A

Major Project-II Report

On

OPTIMIZATION OF DENSITY BASED CLUSTERING DBSCAN USING BAT ALGORITHM

Submitted in the Partial Fulfillment of the Requirement for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE & ENGINEERING

By

NEELAM SINGH

2K14/CSE/23

Under the Esteemed guidance of

Dr. KAPIL SHARMA

DELHI TECHNOLOGY UNIVERSITY

(Formerly Delhi College of Engineering)

Shahabad Daulatpur, Main Bawana Road,

Delhi -110042

JUNE, 2015

CERTIFICATE

This is to certify that Major Project-II Report entitled **"Optimization of Density Based Clustering DBSCAN using Bat Algorithm"** submitted by **Neelam Singh, Roll No. 2K14/CSE/23** for partial fulfillment of the requirement for the award of degree Master of Technology (Computer Science & Engineering) is a record of the candidate work carried out by her under my supervision.

Dr. Kapil Sharma
Associate Professor
Department of Computer Science & Engineering
Delhi Technological University

DECLARATION

We hereby declare that the Major Project-II work entitled "**Optimization of Density Based Clustering DBSCAN using Bat Algorithm**" which is being submitted to Delhi Technological University, in partial fulfillment of requirements for the award of degree of Master Of Technology (Computer Science and Engineering) is a bona-fide report of Major Project-II carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Neelam Singh
2K14/CSE/23

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. My greatest thanks are to my parents who bestowed ability and strength in me to complete this work.

I own a profound gratitude to my project guide Dr. Kapil Sharma who has been a constant source of inspiration to me throughout the period of this project. It was his competent guidance, constant encouragement and critical evaluation that helped me to develop a new insight into my project. His calm, collected and professionally impeccable style of handling situations not only steered me through every problem, but also helped me to grow as a mature person. I am also thankful to him for trusting my capabilities to develop this project under his guidance.

Secondly, I am grateful to Dr. O.P.Verma, HOD, Computer Science & Engineering Department, DTU for his immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Date: NEELAM SINGH

ABSTRACT

Clustering Algorithms are used for the task of classifying spatial databases and also in many other applications like data mining etc. It groups the points such that points within a single group have similar characteristics. There are many clustering algorithms are available for different type of applications. One of them is density based clustering DBSCAN that is used for identifying arbitrary shape of clusters based on their density along with noisy outliers. Secondly, recently many Bio inspired algorithms are used for solving the optimization problems and many other real world complex problems. Bat algorithm is one of the bio-inspired techniques used for solving optimization problems in various fields. It is basically inspired by the echolocation behavior of bats especially micro bats. Bat adjusts its frequency and wavelength accordingly to find its prey's position.

In this proposed work, hybrid of bat algorithm and DBSCAN is used to improve the cluster quality and also time complexity. For achieving this, first the best position of bats in search space is found out i.e. cluster center points, further it groups the other points using those cluster centers i.e. making the clusters according to their density using DBSCAN approach. Results of this work are improved intra cluster distance of clusters and also reduced time complexity of DBSCAN. It may take some extra time to calculate the best position i.e. cluster centers.

Keywords: Clustering, Bio-Inspired Algorithm, Bat Algorithm, and Density based Clustering, DBSCAN, Echolocation, and Data Mining.

List of Figures

List of Abbreviations

DBSCAN	:	Density based spatial clustering of applications with noise
Eps	:	Epsilon
BA	:	Bat Algorithm
minPts	:	Minimum points
BIA	:	Bio Inspired Algorithm

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 DATA MINING

Data mining is one of the important steps in "knowledge discovery in databases" process. This is the process of analyzing data from different sources and converting it into useful information. Information retrieved from this can be used in different type of applications.

Data mining tool is analyzing tool that helps users to analyze data from many different angles and dimensions, also categorize it and build up the relationships identified. Technically data mining is the process of finding similar type of data among dozens of fields in large and spatial databases. It also explores the large hidden predictive information from large databases. Other definition of data mining is the mining knowledge from the data. There is huge amount of data available in information industry and data mining is used to extract the relevant information from that data that is further used in processing of other applications such as Market Analysis, Fraud Detection, and Customer Retention etc. Data mining is very useful in different domains like Market Analysis and Management, Corporate Analysis & Risk Management and Fraud Detection etc. Data mining approach can be implemented rapidly on existing software and hardware platforms to improve the value of existing information resources, and can be integrated with new applications as they are brought on-line.

1.2 CLUSTER

In data mining, Cluster is the set of objects that belongs to the same group. In other words, objects with similar characteristics belong to one cluster and of different characteristics

belongs to other cluster. Definition of cluster varies according to the applications. This one is for the large type of data like spatial databases.

Cluster can also be said a subset of objects which are similar. A subset of objects such that the Euclidean distance or any other measure between any two objects in the cluster is less than the distance between any other object not in the cluster. A connected region of a multidimensional space contains a relatively higher number of objects.

1.3 CLUSTER DISTANCE MEASURE

Each clustering problem is solved on the bases of some kind of "distance" between points. Two type of distance measure are: Euclidean and Non-Euclidean. A Euclidean space consists of real- valued dimensions and points based on density. E.g. of this is L_1 norm, L_2 norm and L_∞ norm. A Non-Euclidean distance is based on properties of points, but not on their "location" in space. E.g. of this is Jaccard distance, Cosine distance and Edit distance. Most common type of distance measure that is used is Euclidean distance.

1.4 CLUSTERING

Recently, large amount of data is capturing market so handling that big data is real problem that information industries are facing. Basically, Clustering is a technique in data mining process to group relevant information according to their similarity. Clustering is unsupervised classification, it means no predefined classes. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Clustering helps users to understand the natural grouping or structure in a datasets. While doing clustering, we partition the set of data into groups based on data similarity and then assign the labels to these groups. Clustering analysis is broadly used in many applications such as market research, pattern

recognition, data analysis, and image processing. Clustering algorithms are important for the task of class identification in spatial databases. Points in one cluster have similar characteristics, while the points in different clusters are dissimilar. In market, there are many different type to techniques are available to solve clustering problems.

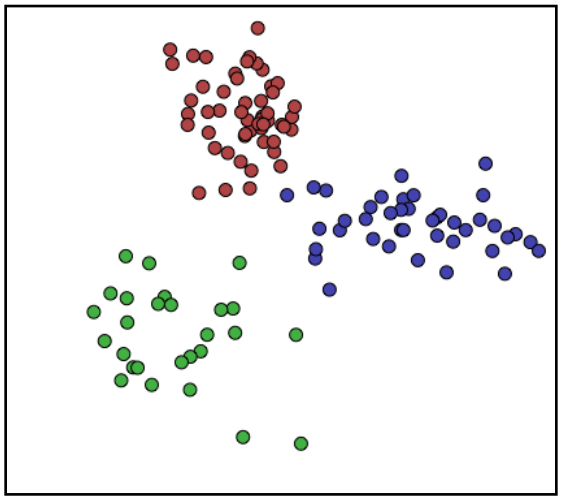


Figure – 1: Group of clusters

1.5 METHODS OF CLUSTERING

Clustering algorithms are classified into three types: partitional, hierarchical and density based clustering (1). Example of partitional clustering is K-Means which finds spherical shaped clusters only based on Euclidean distance. In Hierarchical clustering, clusters are created by merging the closest pair of clusters and in density based clustering; clusters are formed due to dense regions of data points and are separated by sparse regions with respect to given parameters. They can be of arbitrary form of clusters. DBSCAN and OPTICS are the most famous example of density based clustering.

1.6 APPLICATIONS OF CLUSTERING

DATA MINING

Clustering is the first step in data mining analysis. It identifies group of similar records that can be used further. With the help of other techniques, after determining the characteristics of such data group and that information can be used further.

SEARCH ENGINE

While querying on search engine, it returns a huge amount of related data. That is only possible if it groups all data of same type. If using the best clustering technique in search engine, it will group best related data and give best result on front page.

WIRELESS SENSOR NETWORK's BASED APPLICATION

These algorithms are very effective in wireless sensor network's based applications. It helps in finding the cluster head which collects the data from other clusters.

CANCER RELATED DATA

Clustering algorithms are also used on identifying the cancer cells from other normal cells. It can use supervised technique for this task and can use best algorithm for such kind of clustering.

1.7 DENSITY BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE (DBSCAN)

In density based clustering, clusters are formed due to dense regions of data points and are separated by sparse regions with respect to given parameters (2). They can be of arbitrary form of clusters. Some examples of density based clustering are DBSCAN and OPTICS. So for understanding DBSCAN, first understand some terms given below:

Eps- NEIGHBORHOOD OF A POINT

Eps neighborhood of a point means each point should have minimum number of points (MinPts) in its Eps-neighbor (1) such that

$$N\varepsilon\ (p) : \{q|d(p,q) <= \varepsilon\}$$

There are two kinds of points in clusters: first one is core points and other one is border points. Border points contain less number of points than Eps- neighborhood points.

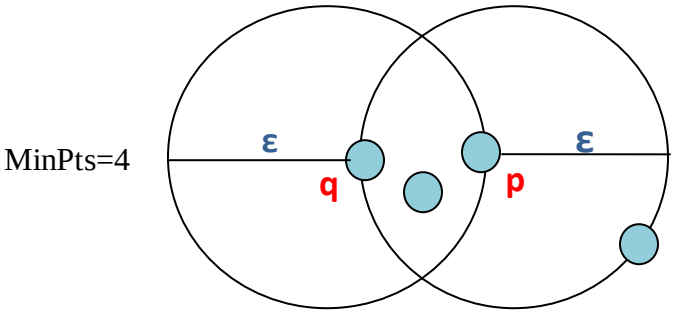


Figure -2: Eps-neighborhood of a point

A point is called a core point if it has more than a specified number of points (MinPts) within Eps - these are points that are at the interior of a cluster. A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point. A noise point is any point that is neither a core point nor a border point.

DIRECTLY DENSITY REACHABLE

An object q is directly density-reachable from object p if p is a core object and q is in p's Eps -neighborhood. i.e.

