Vibrating Particle System Algorithm for Healthcare 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

Shreya Gupta (151250)

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

TABLE OF CONTENTS

List of Figures

List of Graphs

List of Tables

Abstract

Clustering is a tool in data mining which used to obtain the information which is hidden from large sets of data of various structures and clusters. It is has been seen that in the domain of engineering that over the years the focus has now shifted to the use of computing techniques which draw their inspiration from nature. The report puts forward a latest meta heuristic technique, i.e. the Vibrating-Particle System (i.e. VPS) to solve various issues of global optimization. It is essentially a meta heuristic algorithm based on population as well as the damped free vibration of single-degree of freedom system. On a set of five standard datasets which are pertaining to medical care from the UCI-Machine Learning-Repository, we evaluated the proposed algorithm. The overall results which are obtained after implementation and computations have indicated that in terms of calculations and accuracy measures, Vibrating-Particle System outshines the other hi-tech and advanced algorithms.

Chapter-1 INTRODUCTION

1.1 Introduction

The latest trends in technical advances have inevitably given rise to a situation with overflow of data.

Expansion in digital data is increasing by the day due to technological revolution. More info is generated from financial services, scientific experiments, space explorations, biochemistry, telecom and other transactions. A considerably significant quantity of data is generated on the internet in numerous formats be it image, text or some sort of multimedia format. These enormous amounts of data encompass within themselves huge number of hidden and undisclosed trends and info which can prove to be extremely useful in a variety of domains. To store such huge amounts of data, many of the relational-database servers were developed .The online transactional process (OLTP) systems are also being established to reallocate the records to database-servers. For each transaction, these OLTP systems store all transactional data in the database and make decisions on the basis of facts faster. Massive amounts of data are detected in OLTP systems and are pushed for reporting purposes to OLAP systems. With the extremely large volume of data stored in files, databases and other repositories, it is necessary to develop algorithms for examining and analysing such data, and also obtaining the hidden knowledge which might help in decision-making. Therefore the term Data Mining is also known as Knowledge Discovery in Databases, thus refers to the non-trivial extraction of inherent, previously undiscovered and potentially important data from database info. KDD is essentially the procedure of recognizing effective, original, potentially advantageous, and most importantly reasonable trends, patterns or models in information and data. Data mining is a critical step in the process of information discovery comprising of specific data mining algorithms which find patterns or models in data in certain appropriate computing efficiency constraints.

Data Mining

Data mining can essentially be defined as the process of analysis or study of data with an objective to recognize and identify patterns, trends and relations, classify and segregate the data elements and therefore predict and forecast outcomes in significantly huge datasets of organized data. Speaking in general terms Data Mining is one such unique domain where numerous fields such as the computer science, machine learning and statistics merge together in order to achieve one common target. Techniques such as Feature Selection, Clustering and many more come under the domain of Data Mining.

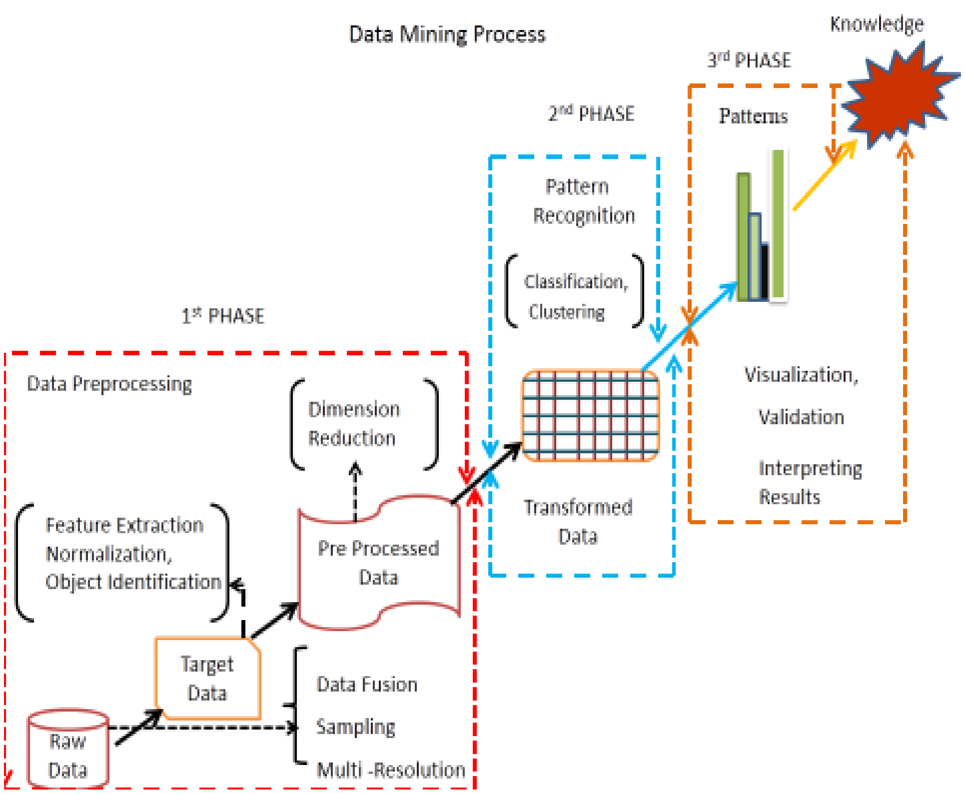


Fig. 1.1: The Process of Data Mining

The phases described in the Fig 1.1 are mentioned below

- Phase 1: Data-Pre-Processing
- Phase 2: Pattern-Recognition
- Phase 3: Interpreting-Results

Each of these three phases are described in detail in the next section

Phase 1: The first phase of Data Mining which is the Data-Pre-Processing phase basically includes the process of transformation of raw or rough data into a format which is essentially in a comprehendible format and can be easily understood. This phase is extremely important because most of the time the data existing in the actual-world is in the raw form which means that it is inconsistent, unfinished, incomplete and does contain a large number of errors. The first phase i.e. data pre-processing is a method which is aimed at resolving all these problems and issues. What this phase does is that it transforms and converts the original raw un-processed data into pre-processed data by applying various techniques of data mining such as: Feature Extraction, Data Enhancement, Data Size Reduction, Data Fusion, Normalization and many more.

Phase 2: In the second stage of the process i.e. the pattern recognition phase, the trends and patterns are identified from the data which was generated in the first phase i.e. the pre-processed data. First, the data which has been processed previously is turned into modified data by decreasing the number of objects as well as features for the sake of better pattern visualization and interpretation. Post this the trends and patterns are recognized from the transformed data set by making use of various data-mining techniques. For instance clustering, association and classification rules.

Phase 3: The third phase which is the result interpretation stage comprises of the validation and visualization of the trends and patterns that have been previously discovered in the predeceasing phase. These patterns that have already been found are visualized and are verified for the purpose of refinement.

You can characterize data mining either as predictive or descriptive.

The former technique follows the approach of forecasting the value of a target-variable on the basis of the previous historical data. Predictive data mining is also many a times referred to as the supervised learning technique or classification. The ultimate objective of this particular approach is to grow and develop a model which encompasses an executable code which will further come in very handy and useful in order to perform a number of tasks relating to data mining. On the other hand the latter technique which is the descriptive one can be described as finding new undiscovered patterns that portray the associations among the various data instances. The Descriptive Approach of Data Mining is also many a times referred to as unsupervised learning technique or more popularly Clustering. The main aim descriptive data mining wants to achieve is that it intends to examine and analyse a system by means of undiscovered trends and patterns and using them in order to find a relationship among these patterns.

Nature Inspired Techniques

For several hundred million years, nature has evolved and progressed and while doing so has found a no. of inventive answers for real world problem solving and adaptation to ever evolving environments. As per the very famous Theory of Evolution put forward by Darwin, The survival of the most fitting species will lead to the variations and triumph of those species that can endure and adapt ideally to the environment Selection therefore is a factor of continuous pressure that continuously energies the system to advance and adapt for the sake of existence. By emulating the successful characteristics of compound systems present in the environment, we can learn a thing or two from nature. In recent decades, numerous nature-inspired optimization algorithms have been created to solve a wide range of optimization problems successfully.

A few Notable examples of Nature-Inspired Algorithms are:

- Differential-Evolution
- Simulated-Annealing
- Genetic-Algorithm
- The Evolutionary-Algorithms
- Particle-Swarm Optimization

Nature-inspired algorithms are still at a rather preliminary early stage with a relatively short history opposed to many conventional, well-established methods such as dividing and conquering, dynamic programming, branch and bound, gradient methods and linear programming. But having said that the nature-inspired algorithms have now shown their Massive potential, versatility and effectiveness with a wide variety of applications. While these nature-inspired techniques have illustrated excellent search capabilities to solve optimization problems ranging from small to medium size , they still continue to face some serious challenges in solving optimization problems of large scale, i.e. problems with include several thousands of variables.

The Nature-Inspired Algorithms or Techniques are basically the techniques which:

- Study of the concepts and implementations of natural events such as the Theory of evolution by Darwin or the concept of gravitation etc.
- Study the basic behaviour of living beings such as insect, birds or even swarms , for instance Ant-colony optimization.
- Analysing the biological progression of numerous living creatures and their reproduction strategy. For instance The Genetic-algorithm or the artificial-neural network and many more.

1.2 Problem Statement

The need for algorithm is illustrated in this portion of the project report. Throughout the past, a significant number of algorithms pertaining to clustering were introduced and subsequently implemented for multiple problems relating to optimization and many of these approaches experienced similar issues which are mentioned below:

I. Failure to attain a cohesive balance among exploitation and exploration processes.
II. Lacking diversity and local optima
III. Quality of Solution

The main purpose is to identify those cluster-centroids which are the best also known as the cluster representative this essentially refers to those cluster centres which have the minimum inter-cluster distances between themselves. Apart from this the other area at which this particular project mainly focusses is on to significantly enhance the overall accuracy of the dataset under consideration. Another objective of our project is to in a way introduce the chaotic maps into the algorithm which is being proposed so as to overcome the issue of randomness and to device and ultimately implement a local search technique which will in a way enhance value of the final result.

The project also aims at overpowering the shortcomings which have been previously mentioned by means of application of the Vibrating-Particle System Algorithm for solving numerous problems related to global optimization. Application of the vibrating-particle System algorithm on a no. of different data-sets related to various diseases have shown that Vibrating-Particle System is indeed a rather competitive clustering-algorithm with reference to the other current meta heuristic algorithms. By means of this project we have been able to successfully apply the Vibrating Particle System algo on the various medical care based optimization-problems and the outcomes which were subsequently obtained after implementation clearly exhibits the effectiveness of this particular to other Problems in the actual world apart from medical care based optimization-problems .To propose a new fangled methodology with the purpose of avoiding the local optima problem.

