria Lopez Vasile Palade & Salvador Garcia, “An insight into the Classification with imbalanced data: the Empirical Results and current trends on using data intrinsic characteristics”, Elsevier Journal, July 2013

[20] Joshua Zhexue Huang, Liping Jing and Michael K. Ng, “The entropy weighting K Mean algorithm for the subspace clustering of HD Sparse Data”, IEEE, June 2007

[21] kumar, Sahoo & Yugal, “A hybrid data clustering approach based on improved cat swarm optimization and *K*-harmonic mean algorithm”, IOS Press, June 2015

[22] Yuwonu, Ngyugen, Su, Moulton, “Data Clustering Using Variants of Rapid Centroid Estimation”, IEEE Explorer, September 2013