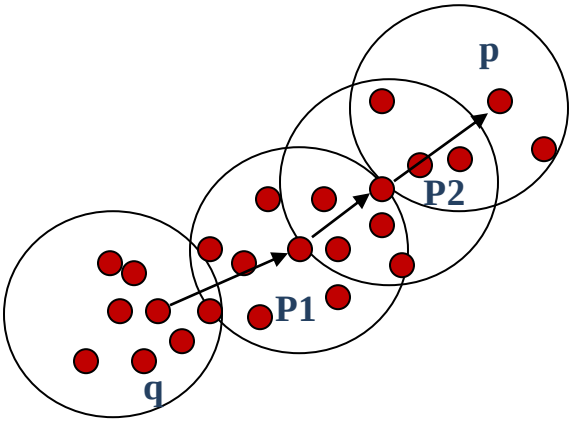
1. $p \in N_{Eps}(q)$ and
2. $|N_{Eps}(q)| >= MinPts$ (core point condition)

In above figure-1.1, q is directly density reachable from p but p is not directly density reachable from p. Density reachable is asymmetric.

DENSITY REACHABLE

A point p is density reachable from a point q wrt. Eps and MinPts if there is a chain of points from p to q such that a point p is directly density-reachable from p_n, p_n is directly density-

reachable from p_{n-1} and so on. Lastly p_1 is directly density reachable from q. i.e. $p \rightarrow p_{1\ldots} \rightarrow p_{n-1} \rightarrow p_n \rightarrow q$.



P is density reachable from q but q is not density reachable from q. (MinPts = 7)

Figure -3: Density Reachable Points

DENSITY CONNECTED

A point p is density connected to a point q wrt. Eps and MinPts if there is a point o such that both, p and q are density-reachable from o wrt. Eps and MinPts. Density-connectivity is a symmetric relation. For density reachable points, the relation of density-connectivity is also reflexive.

Intuitively, a cluster is defined to be a set of density connected points which is maximal wrt. density reachability.

DBSCAN ALGORITHM

Algorithm: **DBSCAN**
Input: Dataset, epsilon, minpts
Output: set of clusters

```
STEP 1:     DBSCAN (SetOfPoints, Eps, MinPts)

STEP 2:         ClusterId := nextId(NOISE);

STEP 3:         FOR i FROM 1 TO SetOfPoints.size DO

STEP 4:         Point := SetOfPoints.get(i);

STEP 5:             IF Point.ClId = UNCLASSIFIED THEN

STEP 6:                 IF ExpandCluster(SetOfPoints, Point, ClusterId, Eps,

                        MinPts) THEN   ClusterId := nextId(ClusterId)

STEP 7:                 END IF

STEP 8:             END IF

STEP 9:         END FOR
```

```
STEP 1:   ExpandCluster (SetOfPoints, Point, CiId, Eps, MinPts): Boolean;
STEP 2:   seeds: = SetOfPoints.regionQuery (Point, Eps );
STEP 3:   IF seeds.size < MinPts THEN
STEP 4:   SetOfPoint.changeClId (Point, NOISE);
STEP 5:       RETURN False;
STEP 6:   ELSE
STEP 7:       SetOfpoints.changeClIds (seeds, ClId);
STEP 8:       seeds.delete (Point);
STEP 9:       WHILE seeds <> Empty DO
STEP 10:          currentP := seeds.first();
STEP 11:          result := setofPoints.regionQuery(currentP, Eps) ;
STEP 12:              IF result.size >= MinPts THEN
STEP 13:              FOR i FROM 1 TO result.size DO
STEP 14:                  resultP := result.get(i);
STEP 15:                  IF resultP.ClId
                              IN (UNCLASSIFIED, NOISE)
                          THEN
STEP 16:                      IF resultP.ClId = UNCLASSIFIED
                              THEN
STEP 17:                          seeds.append (resultP);
STEP 18:                      END IF;
STEP 19:                      SetOfPoints. changeCiId ( resultP,
                          CiId);
STEP 20:                  END IF;
STEP 21:              END FOR;
STEP 22:              END IF;
STEP 23:          seeds.delete (currentP);
STEP 24:      END WHILE;
STEP 25:  RETURN True;
```

1.8 FLOW CHART OF DBSCAN

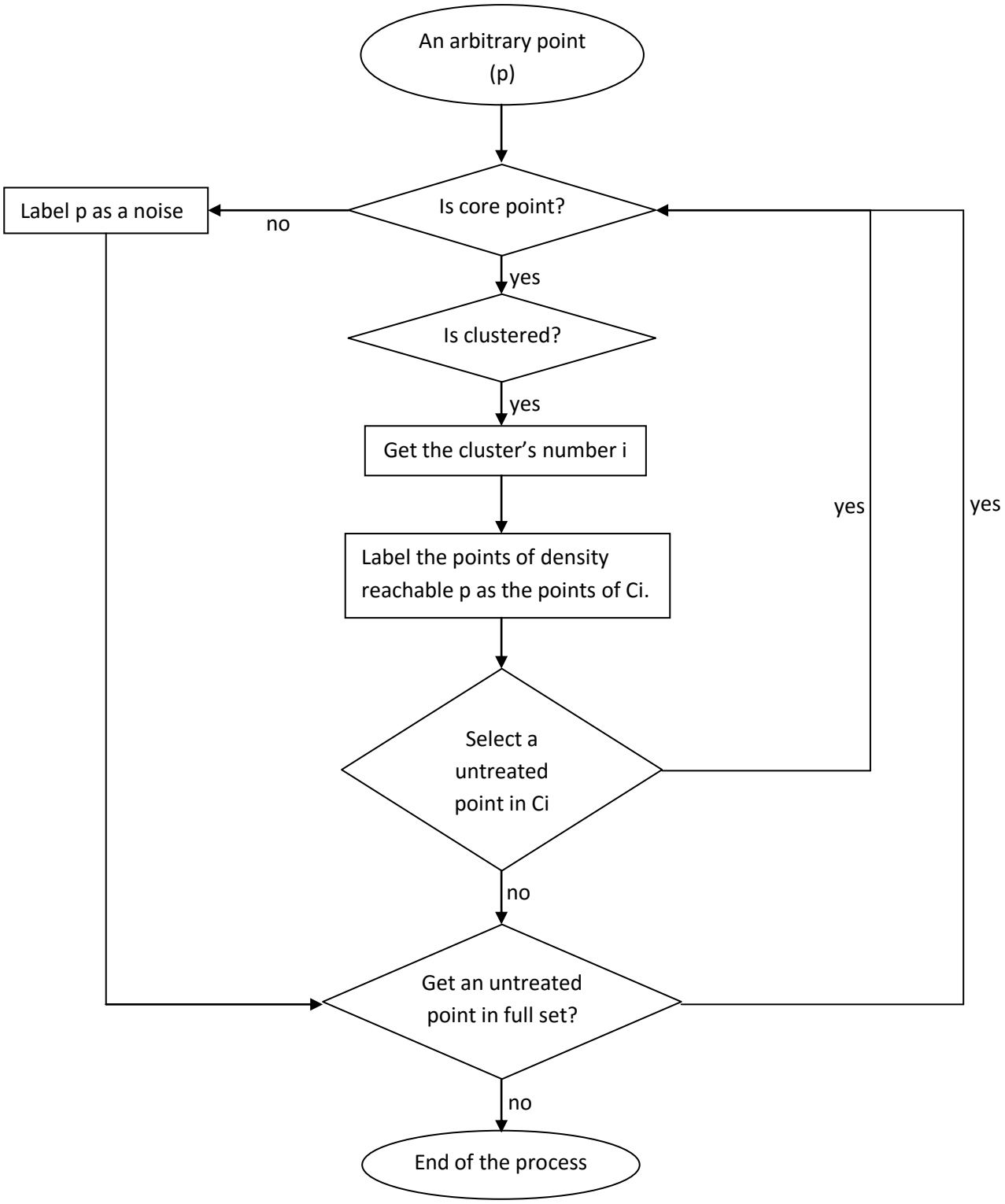


Figure 4 : Flow Chart of DBSCAN

1.9 BIO-INSPIRED ALGORITHM

Bio-Inspired algorithms are recently very interesting research topic in market. Bio inspired algorithms are inherently algorithms that nature has designed to solve computational problems. It is related to the field's biology, computer science and mathematics. Nature has become a great source of solving the hard and complex problems in computer science. Optimization is a commonly faced mathematical problem in all engineering disciplines. It exactly means finding the best desirable/ possible solution. Bio- Inspired algorithms are Meta heuristics that mimics the nature for solving the optimization problems. These kinds of algorithms are totally inspired by the nature. Every step of these algorithms is dependent on natural processes. Some of the examples of such kind of algorithms are Genetic Algorithms, Particle Swarm Optimization, Artificial Bee Algorithms and Bat Algorithms etc.

1.10 BAT ALGORITHM

Bat Algorithm is one of the bio-inspired algorithms that is recently introduced to solve varying optimization problems in different research fields. Basically it is based on the echolocation behavior of bats. Bat algorithm is developed by Xin-She Yang in 2010 (3).

Bats have the special ability to detect insects and avoid obstacles around themselves by using a high frequency sound based system echolocation. Bats emit sound waves with low frequency and high wavelength and listen to the returning echoes. From these, bats can sense their environment. Bats can distinguish the size, shape, and texture of a small prey, in which direction the prey is flying, and also the speed of the prey by using the delayed time and loudness of the response. Bats can change the way they emit the sound pulses. By changing frequency of the pulse, bats can change the traveling range of the pulses. Frequency is inversely proportional

to the wavelength λ and multiplication of these gives us the speed of sound where V = 340 m/s in air.

$$V = f\lambda$$

When bats fly to find their prey, they burst sound pulses with longer wavelength and lower frequency. It searches in large area covered by these waves. When bat detects any prey then it emits waves with high frequency and shorter wavelength. So that it can detect the exact position of the prey. Besides frequency and wavelength, bat can also change loudness and pulse rate of waves while approaching the prey.

Bat completes its search in two phases (4):

First phase is Exploration, when bat searches for prey; it emits sound waves with lower frequency and longer wavelength.

Second phase is Exploitation, when it detects the prey, it increases its frequency and lowers wavelength to find exact position of prey.