Lastly the project also compares the performance of the VPS algorithm against the performance of WWO algorithm.

1.3 **Objectives**

It has been found from the literature-review that in comparison with conventional algorithms, evolutionary algorithms provide better outcomes. The project is aimed at implementing the VPS algorithm on different Medical care datasets with an objective to reduce the inter-cluster distance to a minimum and attain cluster-centres which are optimized and to investigate the performance of this algorithm on the healthcare datasets. Apart from this the project also aims to make a comparison between the VPS algorithms' performance with that of the WWO algorithms' for the same data-sets as given here.

i. WDBC Dataset
ii. Thyroid Dataset
iii. Dermatology Dataset
iv. Heart Dataset
v. Bupa Dataset
vi. BCW Dataset

The source of all the above mentioned datasets is the UCI Machine Learning Repository. Each of the six mentioned datasets have been taken and monitored or examined one at a time by making use of the VPS algorithm for clustering and various functions that are being used.

Thus in a nutshell our ultimate objective for the project is to build clusters and then identify the data from the entire dataset which has a relatively higher accuracy. Different classes are made by selecting the data points are from these data-sets. The criteria used for this purpose is the portion which has been given along each data set that has been obtained from the source of UCI Machine Learning Repository. After the classes are identified, the clusters are then created according to the previously identified classes.

1.4 Methodologies

Clustering is ultimately aimed to identify the likenesses and similarities which is prevalent in the data-point and then subsequently cluster the identified alike data-points together into one single group. Over the years a large number of algorithms have been developed and implemented for the purpose of clustering. In this particular project we have implemented one of the a very well-known and extensively used algorithm in the machine learning domain that is the K-means clustering algorithm. Unsupervised learning is used in our approach in clustering together with data cataloguing. Our project's ultimate job is to identify the top cluster-centroids that have minimum distances amongst themselves in the cluster. Apart from this the main focus is also to improve the accuracy of the data-set under consideration. Computations conducted on the various health care datasets clearly demonstrate that our algorithm is a well performing clustering algorithm when compared with the other current meta heuristic algorithms.

Training of the model is achieved by carrying out work on the algorithm using the algorithm for optimizing the updated VPS and then run some time-to-time functions to figure out the accuracy and fitness of input data-points. The fittest data point is then recognised by identifying the point that has the maximum value of accuracy percentage among all the other data-points.

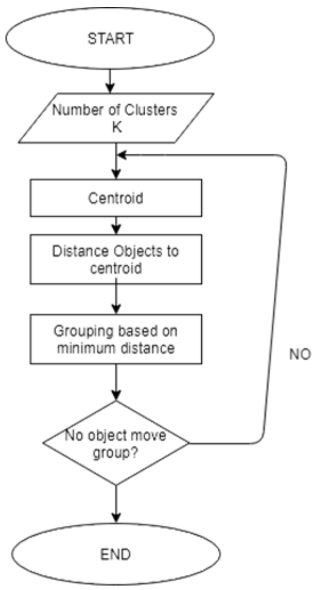


Fig 1.0 Diagram representing K-Means Clustering

1.5 Organization

Chapter-1:

A brief overview about the idea of what we are trying to achieve and by means of completion of this project .Within this chapter we also include the problem statement that we are dealing with and the aim of our project alongside the inclusion of the preliminary concepts which are essentially required in order to accomplish the ultimate objective.

Chapter-2: This chapter primarily focuses on literature-survey. We have gone through a number of research-papers as well as journals from reliable and reputed sources to gain an insight into the topic we are dealing with.

Chapter-3:

This particular chapter deals with the various details regarding system development.

The algorithm we implement with the elaboration for the same and the equations used has been included in this chapter.

Chapter-4:

In this chapter the outputs and analysis of the performance are given in this chapter. On a given dataset, we applied the algorithm and presented the graphs and values of result for the same.

Chapter-5:

We draw the conclusion the project with the results and conclusions in this chapter and reflect the future scope for the further development and implementation of the project.

Chapter-2 LITERATURE SURVEY

2.1 Title: A new meta-heuristic algorithm: Vibrating particle system

Author: A.Kaveh and M.IlchiGhazaan

Publication Years: September 2016

Publisher : Scientia Iranica

The VPS is a current meta heuristic algo which is essentially based on free vibration introduction with viscous damping of single degree of freedom systems. As per the approach in this algorithm the solution candidates progressively come to their balancing positions and are regarded as particles. So in order to attain the accurate amount of equilibrium between diversification as well as intensification, the latest population tends to find their equilibrium positions and also the previously known historically best positions. The method which is being proposed here is used in order to optimize the 4key skeletal structures which notably include frames as well as trusses, for the sake of evaluating their performance. The method being proposed here also indicates its capability to come in useful for solving only a limited set of problems.

2.2 Title: Adaptive Multi subpopulation Competition and Multiniche Crowding-Based Memetic Algorithm for Automatic Data Clustering

Authors: Weiguo Sheng, Gang Xiao, Yujun Zheng, Jiafa Mao
Publication Years: February 2016

Publisher: IEEE-Journal

Sheng et al. [49], reported adaptive multi subpopulation competition and multiniche crowding based memetic algorithm to tackle automatic data clustering. The goal of the proposed algorithm is to identify the high-quality solutions effectively and efficiently for automatic data clustering. In this work three artificial data sets and five real data sets are considered to compute the performance of the algorithm. It is seen that the proposed algorithm shows better and superior performance as compared to other related methods.

2.3 Title: Categorical data clustering: What similarity measure to recommend

Author: Dos Santos and Zarate

Publication Years: February 2015

Publisher: Science Direct

Dos Santos and Zarate[10], reported nine distinct measures with the help of TaxMap clustering mechanism to find is there a similarity measure in categorical variable which is more stable and provides satisfactory results in databases. In these fifteen different databases are considered to compute the experiment by using the clustering quality measures such as NCC, entropy, compactness and silhouette index. It is observed that the similarity measure proposed in this work shows best performance comprehensively.

2.4 Title: Memory enriched Big Bang- Big Crunch algorithm for data clustering

Author: Kayvan Bijari, Hadi Zare, Hadi Veisi, Hossein Bobarshad

Publication Years: March 2018

Publisher: Springer Link

Bijari et al. [7], presented a new heuristic algorithm for solving clustering problems. The presented algorithm reduces the typical clustering algorithm with the advantage of the heuristic nature. The proposed algorithm is based on Big bang-big crunch algorithm and its performance is evaluated by experimental results over six data sets. It is observed that the presented algorithm show dominance over other similar algorithms for solving clustering problems.

2.5 Title: SUBSCALE: Fast and Scalable Subspace Clustering for High Dimensional Data

Author: Kaur and Datta

Publication Years: December 2014

Publisher: IEEE

Kaur and Datta[20] described a new clustering algorithm to find the subspace clusters in the density-based clustering. The aim of the algorithm is to find the non-trivial subspace clusters even in the high dimension with the minimal cost. The proposed algorithm is highly parallelizable and scalable, it shows the best performance as compared to other subspace clustering algorithms. In this work 13 data sets are used to compute experimental results and the algorithm used requires only k database scans for the k- dimensional dataset to find the subspace clusters.

2.6 Title: Unsupervised Metric Fusion Over Multiview Data by Graph Random Walk-Based Cross-View Diffusion

Author: Yang **Wang, Wenjie Zhang, Lin Wu, Xuemin Lin, Xiang Zhao**

Publication Years: December 2015

Publisher: IEEE

Zhang et al. [59], presented a new random walk-based clustering method to find attractor vertices and cluster them. In the presented method the inflation function and normalization functions are adopted to restrict the reach of walking agent. In this work data sets used to are Zachary's karate club, Ego-Facebook graph data, Heterogeneous graph data. The proposed method is able to work in any parallel computing environment and from the experimental results over simulation it is concluded that it is superior as compare to other graph clustering method.

2.7 Title: A charged system search approach for data clustering

Author: Kumar and Sahoo

Publication Years: April 2014

Publisher: ACM

Kumar and Sahoo [25], presented an algorithm which I s inspired from charged system to find solutions for clustering problems. In this paper CSS algorithm is used to find the optimal centroid. In this work two artificial data sets and eight real data sets were used to compute the performance. It is observed that the presented algorithm provides enhanced and more precise results as compared to another algorithm. Gebru et al. [13], introduced weighted-data Gaussian mixture model for clustering problems in heterogenous\multimodal data sets. In this work the two expectation maximization algorithms are derived and these algorithms are based on fixed weights and random weights. In this work four simulated datasets and four publicly available data sets are used and these are MNIST, WAV, BCW, Letter Recognition. In is seen that the derived algorithms provide enhanced and robust results in comparison to the parametric and non-parametric clustering techniques.

2.8 Title: Making kernel-based vector quantization robust and effective for incomplete educational data clustering

Authors: Thi Ngoc Vo, Nyugen

Publication Years: March 2016

Publisher: ACM

Vo et al. [53], adopted VQ_fk_nps a robust solution based on kernel-based vector quantization for incomplete data clustering. The proposed solution for incomplete data clustering adopted the nearest prototype strategy to optimize clusters to reach the resulting cluster with arbitrary shape in the data space. In this work the data sets used are Year 2 for second-year students, Year 3 for third-year students, and Year 4 for fourth-year students. It is observed the proposed solution provides enhanced quality clusters as compared to other existing approaches.

2.9 Title: Meta-learning systems for Clustering

Authors: Ferrari And Castro

Publication Years: March 2015

Publisher: ACM

Ferrari and Castro[12], described the new ways to collect meta knowledge for clustering task. In this paper two concepts are explored to combine the internal indices for ranking algorithms and to characterize clustering problems. In this work several datasets are considered to compute the performance. It is seen that the proposed meta-attribute set is compared with the classical approach and concluded that it provides more enhanced and precise results with high recommendation quality. They presented a novel pattern-based clustering algorithm for numerical datasets. The aim of the presented algorithm is to obtain patterns of numerical datasets without using priori discretization algorithm. In this work twenty data sets are considered to compute the performance. It is evaluated that the proposed algorithm provides better results as compared to other pattern-based clustering algorithms for clustering.