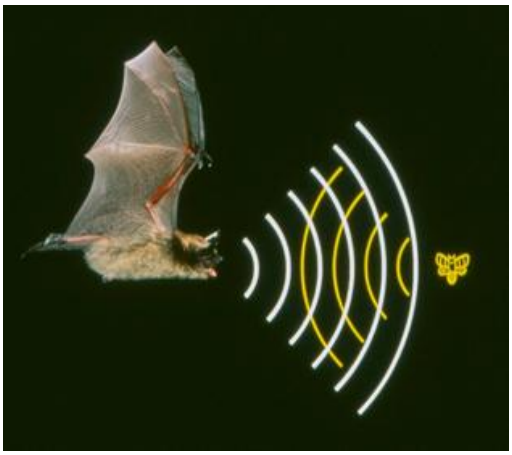


Figure – 5: Echolocation behavior of bats

Bats fly randomly during the search for prey with velocity V_i, with fixed sound pulse frequency f_i, varying wavelength λ, and loudness A_0. Loudness can vary from large value A_0 to minimum fixed value A_{min}. Besides these rules, frequency f varies in a range $[f_{min}, f_{max}]$.

1.11 APPLICATIONS OF BAT ALGORITHM

1.11.1 CLUSTERING ,CLASSIFICATION AND DATA MINING

Bat algorithms are applied on K-Means algorithm for clustering. Its performance is better than other bio-inspired algorithm like genetic algorithm, PSO etc. It can also be applied on other types of clustering. That is still a research area for improvement of performance of these algorithms.

1.11.2 IMAGE PROCESSING

Bat algorithm carries out the loads better than the other algorithms like PSO. Du and Liu (2012) proposed a variation of bat algorithm with mutation for photograph matching, and they proved that their bat-based system is efficient and viable in consider matching than other models of differential evolution and genetic algorithms.

1.12 MOTIVATION

Recently, online data is growing rapidly and handling that data is really a big issue. Many researches are going on how to handle and manage such big data. Some concepts are used in market like map-reduce algorithm and hadoop to manage such kind of data. But to extract important information from that data and driving pattern from this information is a time consuming task. So there is a big need for optimization in existing algorithms. Big data is a multidimensional data which is very complex to manage with the existing technology and algorithms. So this topic is very popular among the researchers. Datasets are growing day by day because of numerous types of information from mobile devices, social sites, blogs, radio channels etc. Nature is the big source of inspiration to solve the complex problems and handle the difficulties in big data. Many bio inspired algorithms are available for optimization. We use clustering to group the same type of data but this is really a difficult task to make out important

information from such a huge amount of data. Existing algorithms are not that much effective so we can apply bio inspired algorithms on the existing algorithms to optimize them to work efficiently and effectively.

Example- Particle swarm optimization is a bio-inspired algorithm and its application in clustering proves that how efficiently we can optimize the existing algorithms' processing.

All these interesting bio-inspired algorithms and the current problem of optimizing existing algorithms to handle big data inspired me to look into it. In this proposed work, optimization of the famous density based clustering DBSCAN using bio-inspired bat algorithm which is based on the echolocation behavior of bats will be done.

In Density based clustering, initially cluster centers are selected and their best position is found out by bat algorithm and then running DBSCAN on those centers. In this work, reduction of the time complexity of DBSCAN is also taking place. In this way, optimization of other clustering algorithms can also be done.

1.13 RESEARCH OBJECTIVE

In previous section motivation, we have already discussed that we are optimizing the density based clustering DBSCAN using bat algorithm. Another objective is we are also reducing its time complexity from n^2 to n. I am taking a real life dataset and organizing it into clusters according to its similarity and showing that the intra cluster distance is better in proposed algorithm than the previous one. It's time complexity is also improved one. In this work, I am first finding the cluster centers with the help of bat algorithm and the running the DBSCAN algorithm only for those cluster points. And also calculating the intra cluster distance and comparing that with the previous algorithm's clusters' intra cluster distance. This will show that the proposed algorithm has optimized the DBSCAN efficiently.

1.14 REPORT ORGANIZATION

I have started this report with the Chapter 1 - Introduction in which I tried to cover all the description and terms that are useful for this work. In Chapter-2, I covered history of the work and all the background details which are required for this work. Like, it has covered the concept of clustering and famous approaches for clustering and their applications and also the concept of bio-inspired algorithm and its implementation over clustering to improve its performance. In chapter-3, I tried to explain the concept of the bio-inspired bat algorithm which is based on the echolocation behavior of bats. Its algorithm is also explained with important parameters. Bat algorithm's applications are also discussed here. In chapter-4, I have explained the proposed work what I am doing in this research. Implementation and result comparison is discussed in chapter-5 and at the last I explained the conclusion and future work in Chapter-6.

CHAPTER 2

LITERATURE REVIEW

2.1 CLUSTERING AND BIO-INSPIRED HISTORY

Cluster analysis is the formal study of algorithms and methods for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity. Clustering is different from classification. Classification is supervised learning technique while clustering is unsupervised learning technique. The aim of clustering is to find structure in data. Clustering has a long and rich history in a various scientific fields. There are many clustering algorithms available in market for different type of applications since mid of 20^{th} century. One of the most popular clustering algorithms is K-Means. The aim of cluster analysis, also known as data clustering, is to discover the natural grouping of a set of points, patterns, or objects. Taxonomists, social scientists, psychologists, biologists, statisticians, mathematicians, engineers, computer scientists, medical researchers, and others who collect and process real data have all contributed to clustering methodology. According to JSTOR data clustering first appeared in the title of a 1954 article dealing with anthropological data. Data clustering is also known as typology, Q-analysis, taxonomy, and clumping depending on the field where it is applied. Clustering algorithms can be divided into three groups: partitional, hierarchical and density based clustering. Partitional clustering algorithms form all the clusters simultaneously as a partition of the data. The most famous and simple partition base algorithm is K-Means. Hierarchical clustering algorithms group the data at each level and form the clusters in form of hierarchy. The most well-known hierarchical algorithms are single-link and complete-link. Density based clustering algorithms groups the data based on the dense region and separate the

groups based on sparse region. The most popular density based algorithms are DBSCAN and OPTICS.

Computer science and biology has a common and shared long history. For many years, to process and analyze biological data (e.g. microarrays), computer scientists have designed algorithms, and likewise, biologists have discovered several operating principles that have inspired new optimization methods (e.g. neural networks). Recently, these two directions have been converging based on the view that biological processes are inherently algorithms that nature has designed to solve computational problems. Most common bio-inspired algorithms are genetic algorithm, PSO, Ant Colony Optimization, Artificial Bee Colony Algorithm and Bat Algorithm etc. Bat Algorithm is developed by Xin-She Yang in 2010. Basically, this works on echolocation behavior of bats. Bio-inspired algorithms are the hot topic in market to research. The first most successful algorithm was genetic algorithm proposed by Holland in 1975. The real beauty of nature inspired algorithms lies in the fact that it receives its sole inspiration from nature. These algorithms are widely used in optimization problems and are very successful in that field. Below figure shows many optimization techniques.

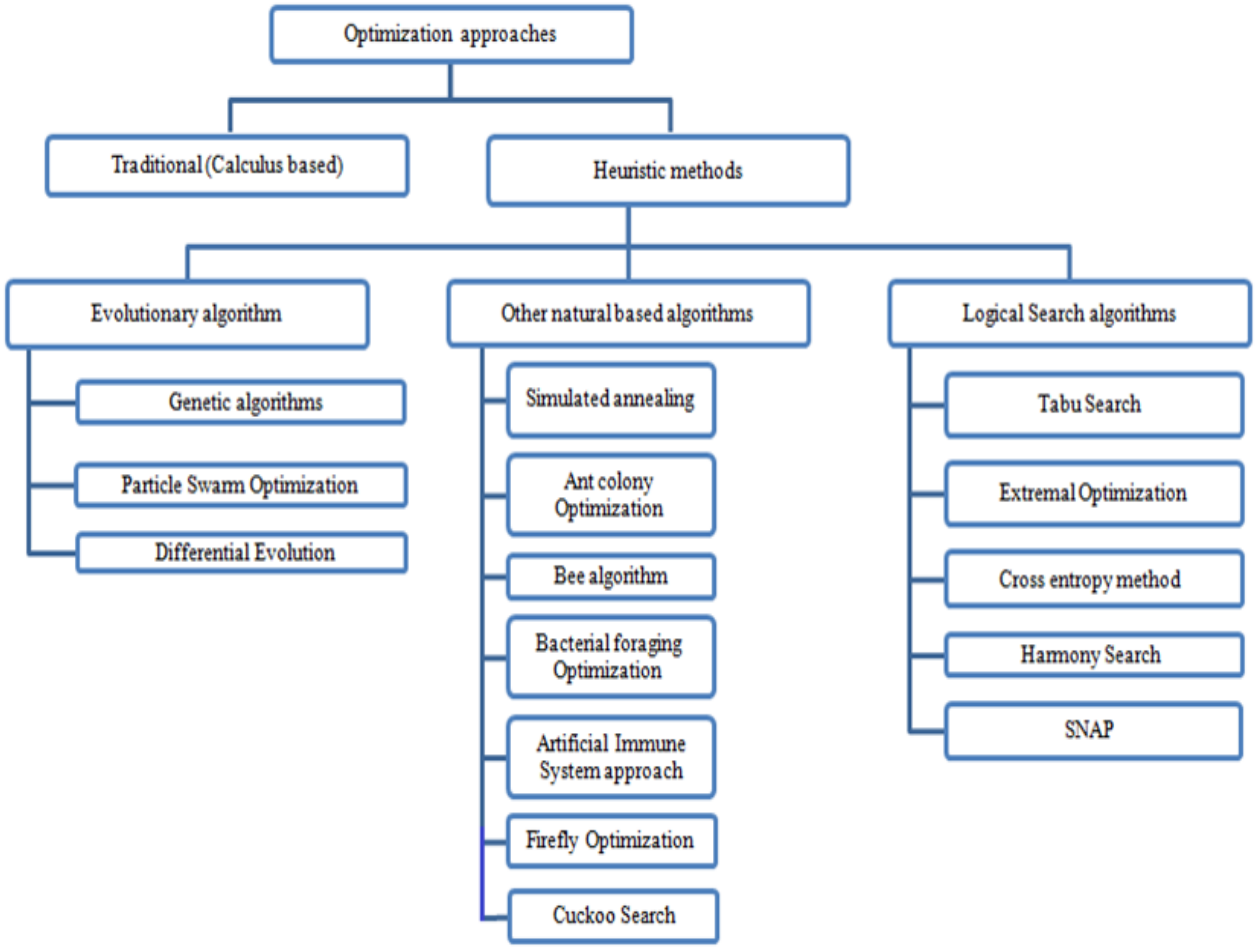


Figure 6: Optimization approaches hierarchy

2.2 CLASSIFICATION, CLUSTERING AND DATA MINING

Classification and Clustering have become an increasingly popular method of multivariate analysis over the past two decades. Optimization is very common problem in information industries. To solve this many clustering algorithms are proposed and many are under research work. Optimization algorithms can be either deterministic or stochastic in nature. Classification is supervised and clustering is unsupervised. Clustering is an important step in data mining that is to get relevant information from huge amount of data. Former methods of solving optimization problems require huge computational efforts, which tend to fail for big data. That is the reason for employing bio inspired stochastic optimization algorithms as computationally

efficient alternatives to deterministic approach. Komarasamy and Wahi (2012) studied bio-inspired bat algorithm and applied it on K-Means clustering and they concluded that the combination of both bat algorithm and K-means can perform efficiently than other available algorithms. So in research many bio- inspired algorithms are applied on clustering.

2.3 DATA CLUSTERING APPROACHES

2.3.1 HIERARCHICAL CLUSTERING

The methods build the clusters either in top-down or in bottom-up manner. This type of clustering consists of various methods. One of them is Agglomerative hierarchical clustering in which initially each object acts like a cluster and later these clusters are merged successively until the desired cluster structure is obtained. Other one is divisive hierarchical clustering in which initially, all the objects belong to one cluster. Then the cluster is divided into sub-clusters, which are further divided into their own sub-clusters. Until the desired cluster structure is obtained, this process continues. Output of these methods is shown by dendrogram. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level. On the basis of similarity, merging or division of clusters is performed.

2.3.2 PARTIONAL METHOD

Partitioning methods changes the position of instances by moving them from one cluster to another. Such methods typically require that the number of clusters will be pre-set by the user. In these algorithms number of cluster centers is known earlier and points change their position according to their distance from the cluster centers. This algorithm try to minimize the intra cluster distance and maximize the inter cluster distance of clusters. K-Means is the most popular and very successful algorithm in this category. Many optimization algorithms are applied over this type of clustering.

2.3.3 DENSITY BASED CLUSTERING

Density based clustering forms the cluster which are dense region of point and are separated by sparse region with respect to given density parameters (1). Clusters formed by this algorithm are of arbitrary types. DBSCAN and Optics are two example of this type of clustering. In Density-based methods, we consider an assumption that the points that belong to each cluster are drawn from a specific probability distribution. Cluster grows as long as its density in the neighborhood exceeds some threshold. In DBSCAN, we set the parameter minPts that is minimum number of points in the neighborhood of the given radius should be greater than that minPts value. We should choose parameters in this algorithm such that the probability of the data being generated by such clustering structure and parameters is maximized. ε-neighborhood of a point in DBSCAN is the set of points in radius ε from that point. Initially, this algorithm selects the arbitrary unvisited point and if its ε-neighborhood contains at least δ points, it is added to the cluster. Then, it expands the cluster by adding more points with the ε-neighborhood of the already added points. This process is repeated if there are points left in the process.

2.3.4 MODEL-BASED CLUSTERING METHOD

Model-based clustering methods attempt to optimize the fit between some mathematical models and the given data. Model-based clustering methods find characteristic descriptions for each group, where each group represents a cluster or class. The most frequently used Model-based clustering methods are decision trees and neural networks. In decision trees, hierarchical tree is used to represent data, where each leaf refers to a class and contains a probabilistic description of that class. Several algorithms produce classification trees for representing the unlabelled data. In Neural Network, each cluster is represented by a neuron or “prototype”. It also represents input data by neurons, which are connected to the prototype neurons. Each connection of such type has a weight, which is learned adaptively during learning.

2.3.5 GRID BASED METHODS

In Grid based methods, we use partition of the space into the finite number of cells that form a grid structure and we perform all the operations of clustering over that. These algorithms are used to making fast the algorithms.

2.3.6 SOFT COMPUTING METHODS

In soft computing methods, we use neural networks in clustering tasks. In earlier clustering approaches we generate simple partitions; in a partition, each object belongs to one and only one cluster. Such a clustering is classed hard clustering and all the clusters are disjoint. Fuzzy clustering is the good example of soft computing. In this each object is associated with every cluster with some membership function.

2.4 APPLICATIONS OF CLUSTERING

2.4.1 DATA MINING

Clustering is one of the important steps in data mining analysis. For exploring further relationships, it creates the group of related data. After getting this information, other process in data mining can create a structure of the information from such cluster that can be used further. We can also say that this is the first and most important step in data mining.

2.4.2 SEARCH ENGINE

Clustering is important in search engine because when there is a query that is related to some specific information. Clustering is used to group that information and can display the most relevant information on the front page and least relevant on other pages.

2.4.3 PATTERN RECOGNITION

Pattern recognition is basically the categorization of data into classes. It is classified based on similarity among the data. So clustering plays a very important role in the pattern recognition. It observes the environment and makes decisions based on the category of patterns.

2.4.4 IMAGE ANALYSIS

Clustering is used in image analysis for high level description of the image content. It classifies the image into classes that provide the information about that image and we can use this information for further processing or in some other applications.

2.4.5 BIOINFORMATICS

Clustering plays a very important role in bioinformatics because it group the tissues that are similarly affected by a disease. It can also group the patients of the similar disease. It groups related functional genes and from that information scientist can build regulatory networks and can discover subtype of disease, etc.

2.4.6 MACHINE LEARNING

Machine learning is related to design and development of Algorithms that allow computers to change behaviors based on empirical data, such as from sensor data or databases. Major research in machine learning is to recognize complex patterns and make decision with the help of that data. Clustering is used to group the information of the same type.

2.4.7 IMAGE PROCESSING

Clustering in image processing is used for high level description of image content. Image information is divided among the classes that can be further used in different task related to image database management.

2.4.8 TEXT MINING AND OTHER APPLICATIONS

In text mining, high quality of information is retrieved through the patterns which can be obtaining by clustering of that text data. Text mining is basically the process for structuring the input text and driving patterns within the structured data. Clustering is also used in many other applications like web cluster engines, weather report analysis etc. Nature behaves like the optimizer to solve these problems. We can also solve NP-hard problems with the solutions provided by nature.

2.5 BIO-INSPIRED ALGORITHMS

Nature is the source for solving many complex and difficult problems and optimizing them. Bio- inspired algorithms mimic the nature and solve the many problems with optimization. First bio-inspired algorithm was genetic algorithm developed in 1975. It maintains the proper balance among its component while solving the complex problem. These algorithms only includes biological components of the nature i.e., humans and animals. Natural components learn from the environment and make changes according to the condition. This is the key idea of the bio-inspired algorithms. These algorithms adopt the natural behavior for solving the complex problems easily and efficiently. Bio- inspired algorithms are categorized in three category- Evolution, Swarm based and ecology. Evolution algorithms contain genetic algorithms, genetic programming, evolutionary strategies etc. Swarm based algorithms have particle swarm optimization, ant colony optimization, artificial bee colony algorithms etc. and the last one ecology algorithms contains BBO, AWC, PS2O etc as shown in below figure.

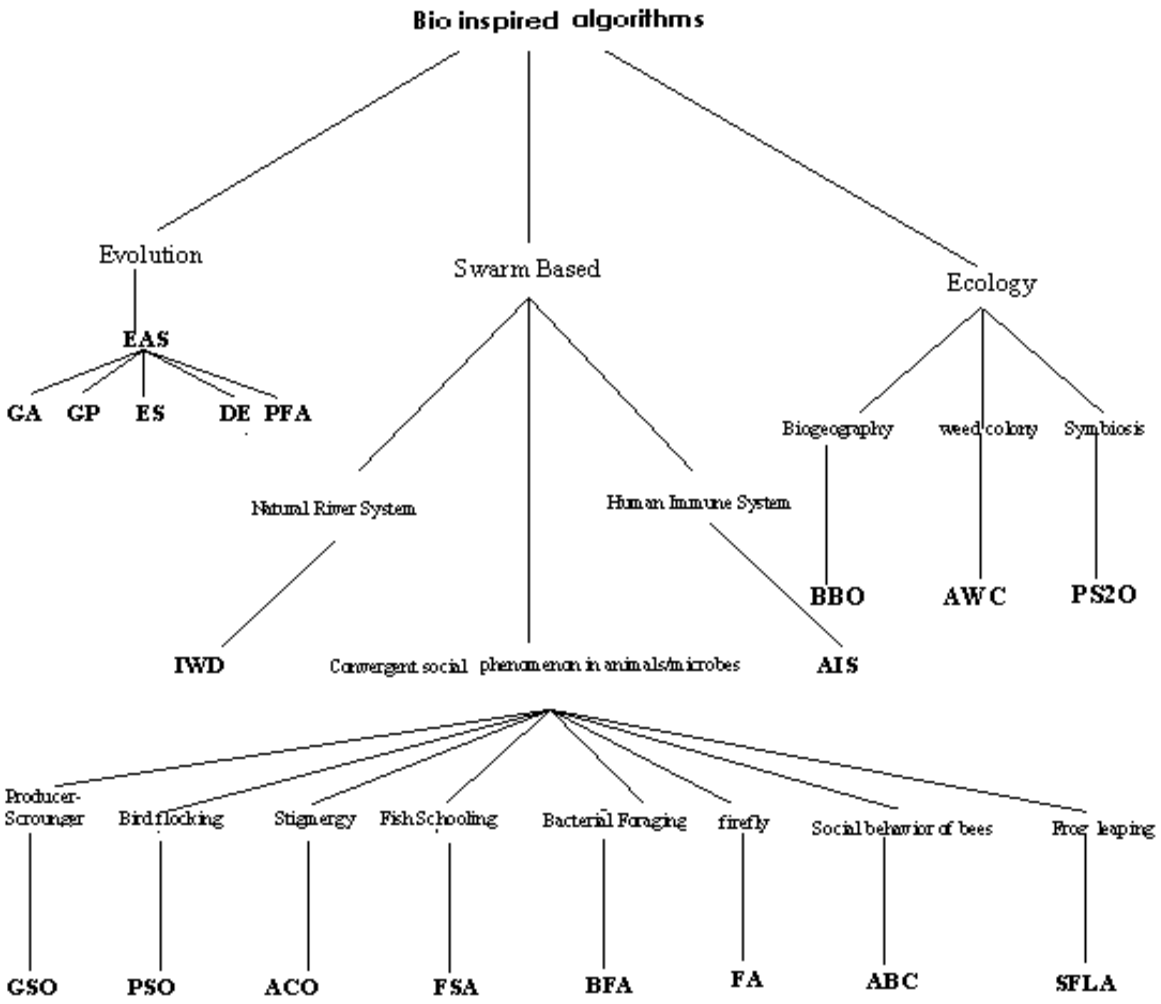


Figure – 7: Taxonomy and nomenclature of various bio inspired optimization algorithms