2.10 Title: Distributed Data Clustering Using Mobile Agents and EM Algorithm

Authors: Safarinejadian and Hasanpour

Publication Years: August 2014

Publisher: IEEE-Xplore

Safarinejadian and Hasanpour[44], reported a MABDEM algorithm for sensor networks. The aim of the reported algorithm is estimation of distributed density and data clustering in sensor networks. This algorithm executes expectation maximization algorithm in distributed manner and able to lower down its number of iterations. In this work the synthetic and real data sets are considered to compute the performance of the algorithm. Allab et al. [2], adopted a new way for data clustering and reduction of dimension simultaneously. The adopted methodology relies over Semi-NMF-PCA. In this work three FCPS data sets, ten document-term data sets and thirteen image and microarray data sets are considered to compute the experimental results. It is seen that the adopted model provides enhanced results as compared to other state-of-the-art algorithms in terms of clustering

2.11 Title: Particle Swarm Optimization Based Hierarchical Agglomerative Clustering

Authors: Alam, Dobble, Riddle

Publication Years: November 2010

Publisher: IEEE-Xplore

Alam et al. [1], discussed Evolutionary PSO and Hierarchical PSO to tackle data clustering in hierarchical manner. The aim of the proposed work is to done more accurate and effective clustering. In this work seven data sets are considered to compute the performance of the proposed work as compared to other algorithms. It is observed that the proposed techniques provide much accurate and efficient results over suggested measures as compared to other techniques. This was improved algorithm for HKA -K for partitional data clustering. The proposed algorithm uses the combination of Heuristic Kalman Algorithm and K-means method. In this work two synthetic and five well known data sets are considered to compute the performance of the proposed algorithm. It is seen that the proposed algorithm is far better than other compared algorithms. The proposed algorithm is modified version of Gravitational Search algorithm and it is inspired form the collective response behaviour of birds. In this work thirteen real data sets are considered to compute the performance of the proposed algorithm.

2.12 Title: A prototype classifier based on gravitational search algorithm

Authors: Abbas Bahrololoum, Hamid Bahrololoum, Saeed

Publication Years: February 2012

Publisher: Science Direct

Bahrololoum et al. [5], proposed a gravity-based algorithm for data clustering. The objective of the proposed algorithm is to reduce effect of noise and enhance clustering. The proposed algorithm is inspired form the Newtonian law of gravity. In this work twelve data sets are considered to compute the performance of the proposed algorithm. It is concluded that the proposed algorithm shows effective and efficient results as compared to other algorithms. The proposed algorithm is relying over standard K-means and K-Harmonic means, these are used as a fitness function in Fish School Search algorithm. In this work thirteen data sets are considered to compute the performance of proposed FSS-SCA algorithm. It is seen that the proposed algorithm shows slightly better and improved results as compared to K-means and PSO algorithms.

2.13 Title: Efficient protocol for data clustering by fuzzy Cuckoo Optimization Algorithm

Authors: Ahsan Amiri, Shahid Mahmaudi

Publication Years: April 2016

Publisher: Science Direct

Amiri and Mahmoudi[3], proposed a new Fuzzy Cuckoo Optimization algorithm for partitional data clustering. The aim of the proposed algorithm is to determine the number clusters. In this work seven data sets are considered to compute the experimental results of the proposed algorithm. It is observed that the proposed algorithm is compared with other data clustering algorithms and concluded that the proposed algorithm provides better performance than other algorithms.

2.14 Title: Sparse Regularization in Fuzzy c -Means for High-Dimensional Data Clustering

Authors: Chang, Liu, Yu Wang, Quwan Wang

Publication Years: December 2016

Publisher: IEEE-Xplore

Chang et al. [8, introduces a new fuzzy c-means (FCM) model with sparse regularization to handle high dimensional data problems. The objective of the proposed model is to identify the relevant features and discovering the cluster structure in high dimensional data. In this work the data sets considered are Yeast, Libra movement, Gesture Phase Segmentation to compute the experimental results. It is seen that from the experimental results that the proposed model provides better and enhanced results as compare to other clustering approaches.

2.15 Title: A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping.

Authors: Ghorbanzadeh et all.

Publication Years: August 2018

Publisher: ACM-Pulication

Ghorbanzadeh et al. [14], adopted an adaptive neuro-fuzzy based correlation model for tumour motion tracking. The aim of the adopted model is to achieve efficient performance and reduce error in tumour motion tracking. In this work to evaluate the performance of the proposed model it is tested over twenty patients. It is seen that the proposed model is much efficient and effective in reducing the tumour tracking errors as compared to Cyberknife model.

2.16 Title: A novel hybrid approach using wavelet, firefly algorithm, and fuzzy ARTMAP for day-ahead electricity price forecasting

Authors: Meng, MAtinez, Amit K. Srivastva

Publication Years: November 2012

Publisher: IEEE-Xplore

Meng et al. [34], investigates the vigilance parameter to make it self-adaptable in Fuzzy ART and reported three algorithms which relies over fuzzy adaptive resonance theory. The objective of the proposed work is to develop high quality clusters in large scale social media data sets with the help of simple parameter settings. In this work, four data sets are considered to compute the experimental results. It is evaluated that the proposed algorithms provide better and enhanced performance as compared to other state of art clustering algorithms.

2.17Title: A dynamic shuffled differential evolution algorithm for data clustering

Authors: Xieng, Zu , Meng

Publication Years: June 2015

Publisher: Science Direct

Xiang et al. [54], proposed a dynamic shuffled differential evolution algorithm for data clustering. The proposed algorithm is inspired by shuffled frog leaping algorithm. The aim of the proposed algorithm is to enhance the convergence performance in data clustering. In this proposed algorithm random multi-step sampling is integrated into it to resolve the problem of premature convergence. In this work eleven data sets are considered to evaluate the performance of the algorithm. It is concluded that the proposed algorithm is most efficient and effective tool for data clustering as compared to other algorithms.

2.18 Title: Multilocal Search and Adaptive Niching Based Memetic Algorithm With a Consensus Criterion for Data Clustering

Authors: Sheng, Chen, Xao, Mao

Publication Years: September 2013

Publisher: IEEE-Xplore

Sheng et al. [48], reported a genetic algorithm for automatic data clustering. The reported algorithm relies over multilocal search and adaptive niching. The goal of the proposed algorithm is to evade possible stagnation and premature convergence. In this work three synthetic and six real data sets are considered to compute the experiment results and to evaluate the performance of the algorithm. It is observed that the proposed algorithm is superior and enhanced as compared to other algorithm in respect to performance.

2.19 Title: Parameter adaptive harmony search algorithm for unimodal and multimodal optimization problems

Authors: Vijay Kumar, Jitender Kumar chabra,Dinesh Kumar

Publication Years: March 2014

Publisher: IEEE-Xplore

Kumar et al. [23], utilizes the parameter adaptive harmony search (PAHS) as an optimization strategy for automatic data clustering. The main aim of this work is to detect the number of clusters automatically and find optimal centroid. In this work five data sets are considered to evaluate the experimental results. It is concluded that the proposed work provides better clustering results as compared to other clustering techniques. It is also observed that the clusters produced by ACPAHS are well separated and compact.

2.20 Title: Adaptive Clustering for Dynamic IoT Data Streams

Authors: Daniel Puxhmann, Rahim Tafazolli
Publication Years: October 2016

Publisher: IEEE-Xplore

Puschmann et al. [43], introduced adaptive clustering method for dynamic IoT data streams. The objective of the proposed method is to find how many clusters can be formed from data stream. In this work synthetic data sets are considered to perform experiments and compute results. It is evaluated that the proposed adaptive algorithm produces more enhanced clusters as compared to non-adaptive algorithm in contrast to cluster quality. The proposed work is applied on eleven data sets and performance is evaluated in contrast to particle swarm optimization algorithm and artificial bee colony algorithms. The proposed model is able to handle the heterogeneity and introduces variable neighbourhood search algorithm to find solution efficiently even for the large problems. In this model the best clustering solution is determined for the similar clustering group in which the hetero individual is assigned. It is observed from the experimental evaluation that the clustering structure can be recovered from the available datasets.

Authors	**Algorithm**	**Exploitation (Global Search)**	**Exploration (Local Search)**	**Shortcoming**	**Improvement**	**Bench marks**	**Performance Parameters**	**Compared Algorithms**	**Statistical Test**
Han et.al., (2017)	Bird Flock Gravitational Search Algorithm (BFGSA)	The cluster centroids are updated using velocity and coordinate updating formula of GSA.	New search space is explored using nearest neighbour method (mean of 7 nearest neighbours).	Local Optima, unable to handle multidimensional data, and premature convergence.	Introduced new diversity mechanism	Balance, Cancer, Cancer-Int, Credit, Dermatology, E. Coli, Diabetes, Glass, Heart, Horse, Iris, Thyroid and Wine.	Average intra-cluster distances and Error rate.	Standard GSA, PSO, ABC, FA, K-means, NM-PSO, K-PSO, K-NM-PSO and CPSO	Wilcoxon signed-rank
Kumar and Singh (2017)	Improved cat swarm optimization algorithm (ICSO)	The global best position of catis achieved in tracing mode	The global best position of cat is used to guide the positions of cats in tracing mode	Inappropriate balance between exploration and exploitation, lack of diversity, and slow convergence rate.	New amendments like accelerated velocity equation, position update equation has introduced in ICSO to handle clustering problems.	Iris, Wine, CMC, Cancer and Glass.	Cluster quality (Best, average, worst and standard deviation), and F Measure	K-means, PSO, ACO, CSO andTLBO	Friedman test, Wilcoxon signed-ranks test
Rana et.al., (2013)	Boundary Restricted Adaptive Particle Swam Optimization (BR-APSO)	Updating the velocity and position of particle using boundary restricted strategy.	Calculating the inertia weight exponentially	Local optima and Outliers	Introduced Boundary restriction strategy to handle outliers.	Art1, Art2, Vowel Iris, Crude oil, CMC, Cancer, Glass and Wine	Sum of intra cluster distance and Error rate	K-means, PSO NM-PSO, K-PSO, K-NM-PSO, LDWPSO and ALDWPSO	
Tsai and kao(2011)	Selective regeneration particle swarm optimization (SRPSO)and Hybrid K-means and selective regenerated particle swarm	Updating the velocity and position of particle	Selective particle regeneration technique	Local optima, convergence speed	Introduced unbalanced parameter setting for fast convergence and particle regeneration operation to escape	ArtSet1, ArtSet2, Iris, Crude oil, Cancer, Vowel, CMC, Wine and Glass	Sum of intra-cluster distances and Error rate (ER)	KSRPSO, PSO and K-mean	