2.6 VARIOUS TYPES OF BIO-INSPIRED ALGORITHMS

2.6.1 GENETIC ALGORITHMS

GA is an evolutionary based stochastic optimization algorithm that is part of bio-inspired algorithm with a global search approach proposed by Holland in 1975. This algorithm solves both constrained and unconstrained optimization problems. This algorithm is very popular and successful and is inspired by evolutionary ideas of natural selection. In this algorithm, first of all population initialization takes place and then for each chromosome, fitness value is calculated

using fitness function and then best chromosome is selected for mating pool. Then crossover and mutation take place. Mostly it is used in solving global optimization problems (5).

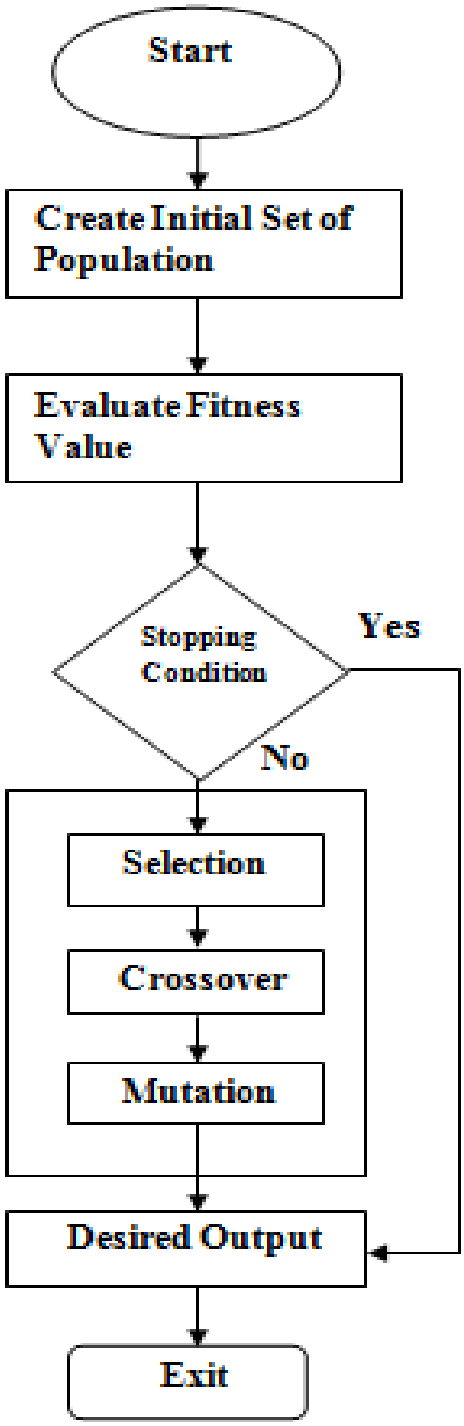


Figure 8: Genetic Algorithm Flow Chart

2.6.2 PARTICLE SWARM OPTIMIZATION

Particle swarm optimization (PSO) is a stochastic, computational intelligence oriented; popul-ation based global optimization technique proposed by Kennedy and Eberhart in 1995 (6). It is basically based on the bird flocking behavior for searching food. It is applied on many optimization areas because of its simple and different searching approach. In PSO, we refer population members as particles. Each particle in the swarm represents a solution in a high-dimensional space with four vectors, 1st one is its current position, 2nd is best position found so far, 3rd is the best position found by its neighborhood so far and last one is its velocity and changes its position in the search space based on the best position found by itself and on the best position found by its neighborhood during the search process.

2.6.3 ANT COLONY OPTIMIZATION

Ant Colony Optimization is proposed by Dorigo & Di Caro in 1999 and this is most successful algorithm among swarm based algorithms (5). ACO is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. It is a meta heuristic technique inspired by the foraging behavior of ants known as stigmergy. Stigmergy refers to the indirect communication amongst a self-organizing emergent system via individuals modifying their local environment. The most important aspect of the collaborative behavior of ant species is their ability to find shortest paths between the ants 'nest and the food sources by following pheromone trails Then, ants choose the path to follow by a probabilistic decision dependent on the amount of pheromone: the stronger the pheromone trail, the higher its desirability.

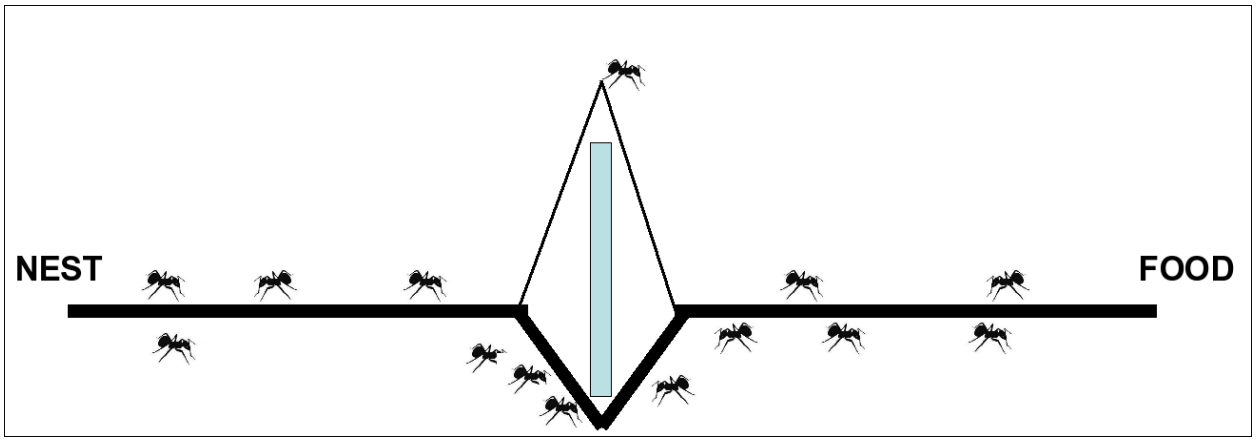


Figure 9: Ant Colony Optimization

2.6.4 ARTIFICIAL BEE COLONY ALGORITHM

ABC is simulating the intelligent foraging behavior of a honeybee swarm, proposed by Karaboga and Basturk (5). In ABC algorithm, the colony of artificial bees contains three groups of bees: employed bees, onlookers and scouts. A onlooker bee waits on the dance area for making a decision to choose a food source and one going to the food source visited by it before is named employed bee. The other kind of bee is scout bee searches for discovering new sources. The position of a food source represents a possible solution to the optimization problem and the nectar amount of a food source corresponds to the quality (fitness) of the corresponding solution.

Initially, some (scouts bee) fly and choose the food sources randomly without using experience. If the nectar amount of a new source is higher than that of the previous one in their memory, they memorize the new position and forget the previous one. Whole algorithm is based on this phenomenon.

2.6.5 FISH SWARM ALGORITHM

Fish Swarm Optimization Algorithm (FSOA) inspired by the collective movement of the fish and their various social behaviors (7). This is based on stochastic search. Based on a series of instinctive behaviors, the fish demonstrate intelligent behaviors by maintaining their colonies.

Immigration, searching for food and dealing with dangers all happen in a social form and interactions among the fish in a group will result in an intelligent social behavior. This algorithm has many advantages including high convergence speed, fault tolerance, flexibility and high accuracy. Initially fish is at its current position and it also has some visual position. If the state at visual position is better than the current state, it goes forward a step in that direction and arrives the next state; otherwise, continues in the vision scope. For finding the better overall states of the vision, it has to inspect tour efficiently. It does not need to move to infinite states, which helps to find the global optimum solution.

2.6.6 INTELLIGENT WATER DROPS ALGORITHM

Intelligent Water Drops is population based method proposed by Hamed Shah-hosseini in 2007 (7). It is inspired by the processes in natural river systems having the actions and reactions that take place between water drops in the river and the changes that happen in the environment when river is flowing.

This Intelligent Water Drop has two important characterisitcs:

1. The velocity that it is moving now, Velocity (IWD).

2. The amount of the soil it carries now, Soil (IWD).

IWD algorithm is composed of two parts: a graph that works as distributed memory for preserving the soils of different edges, and the moving part of the IWD algorithm, which is a few number of Intelligent water drops. These Intelligent Water Drops (IWDs) both compete and cooperate to find better solutions and by changing soils of the graph, the paths to better solutions become more reachable. IWD-based algorithms need at least two IWDs to work.

2.6.7 FIREFLY ALGORITHM

Firefly algorithm is proposed by Yang and it is considered as an unconventional swarm-based heuristic algorithm for constrained optimization tasks (7). It is inspired by the flashing

behavior of fireflies. The algorithm contains a population-based iterative procedure with many agents (perceived as fire flies) together solving a considered optimization problem. Agents communicate with each other via bioluminescent glowing which enables them to explore cost function space more effectively than in standard distributed random search. Intelligence optimization technique is based on the assumption that solution of an optimization problem can be obtained as agent (fire fly) which glows in proportion to its quality in a considered problem setting. Consequently each brighter fire fly attracts its partners (regardless of their sex), which makes the search space being explored more efficiently.

2.7 BIO-INSPIRED ALGORITHMS USES IN CLUSTERING

Bio- inspired algorithms are very useful in optimization problems. Bio-inspired algorithms are applied on many problems including NP-hard problems and gave very positive results in improving their performances. Many bio-inspired algorithms are applied on clustering to improve its performance. Some of them are genetic algorithms, particle swarm optimization, ant colony algorithms, artificial bee colony algorithms etc. All of them are giving very positive results. They also improve the time complexity of the algorithms of clustering. Bio-inspired algorithms use the natural methods to find the solution in the best and efficient way.

Many researches are going on to apply bio- inspired algorithms in most of the fields of clustering. In this work, I am also optimizing density based clustering DBSCAN using bat algorithm which is inspired by the echolocation behavior of bats.

CHAPTER 3

BAT ALGORITHM

3.1 PARAMETERS OF BAT ALGORITHM

Bat Algorithm works on the echolocation behavior of bats. Bats are blind and they find their prey by emitting sound waves with low frequency and longer wavelength. Then, they hear the echoed waves and sense the exact location of prey and the background. Initially, Bats fly randomly during the search for prey with velocity V_i, with fixed sound pulse frequency f_i, varying wavelength λ, and loudness A_0. Loudness can vary from large value A_0 to minimum fixed value A_{min}. Besides these rules, frequency f varies in a range $[f_{min}, f_{max}]$.

Initially for the i th bat, every parameter is determined randomly and later to update them we follow some equations (4):

$$f_i = f_{min} + (f_{max} - f_{min})\,\beta,$$

$$V^t{}_i = V^{t-1} + (x^t{}_i - x_*)\,f_i,$$

$$x^t{}_i = x^{t-1}{}_i + V^t{}_i,$$

Where $\beta \in [0,1]$, a random vector and x_* is the current global best position at time step t. For local search phase, local solution is generated by random walk.

$$x_{new} = x_{old} + \varepsilon A^t,$$

Where $\varepsilon \in [0,1]$ is a random number, $A_t = \langle A^t{}_i \rangle = 1/N \sum^N{}_{i=1} A^t{}_i$ is the average loudness of all the bats and N is the total number of bats.

Now loudness A_i and the rate of pulse r_i changes according to below equations:

$$A^{t+1}{}_i = \alpha A^t{}_i\ ,$$

$$r^{t+1}_i = r^0_i [1 - \exp(-\gamma t)] ,$$

Where α and γ are constants. We usually take its value 0.9 to 0.98. For any $0 < \alpha < 1$ and $\gamma > 0$, we have,

$$A^t_i \to 0, r^t_i \to r^0_{i,} \text{as } t \to \infty$$

3.2 FLOW CHART OF BAT ALGORITHM

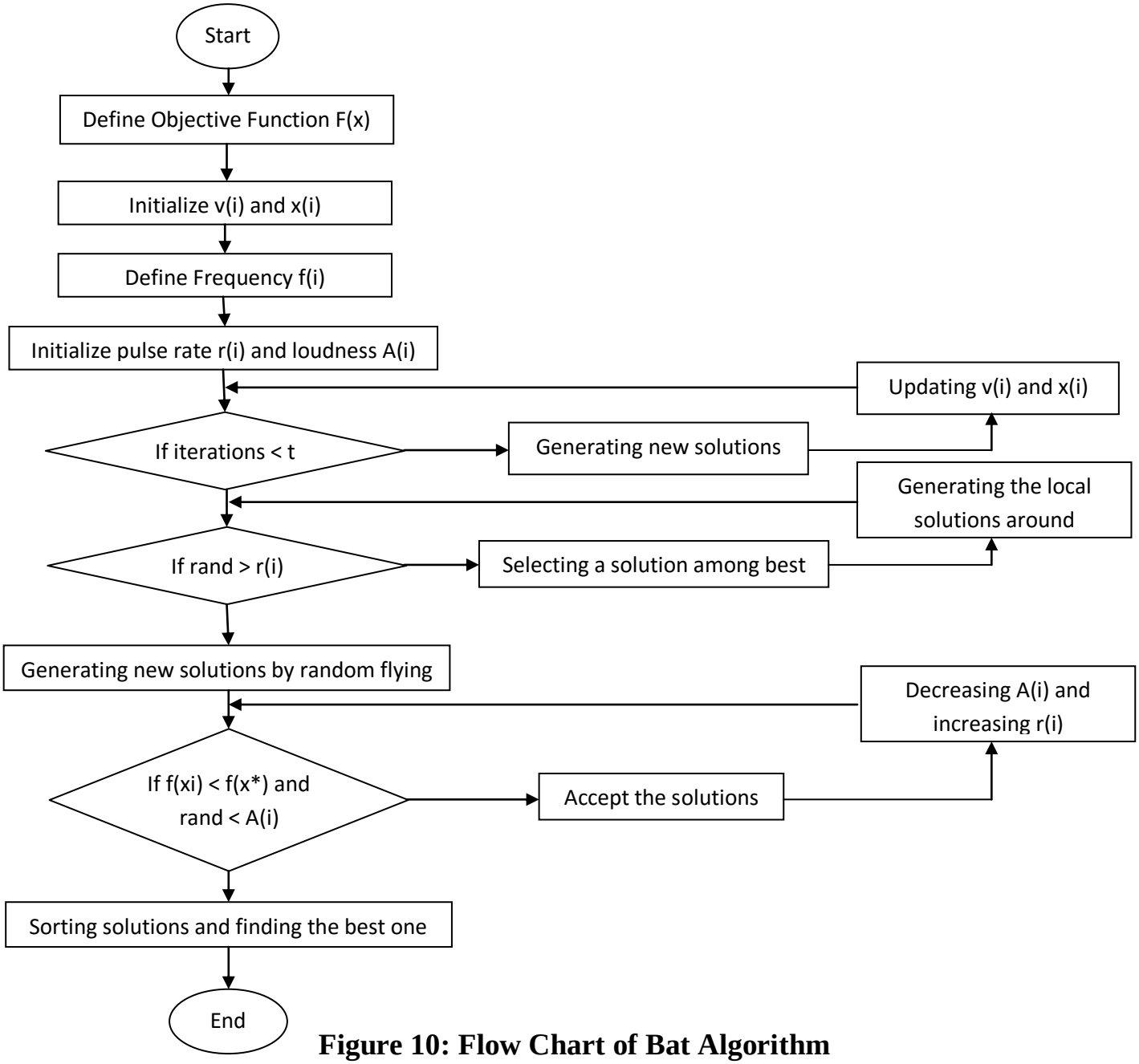


Figure 10: Flow Chart of Bat Algorithm

3.3 BAT ALGORITHM

STEP 1:	Objective function f(x) , $x = (x_1 , . . . , x_d)^T$
STEP 2:	Initialize the bat population $x_i , v_i (i = 1 , 2 , . . . , n)$
STEP 3:	Define pulse frequency (f_i) at x_i
STEP 4:	Initialize pulse rates r_i and the loudness A_i
STEP 5:	while (t < Max number of iterations) do
STEP 6:	Generate new solutions by adjusting frequency,
	and updating velocities and locations/solutions
STEP 7:	if (rand > r_i) then
STEP 8:	Select a solution among the best solutions
STEP 9:	Generated a local solution around the selected best solution
STEP 10:	end if
STEP 11:	Generate a new solution by flying randomly
STEP 12:	if (rand < A_i & f (x_i) < f (x_*)) then
STEP 13:	Accept the new solutions
STEP 14:	Increase r_i and reduce A_i
STEP 15:	end if
STEP 16:	Rank the bats and find the current best x_*
STEP 17:	end while

In the main loop, bats find the global best position and moves towards that position and they also changes their loudness and rate of pulses after finding global best position.

3.4 VARIANTS OF BAT ALGORITHM

The basic bat algorithm has many advantages, and one of the main advantages is that it can be converged very quickly by switching from exploration to exploitation stage. This makes it an efficient approach for many applications such as classifications and also when a quick solution is needed. However, if in this algorithm, the switch from exploration to exploitation stage is taking place too quickly by varying A and r often, it may lead to stagnation after some iterations. In order to improve the BA's performance, many approaches have been attempted to increase and enhance the performance, which developed a few good variants of bat algorithm.

Some of the variants of BA are as follows:

- **Fuzzy Logic Bat Algorithm (FLBA):** Khan et al. (2011) introduced a variant by applying fuzzy logic into the bat algorithm; they called this variant, fuzzy bat algorithm. This is better than the standard BA and Genetic Algorithm.

- **Multiobjective bat algorithm (MOBA):** Yang (2011) proposed one variant of BA to deal with multiobjective optimization (8). It uses simple weighted sum with random weights. It solves the nonlinear, global optimization problems and multiobjective design problems.

- **K-Means Bat Algorithm (KMBA):** Komarasamy and Wahi (2012) proposed a combination of K-means and bat algorithm (KMBA) for efficient clustering. In this it randomly selects clusters and then find the best one among all and converges to global best.

- **Chaotic Bat Algorithm (CBA):** Lin et al. (2012) presented a chaotic bat algorithm that is used in solving integer programming problems (9). The proposed algorithm

uses chaotic behavior to find a candidate solution in behaviors similar to acoustic monophony.It is using chaotic maps and L´evy flights to carry out parameter estimation in dynamic biological systems.

- **Binary bat algorithm (BBA):** Nakamura et al. (2012) introduced a discrete version of bat algorithm to solve classifications and feature selection problems (10). The wrapper approach combines the power of exploration of the bats together with the speed of the Optimum-Path Forest classifier to find the set of features that maximizes the accuracy in a validating set.

- **BAT- FLANN:** Sashikala et al. in (2012) proposed BAT Algorithm embedded with FLANN to solve the classification of gene expression data. Meta- heuristic Framework was designed with this model and Protein Structure Prediction was also done by this and later optimized by FLANN network.

- **Differential Operator and L´evy flights Bat Algorithm (DLBA):** Xie et al. (2013) introduced a variant of bat algorithm using L´evy flights and differential operator to solve function optimization problems (9).

- **Directed Artificial Bat Algorithm (DABA):** Rekaby in Aug 2013 proposed Directed Artificial Bat Algorithm (11). This algorithm explains the echo system of that bats, and how they use this echolocation system in prey finding and obstacle avoidance.

- **Improved bat algorithm (IBA):** Jamil et al. (2013) extended the bat algorithm with a hybrid of L´evy flights and subtle variations of pulse emission rates and loudness (12). They tested the IBA versus over 70 different test functions and proved to be very efficient.