	Optimization(K-SRPSO)				from local optima				
Zou et.al., (2010)	Cooperative Artificial Bee Colony Algorithm(CABC)	Each Individual bee (Employed and onlooker bees)	Virtual bee	Initial cluster centre selection, Local optima and poor convergence speed.	Cooperative approach	Motor cycle, Iris, Wine, CMC, Cancer and Glass,	Intra cluster distances (Average, best, worst and standard deviation)	ABC, PSO, CPSO and K-means	
Dowlatshahi and Nezamabadi-pour(2014)	Grouping Gravitational Search Algorithm (GGSA)	Reinsertion phase	Inheritance phase	Local optima and redundancy	Special group encoding scheme	Balance, Cancer Cancer-Int, Credit Dermatology, Diabetes, E. Coli, Glass Heart, Horse Iris and Thyroid Wine	Classification Error and Rank.	Standard GSA, MLP-ANN, Bayes Net, Bagging, NBTree, KStar, Ridor, PSO, ABC, VFI, MultiBoost, RBF-ANN and FA	Wilcoxon signed-rank test
Jensi and Jiji (2016)	Improved Krill Herd algorithm(IKH).	Greedy selection technique for better krill position	Original KH algorithm steps (Foraging action, Physical diffusion, and crossover operator)	Poor at Exploitation, local optima	Global search operator and elitism strategy	Iris, Wine, Glass, Cancer, CMC, Vowel and Livor Disorder(LD)	Intra-cluster distance	K-means, K-means++, GA, SA, TS, ACO, HBMO, PSO, KH	Wilcoxon rank sum test
Malinen et. al.,(2014)	K-means*			Local Optima and Empty clusters	Inverse Transform step and random swap strategy	s1, s2, s3, s4, a1, DIM32, DIM64, DIM128, DIM256, Bridge,Missa,House, Thyroid,Iris, Wine, Breast, Yeast, wdbc and Glass	MSE, NMI, Normalized Van Dongen and Incorrect clusters	K-means, Repeated k-means, K-means++ and FastGKM	
Ji et. al.,(2013)	Improved k-prototypes clustering algorithm			Mixed data	Combined mean with distribution centroid	Iris, Soybean Heart Disease and Credit Approval	Accuracy	K-prototypes, SBAC, and KL-	

					to represent prototypes of clusters			FCM-GM		
Sarma et. al.,(2013)	lk-means clustering method with Varying Threshold (lkmeans-CMVT).			Speed and Empty cluster formulation	lk-means clustering method with Varying Threshold (lkmeans-CMVT).	PENDIGITS, OCR, LIR and Synthetic data-sets SD2 to SD40	Running time	lk-means-CMFT, filtering method, lloyd's k-means		
Li et. al.,(2012)	Chaotic particle swarm fuzzy clustering (CPSFC) algorithm	Gradient method	Gradient method and chaotic local search	Local optima and convergence speed	CPSO algorithm for local optima and gradient operator introduced for faster convergence	Circular_4_2 Sphere_4_3 Circular_5_2 Sphere_5_3 Circular_6_2 Sphere_6_3 Iris Wine Vowel Glass Ecoli Liver disorder Vowel	Optimal values (Mean, Standard deviation)	FCM, GAFCM, and PSOFCM		

Chapter 3

SYSTEM-DEVELOPMENT

3.1 Flowcharts

The figure listed as Fig. 3.1 is the diagrammatic representation of the flowchart of the VPS algorithm for clustering. The In this flow diagram we start by initialization of the initial cluster centres. After this initialization step we go on to make a call to function labelled as Class-call . This Class-call function will then return the class variables accordingly to each of the cluster centre. After this has been done i.e. class-call has completed its execution, the call to next function, that is,the 'accsum' function is made. value The call to the accusum function will eventually return the value of accuracy of each of the cluster centres and apart from this will also return inter-cluster distance. Until the maximum iterations are complete and the condition of termination is not reached, the fitness of the data particles is found. We sort the fitness after calculating the fitness of the data particles and Ybest is the fitness max and Ybad is the fitness minimum. After this we calculate the average in order to find out theYgood particle. By values making use of the rand function, we go on to initialize the constant P and Q whose value is selected after finding all three positions. We calculate the values of w1,w2,w3 based on the value of the constant P. Next we calculate the value of A using all these parameters using equation 11. After this we calculate the value of B new to enable us to find the new cluster centres. This is preceded by examining the value of new cluster centers with the boundary conditions and producing updated cluster centers in B up. When the termination condition is finally reached we plot the graph between total number of iterations and the inter cluster distance.

Figure-3.3 shows the accsum function flowchart. When ever the main function makes a call the accusum function, it first starts by identifying the algorithm's ' accuracy ' by measuring the assigned class correctly and incorrectly.