- **Dynamic Virtual Bat Algorithm (DVBA):** Ali Osman Topal in Mar 2016, proposed this algorithm in which only two bats handle the complete search including exploration and exploitation phase by changing wavelength and frequency accordingly (4).

There are other variants and improvements of bat algorithm are also available.

3.5 APPLICATIONS OF BAT ALGORITHM

3.5.1 CONTINUOUS OPTIMIZATION

Continuous optimization demonstrated that BA can handle nonlinear problems efficiently and can find optimal solutions easily. Case studies include car side design, pressure vessel design, truss systems, spring and beam design tower and tall building design and others.

3.5.2 SCHEDULING

Bat Algorithms can be used in wide area of problem solving. Musikapun and Pongcharoen (2012) solved multi-stage, multi-product, multi-machine sche-duling problems using bat algorithm, and they solved a class of non-deterministic polynomial time (NP) hard problems with a wide and detailed parametric study.

3.5.3 COMBINATORIAL OPTIMIZATION

Combinatorial problems are really hard, often non-deterministic polynomial hard (NP-hard). Ramesh et al. (2013) presented a detailed study of combined emission dispatch and economic load problems using bat algorithm. They compared bat algorithm with hybrid genetic algorithm, ant colony algorithm (ABC) and other methods and they found that bat algorithm can be implemented easily and efficiently on various applications.

3.5.4 CLASSIFICATION AND CLUSTERING

Komarasamy and Wahi (2012) presented the hybrid of K-means clustering and bat algorithm and they concluded that the combination of both K-means and BA can achieve higher

efficiency and thus performs better than other algorithms like genetic algorithms, PSO etc. On the other hand, Mishra et al. (2012) used bat algorithm to classify microarray data.

3.5.5 IMAGE PROCESSING

Abdel-Rahman et al. (2012) studied full body human pose estimation using bat algorithm, and they concluded that BA performs better than particle filter (PF), particle swarm optimization (PSO), and annealed particle filter (APF). Later Du and Liu (2012) proposed a variant of bat algorithm with mutation for image matching, and they concluded that their bat- based algorithm is more feasible and effective in imagine matching than other models such as differential evolution and genetic algorithms.

3.5.6 FUZZY LOGIC AND OTHER APPLICATIONS

Reddy and Manoj (2012) used bat algorithm in a study of optimal capacitor placement for loss reduction in distribution systems. It combines with fuzzy logic in finding optimal capacitor sizes so they can minimize the losses. Their results showed that the real power loss can be reduced significantly. There are also many other applications available that used fuzzy logic with bat algorithm.

.

CHAPTER 4

PROPOSED WORK

4.1 PROBLEM STATEMENT

The algorithm DBSCAN groups the points on the basis of density around the points and having some parameters. But their cluster's quality cannot be measured by intra cluster distance because there is no proper set of cluster centers. DBSCAN algorithm runs for every point and finds its neighbors in given radius and if they are greater than MinPts. So it's time complexity is n^2. We want to improve this time complexity as well as intra cluster distance of clusters in density based clustering DBSCAN.

Bat algorithm is a bio- inspired algorithm that is used in optimization in many applications. What Bat algorithm can do in DBSCAN:

1. It will select the cluster points from the datasets randomly.
2. Then this tries to find the best possible solutions by random walk.
3. After selecting the cluster centers, we will run the DBSCAN only for cluster centers rather than running this for all the points.
4. Then we will calculate the intra cluster distance of all the clusters.
5. We will also see that now DBSCAN will run only for cluster centers rather than every point that will reduce its time complexity too.

4.2 PROPOSED WORK

In this work, I am optimizing the density based clustering DBSCAN with the help of Bat Algorithm. Suppose we have a dataset of 1000 points and it has two clusters division based on density. Now we want to divide these 1000 points into two clusters based on the density of these clusters as shown in below figure:

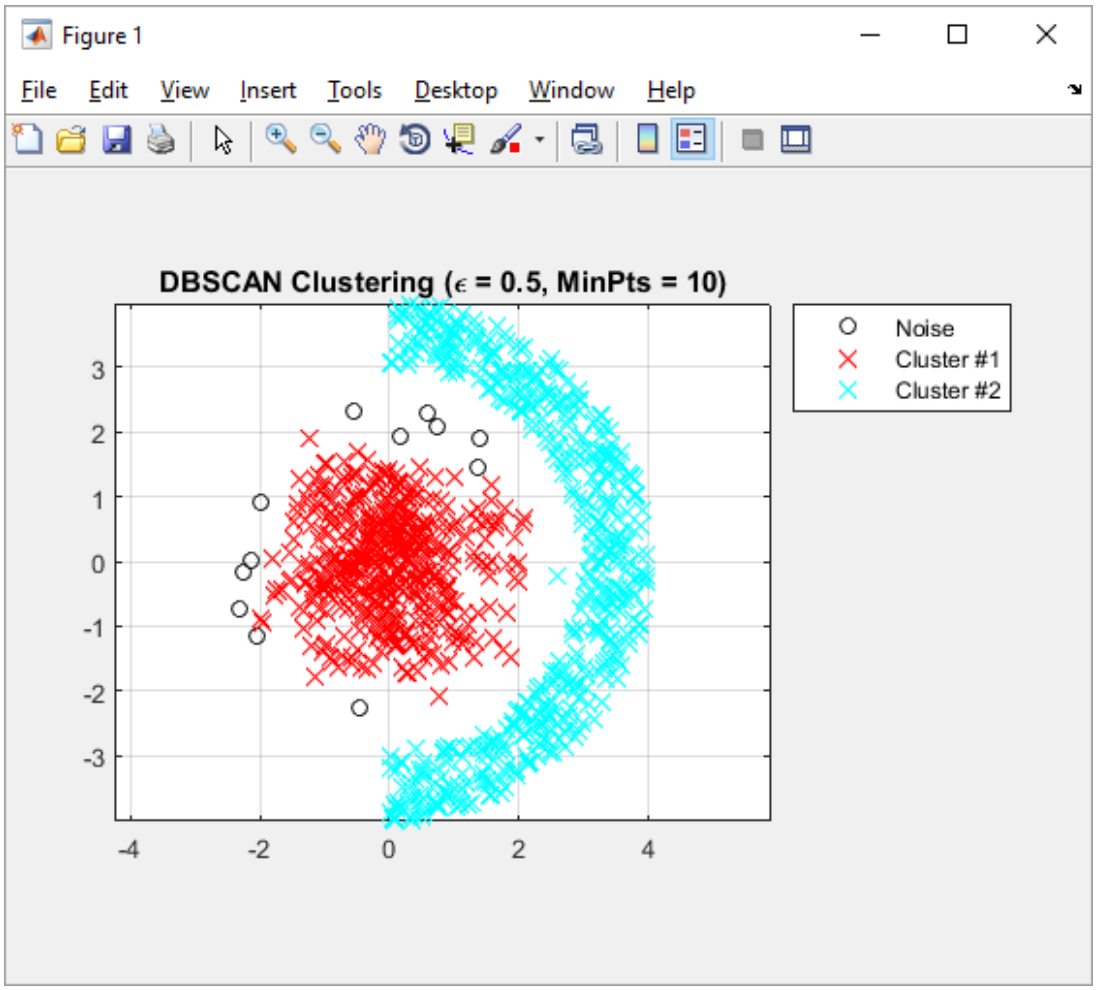


Figure 11: Clusters based on density

In our algorithm, we will first choose the cluster centers from the given datasets then we will find the best centers near those selected centers using bat algorithms. Now we will run DBSCAN only on those centers. Its cluster quality will get improved from the DBSCAN as we are getting minimized intra-cluster distance of the clusters.

4.3 FLOW CHARTS

I am using the approach that will optimize the DBSCAN algorithm and also will reduce the intra cluster distance of clusters. Now I am going to draw the flow chart of my algorithm that will help to understand my work easily.