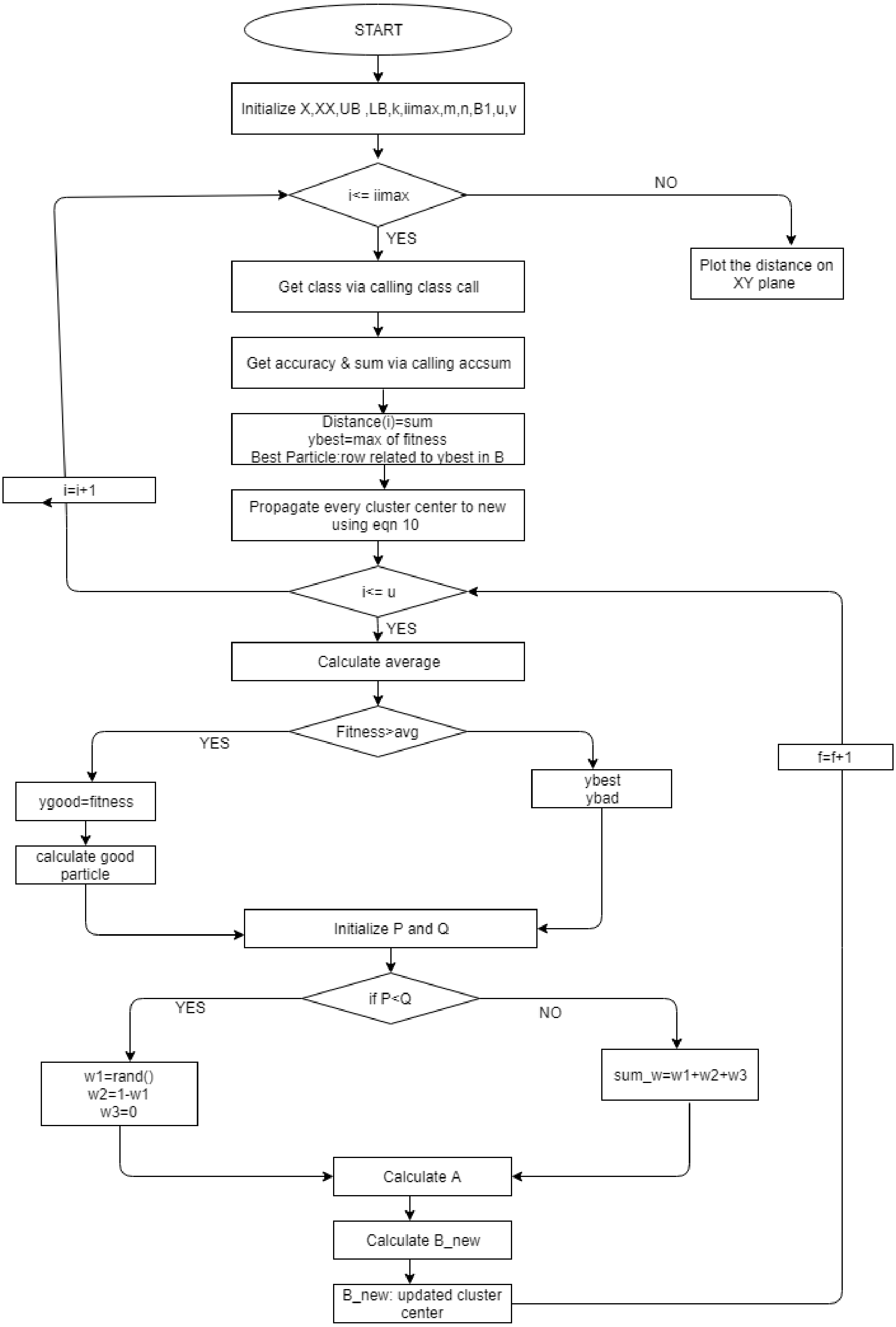


Figure-3.1. flow-chart of the Vibrating Particle System algorithm

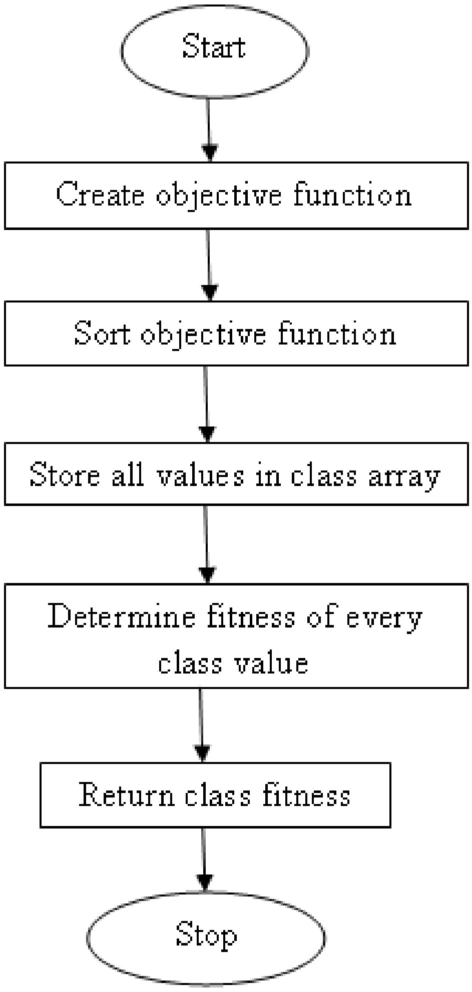


Figure-3.2. flow-chart of the class-call function

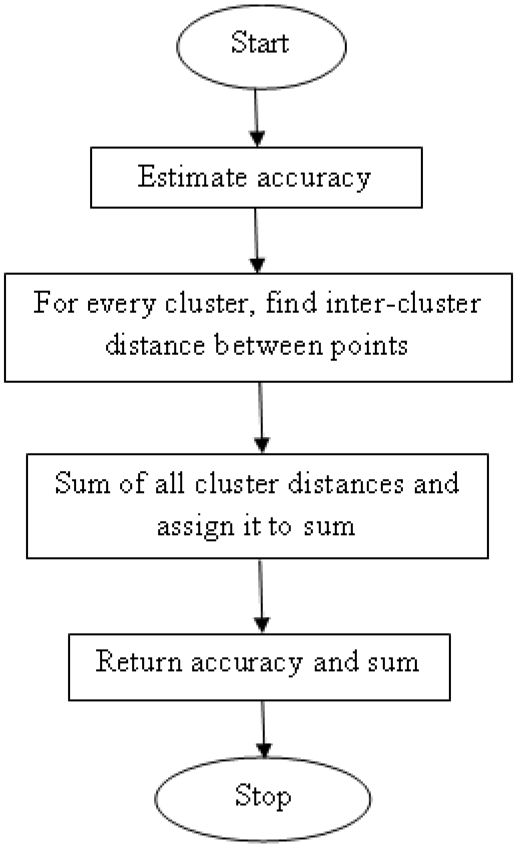


Figure-3.3. flow-chart of the accuracy function

The diagram which has been labelled as Figure-3.4 is the flowchart for the WWO algorithm. First, we begin by assigning a value to the initial cluster-centers as can be seen in Figure-3.4. Now we are going to call for the function ' classfit'. The class-fit function returns data point fitness and class variables (which have been named 'fitness1') by cluster centre. After this the next call is made to the next function, that is, the ' accu-sum ' function that returns the cluster centre accuracy and inter-cluster distance.

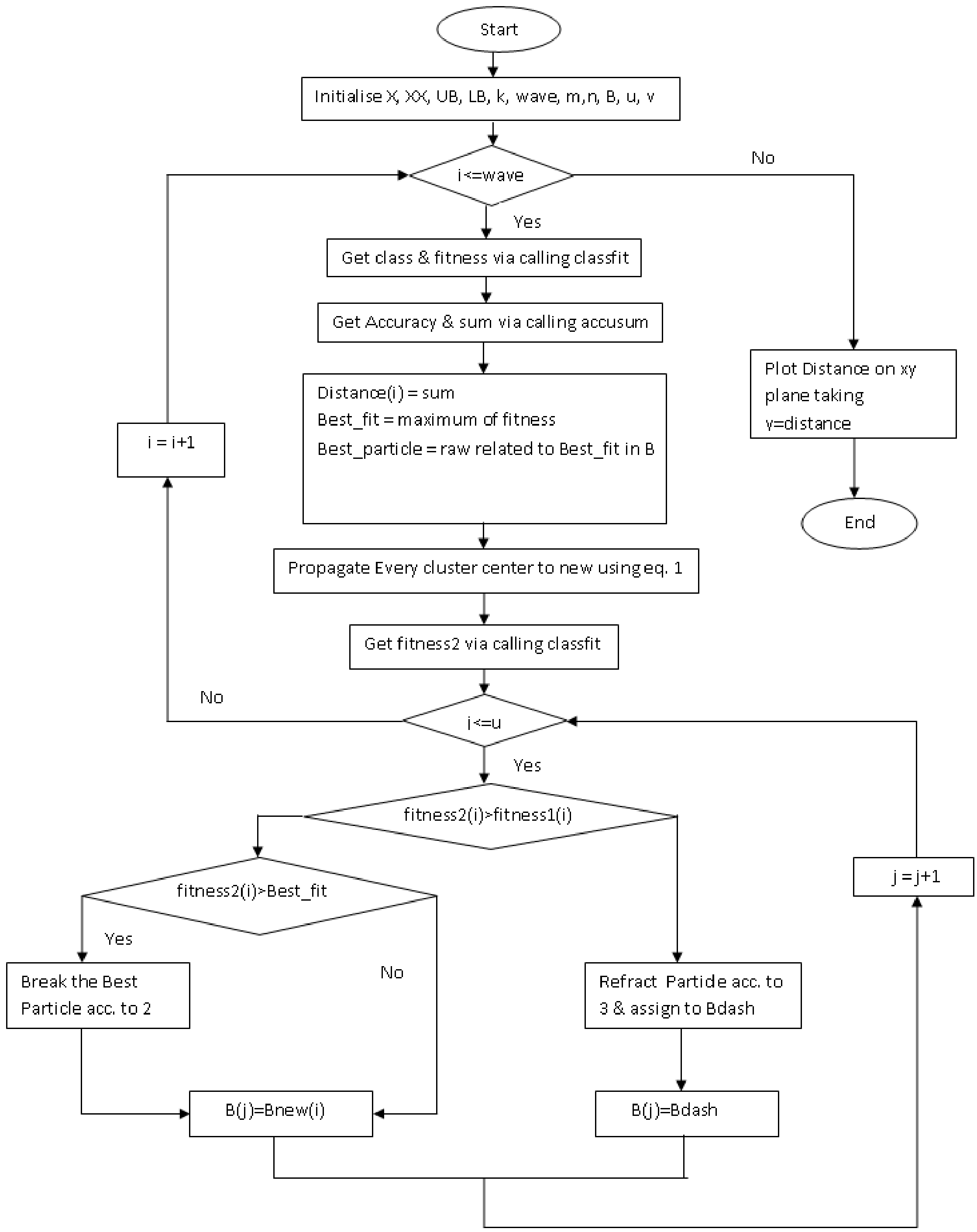


Figure-3.4. flow-chart of the WWO algorithm

3.2ALGORITHMS

The algorithm used is VPS which is one of the latest population-based meta heuristic algorithm that is based on a single degree of freedom system's damped free vibration.The number of candidate solutions representing the particle system is included. The particles gradually approach their position of balance and are randomly initialized in a n-dimensional search space.

The following steps are taken into consideration for the algorithm:

Step 1: Initialization: All VPS parameters are set and the initial positions in a n-dimensional search space are determined randomly.Different parameters are included in the algorithm to calculate the accuracy, fitness, intra cluster distance and are initialized.

Step 2: Evaluation of Candidate Solution: An objective function is created in the algorithm whose value is calculated using the number of clusters, the size of the data set selected for the number of iterations specified for loops and parameters.

Step 3: Updating the particle positions: The particle tends to approach three different equilibrium positions with different weights such as

1. The best position in the population that is Best Particle

2. A good Particle

3. A bad Particle

To find the GP and BP for each particle the objective function is used to sort the current population in an increasing order. A descending function that is proportional to the no. of iterations is used for the optimization algorithm. The equation for the following is :

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

The present iteration number in use in the above equation iteration, itermax is the largest number of iterations being used and α is a constant being used.

The equation used to update locations is given as follows:

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

In the above equation x_i^j is the j^{th} variable of particle . To measure the relative importance of GP,BP and HB, three parameters w_1,w_2,w_3 are used. To compute x_i^j the eqn mentioned here is made use of:

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

$$w_1 + w_2 + w_3 = 1$$

A parameter p is defined within (0,1) range and is compared with rand for each particle. The GP and BP are chosen after that if $p<rand$ then $w_3=0$ and $w_2=1-w_1$

Step 4: In order to find better results in the search space the particles may violate the side constraints. To resolve this violation the boundary must be regenerated by harmony search based side constraint handling approach.

Step 5: Terminating criteria Controlling: Until the termination criteria is fulfilled the step 2-4 are repeated. In this algorithm the number of iterations is considered ad the terminating condition but it can be some other condition also.

	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

Table 3.1 VPS Algorithm (Pseudocod

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*	

Table 3.2 WWO Algorithm (Pseudocode)

There are three wave operations that are used in the algorithm, i.e. refraction, braeking, and propagation.

i. Propagation:

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

ii. Refraction:

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

iii. Breaking:

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

.

3.3 Test Plan

In this section we discuss the various data sets that we use to implement our algorithm in order to obtain the optimized cluster-centres.

1. BUPA
2. Heart
3. BCW
4. WDB

5. Thyroid
6. Diabetes

3.3.1 Data-Sets

Detailed description of the data sets:

1. BCW-Dataset

Dataset information: As Dr.Wolberg claims in his clinical cases, the samples reported in the dataset are received periodically. On July 15, 1992, the + dataset was provided. Below is the chronological order of receipt of the data samples.

Group 1: 367-instances (Jan-1989)

Group 2: 70-instances (Oct-1989)

Group 3: 31-instances (Feb-1990)

Group 4: 17-instances (Apr-1990)

Group 5: 48-instances (Aug-1990)

Group 6: 49-instances (Jan-1991)

Group 7: 31-instances (Jun-1991)

Group 8: 86-instances (Nov-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)

Table 3.3 Information regarding BCW Dataset

2. WDBC

Dataset Information:

Breast mass characteristics are measured by using the digitized image of a fine needle aspirate also known as the FNA. It explains the qualities of the cell nuclei visible 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.

Table 3.4 Information about 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

Table 3.5 Information about heart disease dataset

4. BUPA

Dataset information: The samples in the provided dataset constitute each and every male record. The primary five variables are the ones corresponding to the blood sample test-results and are expected to be responsive to liver illnesses that could result in excessive amounts of alcohol consumption.

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

Table 3.6 Information about 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

Table 3.7 Information about Diabetes dataset

6. THYROID

Dataset information: Gravan Institute provided a total no. of ten distinct data-sets, one of which is used here. Stefan Aeberhard offers the data set.

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

Table 3.8 Information about Thyroid dataset

3.4.2 Metrics

1. Accuracy-Matrix

Accuracy-Matrix is also a single row matrix where the total no. of columns equivalent to the total no. of iterations. The Accuracy-Matrix Illustrates the correctness of our forecasted cluster centers and how much algorithm-assigned class variables Are assigned correctly by equating with the class file we already have. Broadly speaking, the Accuracy-matrix demonstrates a rising trend with every next iteration. The cluster-centres adjust in accordance to the class variables assigned fluctuations and are positioned in roughly the correct positions where they should go, resulting in increased overall accuracy.ClusterCentre Matrix

A cluster-matrix demonstrates the data-centres it will accomplish during the no. of iterations possible. Cluster-centre is a matrix point or value assumed to be the center of similar type of data points. According to the above data set, we will have three rows in the cluster matrix as we have three class options. New cluster centers are produced from each iteration by making use of the propagation equation. Old cluster centers are then revised appropriately to new cluster centers.

2. Distance-Matrix:

   Distance-Matrix is a single row matrix where the no. of iterations is equivalent to the total no. of columns. The Distance-Matrix illustrates the distance in the cluster as far as data points and cluster centers are concerned..As we calculated the cluster centers, using the technique of root mean square to calculate every data point distance between the cluster center and that point to each cluster center. Then in the Distance Matrix, sum of all values is designated, this matrix shows steep declines particularly in comparison to the matrix of accuracy.

3.4.3 Test- Setup

This particular test-setup is unlike any other types of testing wherein we do offer the test data and then see whether the obtained output is correct or not. Therefore, post developing and implementing the algo , here we are plotting the graph between the inter cluster on the labelled x axis and the total number of iterations which is plotted against the y axis.If the graph in some way is showing the decreasing trend then it is indicative of the very fact that the algorithm which has been designed and implemented is working accurately for given data set. To ensure this we can further go on to implement our algorithm for more no. of datasets in order to verify if the output graph obtained in each of the case is also according to what they need to be i.e. they too show a decreasing trend .If the case turns out to be so are ,then We can continue to say that the engineered algorithm works properly.

CHAPTER- 4 RESULTS AND PE RFORMANCE ANALYSIS

Following are the results that were noted after the successful implementation of the VPS algorithm used on the different medical data-sets. In order to analyse the performance as well as the accuracy of VPS we applied this algorithm on six distinct data-sets relating to healthcare. After implementing the algorithm on the six different sets of data we plotted and obtained the resultant graphs which were produced by plotting the intra cluster distance on the X axis against the total number of iterations used which is plotted on the Y axis. Following this we drew out a comparison between the precision accuracy and the performance of Vibrating Particle System algo against that of WWO algo. This was achieved by applying the two nature inspired clustering algorithms on the matching 6 data-sets for equal number of iterations. This was done in order to draw a realistic and a distinguishable comparison between the two algorithms ' accuracy which are being considered here.

Mentioned in the next section are the outputs along with the graphs which were produced for the two algorithms. When implemented individually on the multiple medical care data-sets

1.Thyroid-Dataset

Upon running the Water Wave Optimisation and the Vibrating Particle System algorithm on the BCW Healthcare data-set obtained from the UCI Repository the below mentioned results and output graphs were found

The following Accuracy-Matrix was produced:

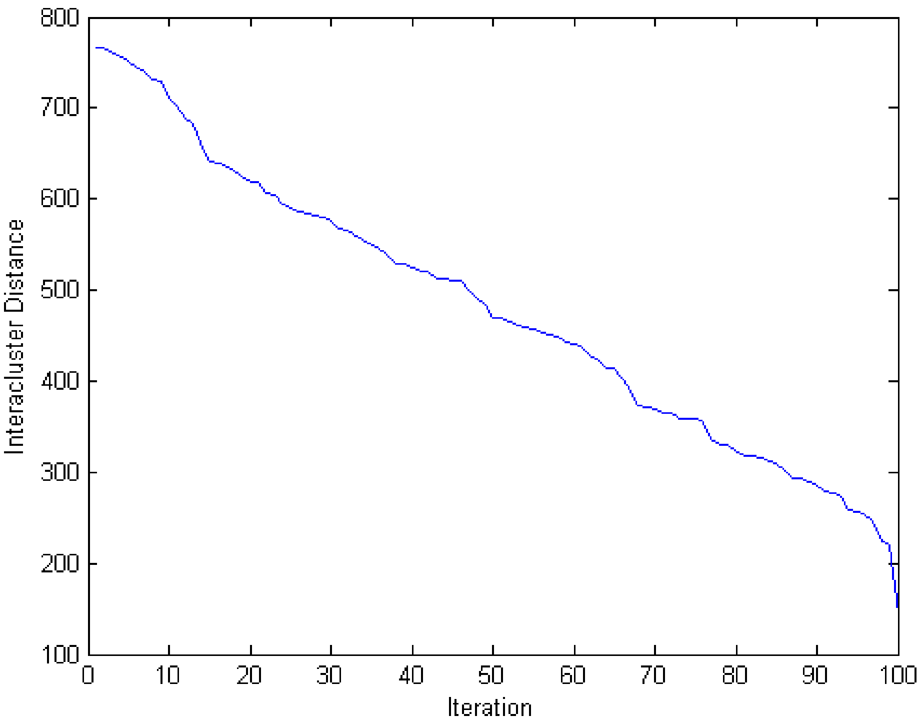
		85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
1	574	70.2326	71.1628	71.1628	71.1628	71.1628	72.0930	73.0233	73.0233	73.0233	73.9535	74.4186	76.2791	77.6744	78.1395	78.6047	80.4667

Fig 4.1 Accuracy-Matrix for thyroid dataset (VPS Algorithm)

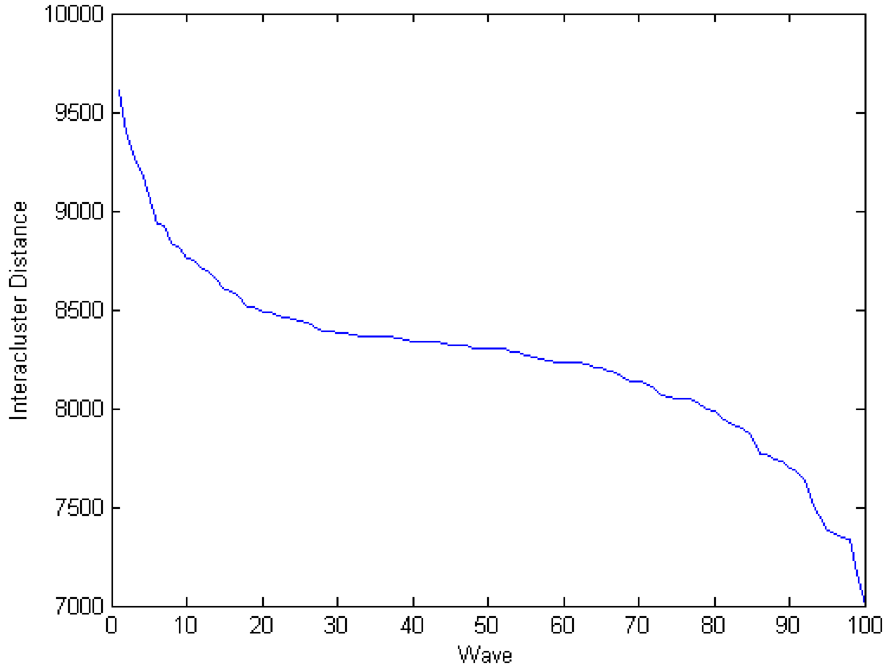
The following Distance-Matrix was produced:

	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 4.2 Distance-Matrix for Thyroid-Dataset(VPS Algorithm)



Graph 4.1 Performance of VPS (Thyroid-Dataset)



Graph 4.2 Performance of WWO (Thyroid-dataset)

After the implementation of the two nature inspired algorithms on the Thyroid disease dataset for total number of iterations equal to hundred we have finally got the below mentioned results.

Value of accuracy for the Thyroid-disease dataset turned out to be as mentioned:

For Vibrating Particle System: 80.36%

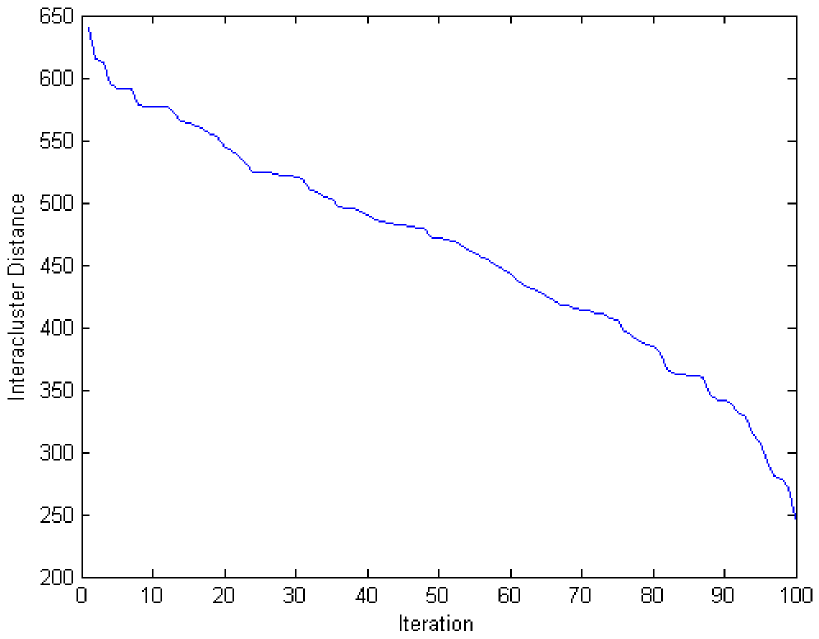
For Water Wave Optimization:82.69%

As depicted pretty evidently in graph, in Graph-4.1 and Graph-4.2 it is quite clear that Inter Cluster Distance is reducing as the number of iterations is increasing. Also one can take clear notice from both the listed graphs that initially we see a sharp and abrupt reduction in the Inter-Cluster Distance on the Y axis. This is because of one reason that is initially we pick the cluster centre in a random fashion because of which the inter cluster distance turns out to be is significantly high but then gradually as the total number of iterations increase along the X axis the cluster centres progressively come to the correct location. Because of this very reason the Inter-Cluster Distance is subsequently becoming less and less. Also we see from the output graph that beyond a certain specific no. of total iterations total reduction in the Inter-Cluster Distance also starts slowing-down as there is a very small amount of shift or drift in the location/position of the cluster-centre which eventually result in a low rate of reduction or decrease in the inter cluster distance. Also as previously mentioned in the prior sections, one can clearly note and see that with each and every iteration that is taking place the value of accuracy is also growing which in a way hints and points at the increasing trend.

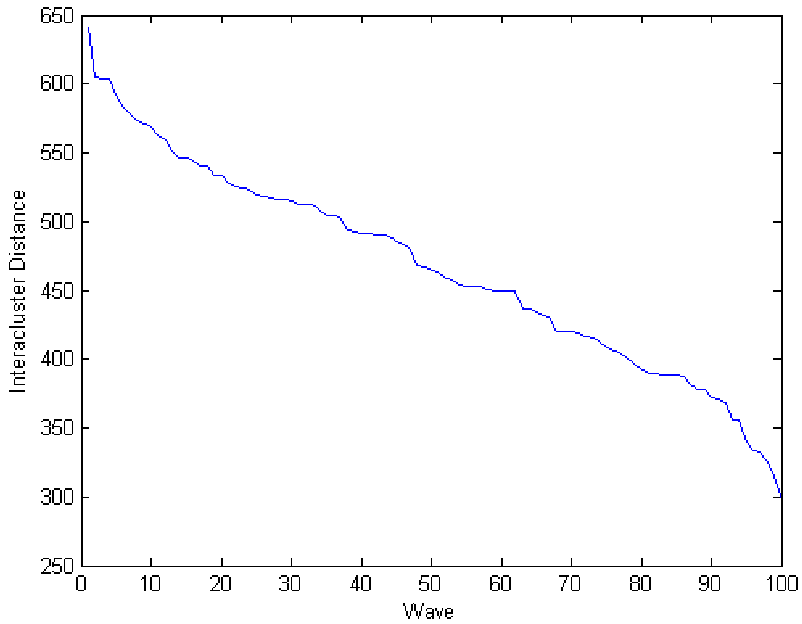
Also on taking a closer look at the value of accuracy obtained after the implementation of both these algorithms for the thyroid data set we happen to see that value of accuracy which we get for VPS-Clustering algorithm is around 80.36% whereas for the WWOthe accuracy which was obtained for a maximum of hundred iterations in was 82.69% which is clearly indicative of the very fact that for this particular dataset Water Wave Optimization algorithm's overall performance is somewhat better than the VPS algorithm's general performance. Having said that the variance in accuracy is not quite significant as it is very small and is only a mere 2.33%.

2. BCW-Dataset

Upon running the Water Wave Optimisation and the VPS algorithm on the BCW Healthcare data-set obtained from the UCI Repository the below mentioned results and output graphs were found



Graph 4.3 Performance of VPS (BCW-dataset)



Graph 4.4 Performance of WWO (BCW-dataset)

Post the implementation of the two nature inspired algorithms on the BCW Healthcare dataset for total no. of hundred iterations we have obtained the below mentioned results.

Value of accuracy for the BCW Healthcare dataset turned out to be as mentioned:

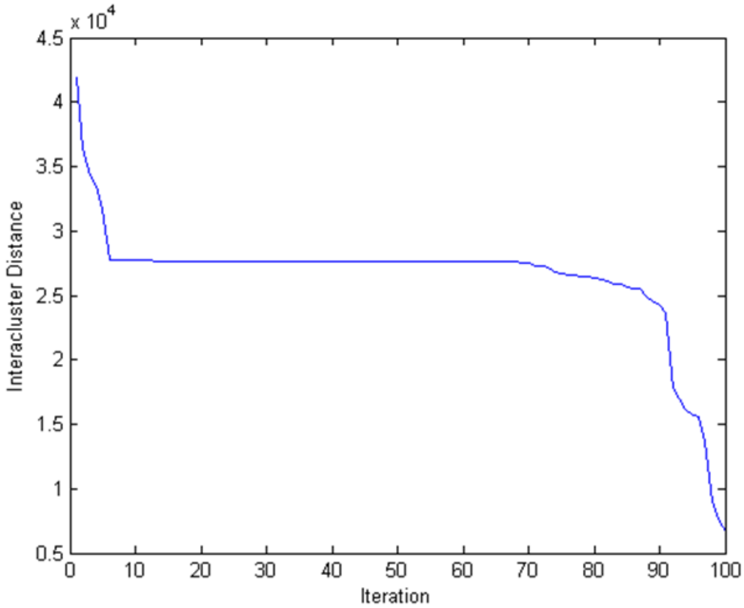
For Vibrating Particle System: 92.66%

For Water Wave Optimization: 81.23%

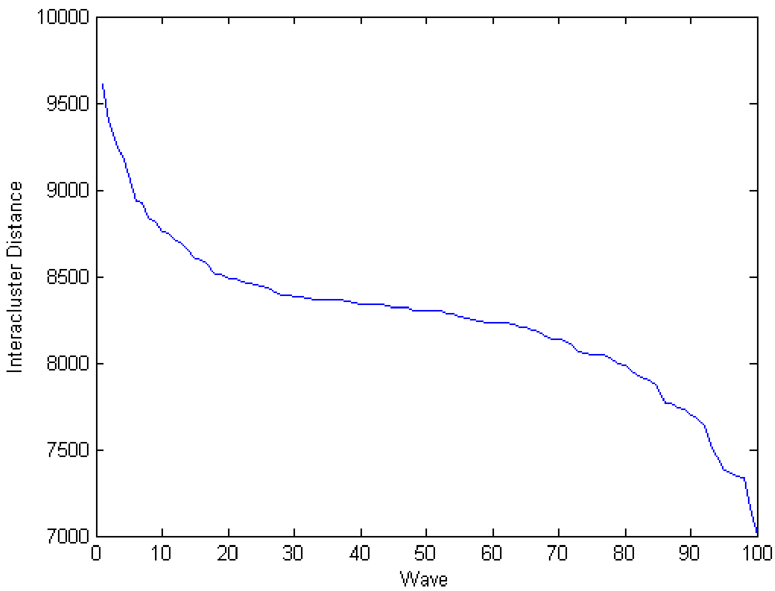
One can note from the Output-graph enlisted as the Graph- 4.3 and Graph-4.4, that Inter Cluster Distance which has been plotted along the X axis is falling with as the total number of iterations which has been plotted along the Y axis is increasing. The reason behind this particular trend has been stated previously in the last section. Another important thing to be considered is that the accuracy for the VPS-Clustering algorithm has come out to be an impressive 92.56% whereas the accuracy for the WWO is 81.23% which visibly shows that for this particular dataset of BCW overall the VPS algorithm's performance is notably a lot better in comparison to Water Wave Optimization algorithm's performance .

3. WDBC-Dataset

The preceding outcomes were produced when we ran the algorithm on the Bupa Disease Data-Set:



Graph 4.5 Performance of VPS (WDBC-dataset)



Graph 4.6 Performance of WWO (WDBC-dataset)

Post the implementation of the two nature inspired algorithms on the WDBC-dataset for a total no. of hundred iterations we have finally obtained the below mentioned results.

The value of Accuracy for the WDBC data-set turned out to be:

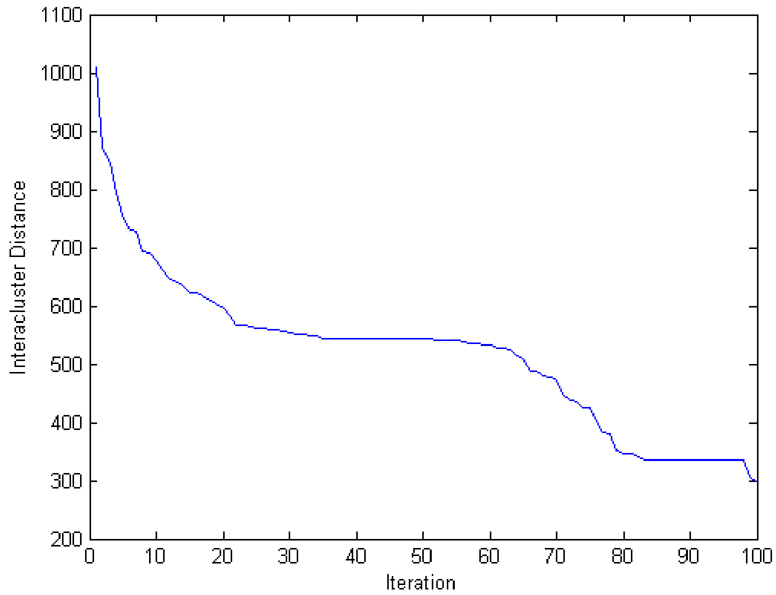
For Vibrating Particle System: 79.25%

For WWO: 81.29%

Also it is evident from the output graph as depicted in Graph-4.5 and Graph-4.6, that when the no. of iterations the inter-cluster distance is reducing. Cause of this trend seen in each of the output graph is the same and has been previously discussed in elaborate detail. Also as we can clearly see that the accuracy which is obtained for the VPS-Clustering algorithm is 79.25% whereas the accuracy for WWO- clustering algorithm is slightly higher than the former i.e. 81.29%. These values of accuracy clearly indicates that speaking specifically for the WDBC dataset the Water Wave Optimization algorithm's performance is marginally higher in comparison to Vibrating Particle System algorithm's performance .But also to be noted is the fact that the variance in the accuracy's value is not quite noteworthy as it is only a mere 2.04%.

4. Heart-Dataset

The preceding outcomes were produced when we ran the algorithm on the Heart Disease Data-Set:



Graph 4.7 Performance of VPS(Heart-dataset)

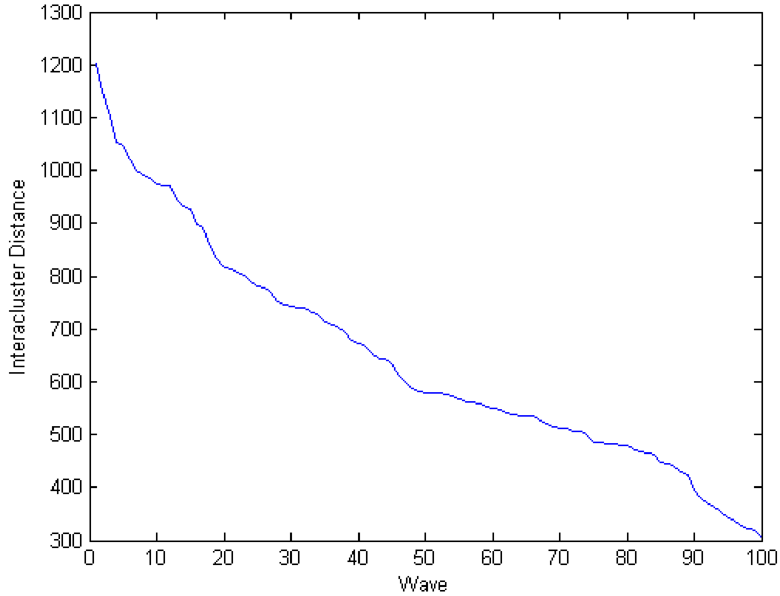


Fig 4.8 Output-Graph Produced for Water Wave Optimization (Heart-dataset)

Post the implementation of the two nature inspired algorithms on the Heart-dataset for a total no. of hundred iterations we have finally obtained the below mentioned results.

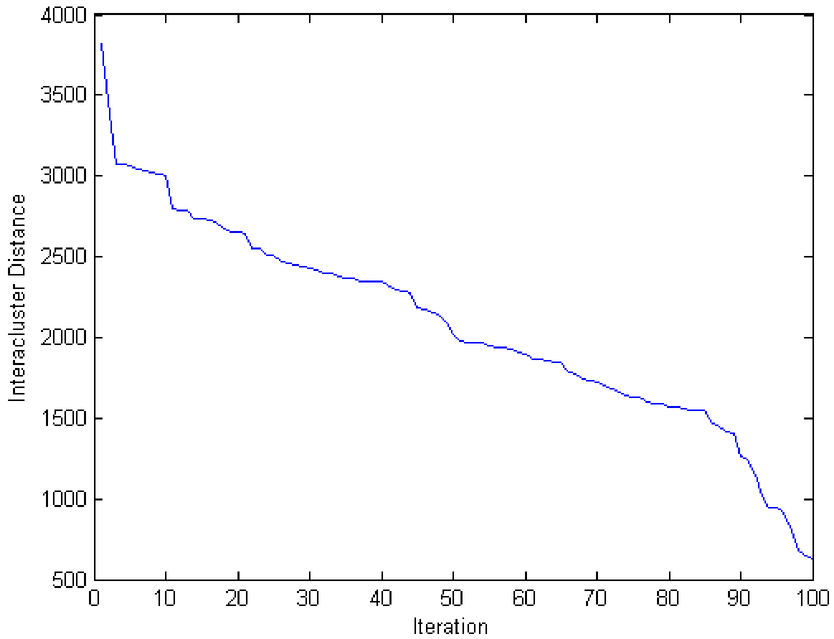
The value of Accuracy for the Heart dataset turned out to be:

For Vibrating Particle System: 62.58%

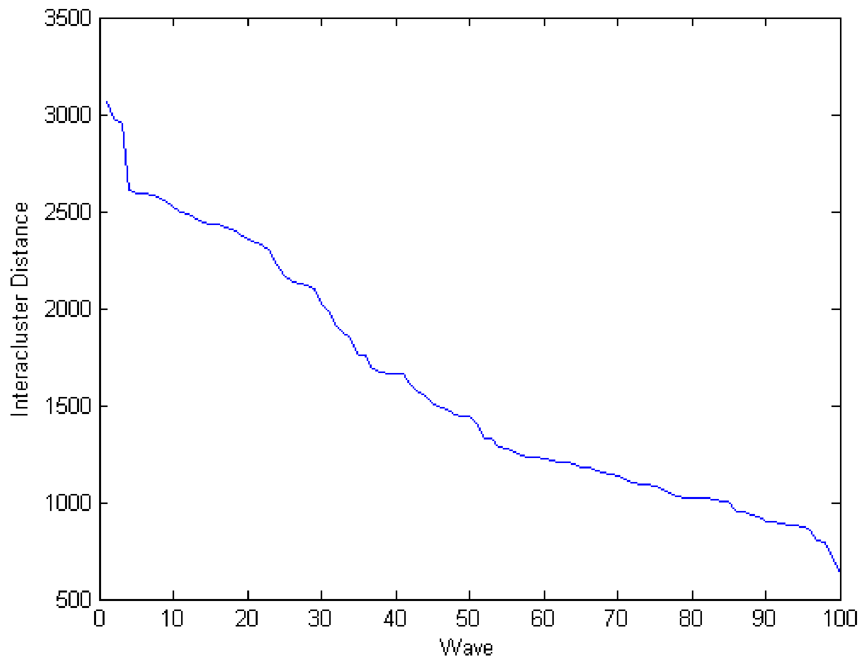
For Water Wave Optimization: 59.62%

5. BUPA Dataset

The preceeding outcomes were produced when we ran the algorithm on the Bupa Disease Data-Set:



Graph 4.9 Performance of VPS (BUPA-dataset)



Graph 4.10 Performance of WWO (BUPA-dataset)

Post the implementation of the two nature inspired algorithms on the BUPA data-set for a total no. of hundred iterations we have finally obtained the below mentioned results.

The value of Accuracy for the BUPA dataset turned out to be:

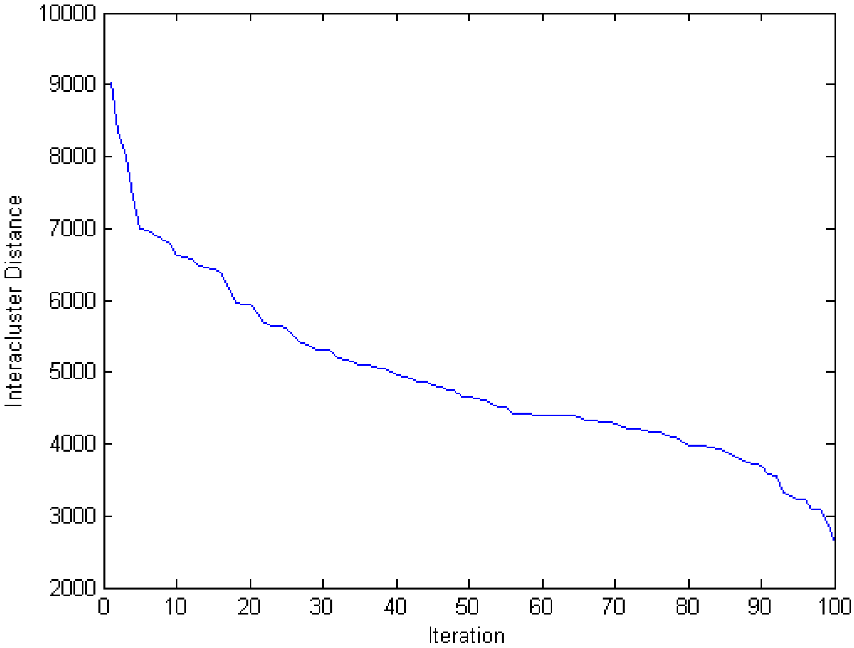
For Vibrating Particle System: 61.57%

For Water Wave Optimization: 69.69%

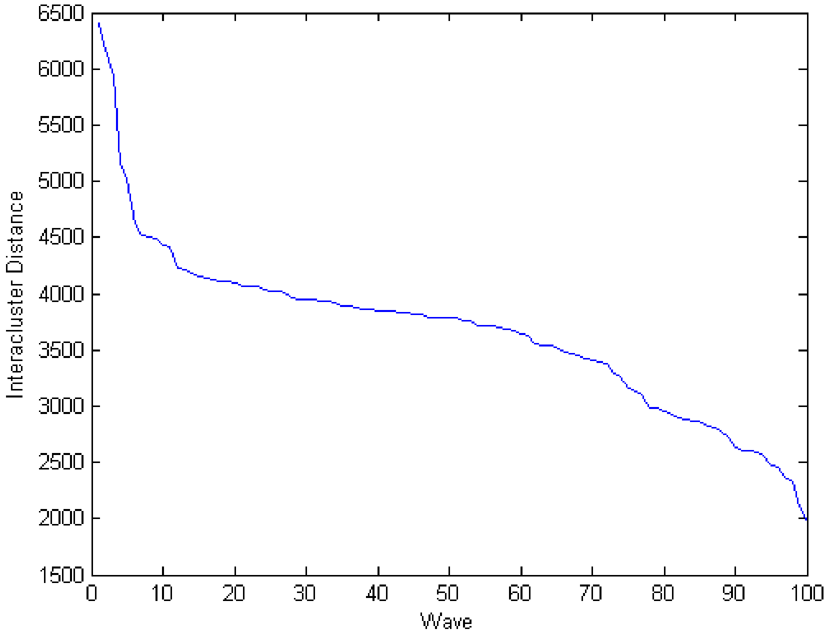
We can note from the output Graph- 4.9 as well as the Graph-4.10, the Inter-cluster- distance is becoming less and less with a rise in the total no. of iterations(Plotted along Y axis). Also to be noted is that the accuracy for the case of the VPS Clustering-algorithm is around the value of 60.57% while on the other hand the value of accuracy for WWO is 68.69% which points to the fact that Water Wave Optimization algo's performance is comparatively much better compared to the VPS algorithm's performance for the data-set with a value of difference in accuracy standing at a significant 8.12%.

6. Diabetes

After implementing both of the algorithms on the Diabetes Disease data-set the below mentioned graphs and results were obtained:



Graph4.11 Performance of VPS (Diabetes-dataset)



Graph4.12 Performance of WWO (Diabetes-dataset)

Post the implementation of the two nature inspired algorithms on the Diabetes disease data set for a total no. of hundred iterations we have finally obtained the below mentioned results.

The value of Accuracy for the Diabetes disease dataset turned out to be:

For Vibrating Particle System: 72.91%

For Water Wave Optimization: 74.35%

Graph- 4.11 and graph 4.12 show that the inter-cluster distance is becoming less and less with anrise in the total no. of iterations. Also evident from the output graphs visually is that at initial stages a sharp reduction in the inter- cluster distance is present. That's because we chose a random cluster centre at first due to which the distance is notably large but thereafter progressively the cluster centres tend to move towards the correct positions and hence the distance between clusters eventually decreases

CHAPTER 7 CONCLUSION

After the successful implementation of both the clustering algorithms i.e. The Vibrating Particle System Algorithm and the WWO on the six different medical data-sets, description of which has been mentioned previously. The observation that we have made by analysing the final outcomes is that there is no clear cut winner as far as performance comparison is considered. This is because for certain data-sets VPS performs comparatively better than WWO algorithm while for some of the heath care datasets WWO performs relatively better. So in order to predict which of the two algorithms is better from the other , we need to examine and compare the values of accuracy obtained for both the algorithms.

We observe that after implementation of the two algorithms on the six health care datasets it turns out that the VPS algorithm provides higher value of accuracy for two data-sets out of six which are Heart disease dataset and the BCW dataset. On the other hand the WWO algorithm provides higher value of accuracy for rest of the four medical datasets which are Thyroid dataset, the Diabetes dataset, BUPA dataset and also WDBC dataset.

The four out of six performance for the WWO algorithm clearly indicate that the Water Wave Optimization algorithm has a comparatively better performance for the Medical Datasets than the VPS algorithm.

The project can be further expanded and its future scope can be enhanced by making use of these two nature inspired clustering algorithms for Multi-partitioning-clustering and also for various real world prediction problems. This project holds great potential in the field of healthcare. Where it can be as be used to forecast whether a disease is present or absent in a patient in initial stages where prognosis is otherwise quite difficult and rare .If this becomes possible then it will revolutionize the entire healthcare industry.