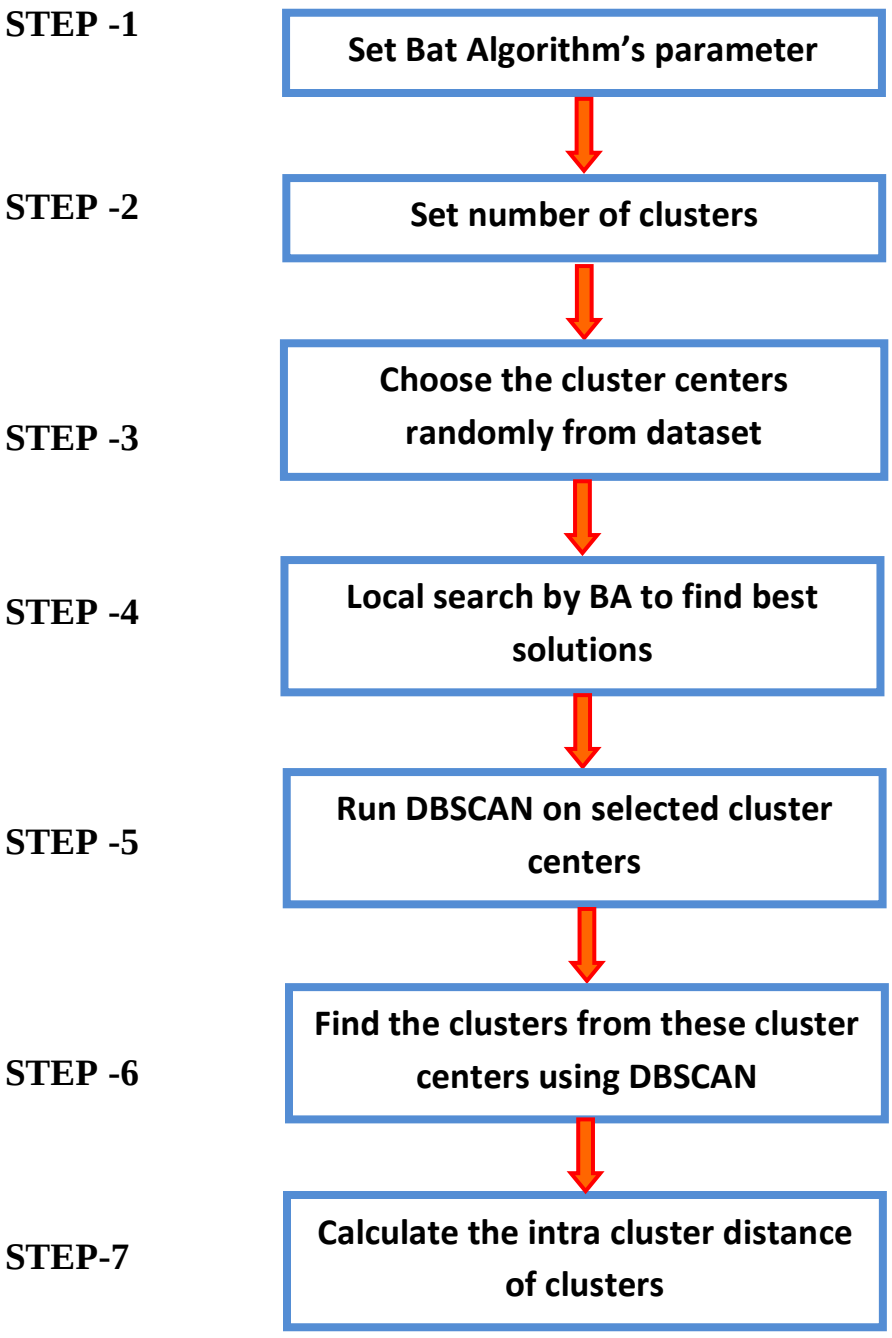


Figure 12: Flow Chart of proposed work

4.4 ARCHITECTURE OF PROPOSED ALGORITHM

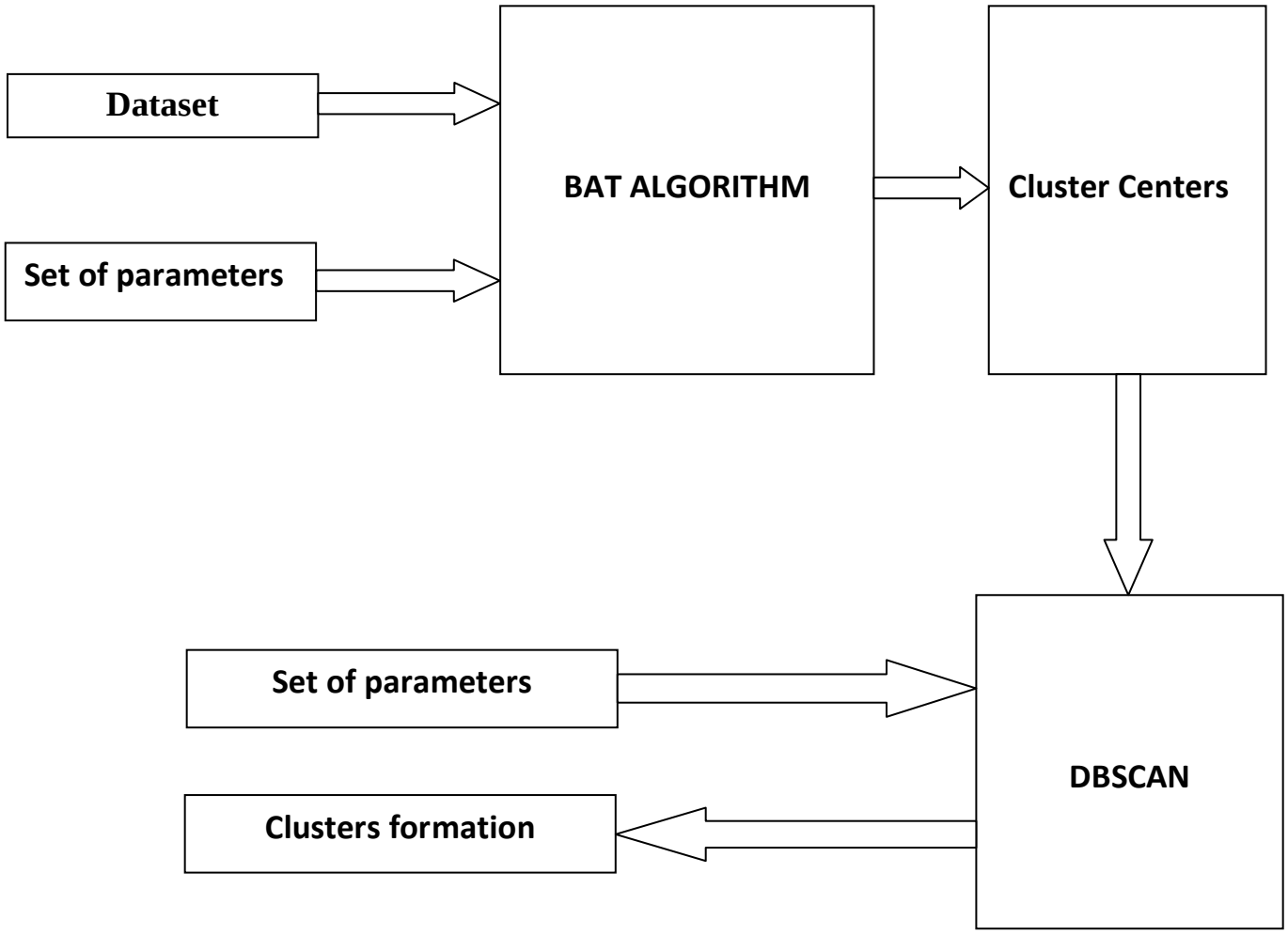


Figure 13: Architecture of proposed work

CHAPTER 5

IMPLEMENTATION, TESTING AND RESULT ANALYSIS

	epsilon	minPts	No of Iterations	No of clusters	Dimensions
Synthetic Dataset 1	0.5	10	100	2	2
Synthetic Dataset 2	0.7	10	100	2	2

Figure 14: Parameters for datasets

Datasets	Synthetic Dataset 1	Synthetic Dataset 2
No of Points	1000	1012
Intra-cluster Distance from proposed algorithm	2002.1	3428.5
Intra-cluster distance from DBSCAN	2597.5	3876.6

Figure 15: Intra-cluster distance of datasets

Clusters of Dataset 1:

In this we have 1000 points with 2 dimenstions. We are getting 2 clusters from these points based on the density. We have its parameters value as shown in above table and we are improving its cluster quality from the proposed algorithm.

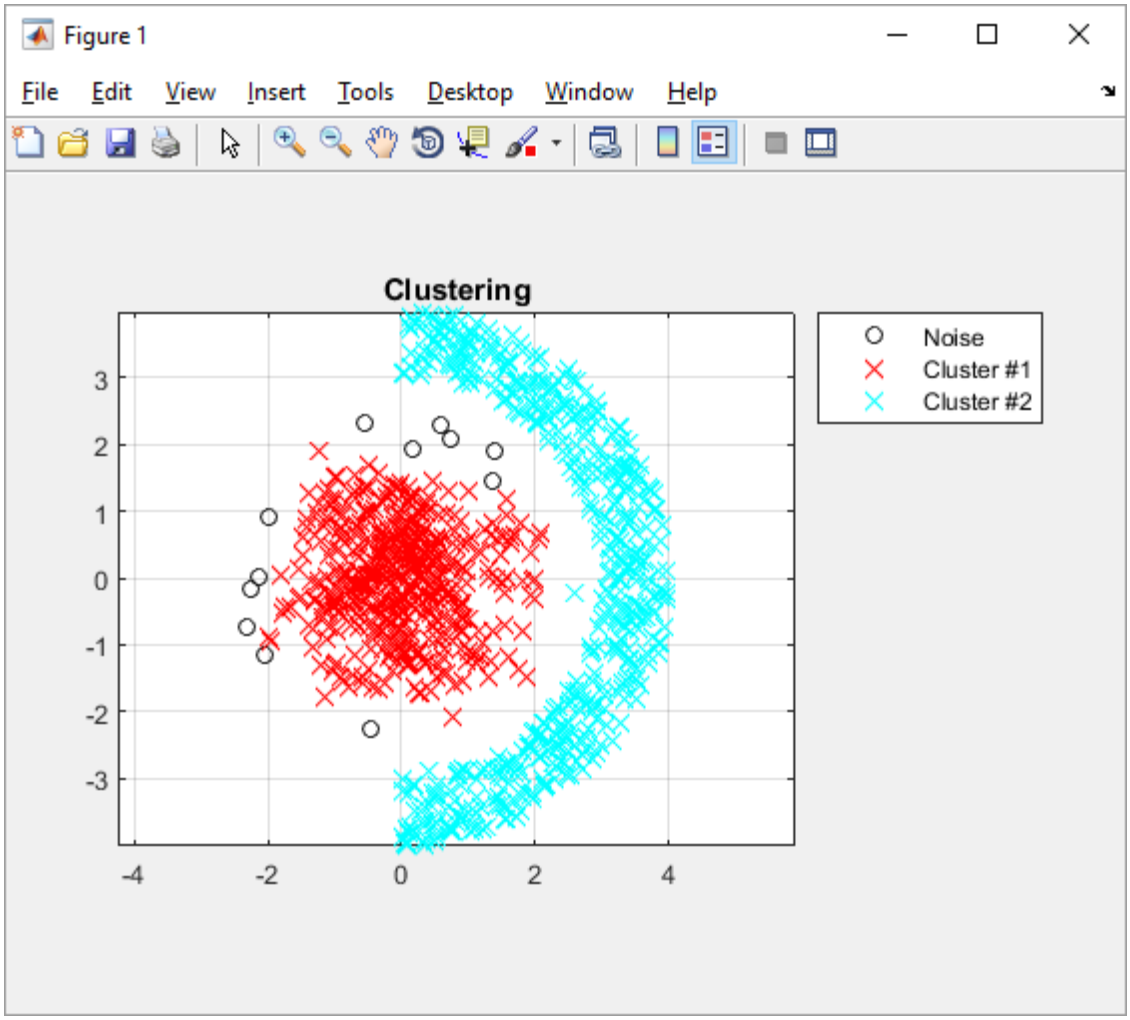


Figure 16: Clusters of dataset 1

Clusters of Dataset 2:

In this we have 1012 points with 2 dimenstions. We are getting 2 clusters from these points based on the density. We have its parameters value as shown in above table and we are improving its cluster quality from the proposed algorithm.

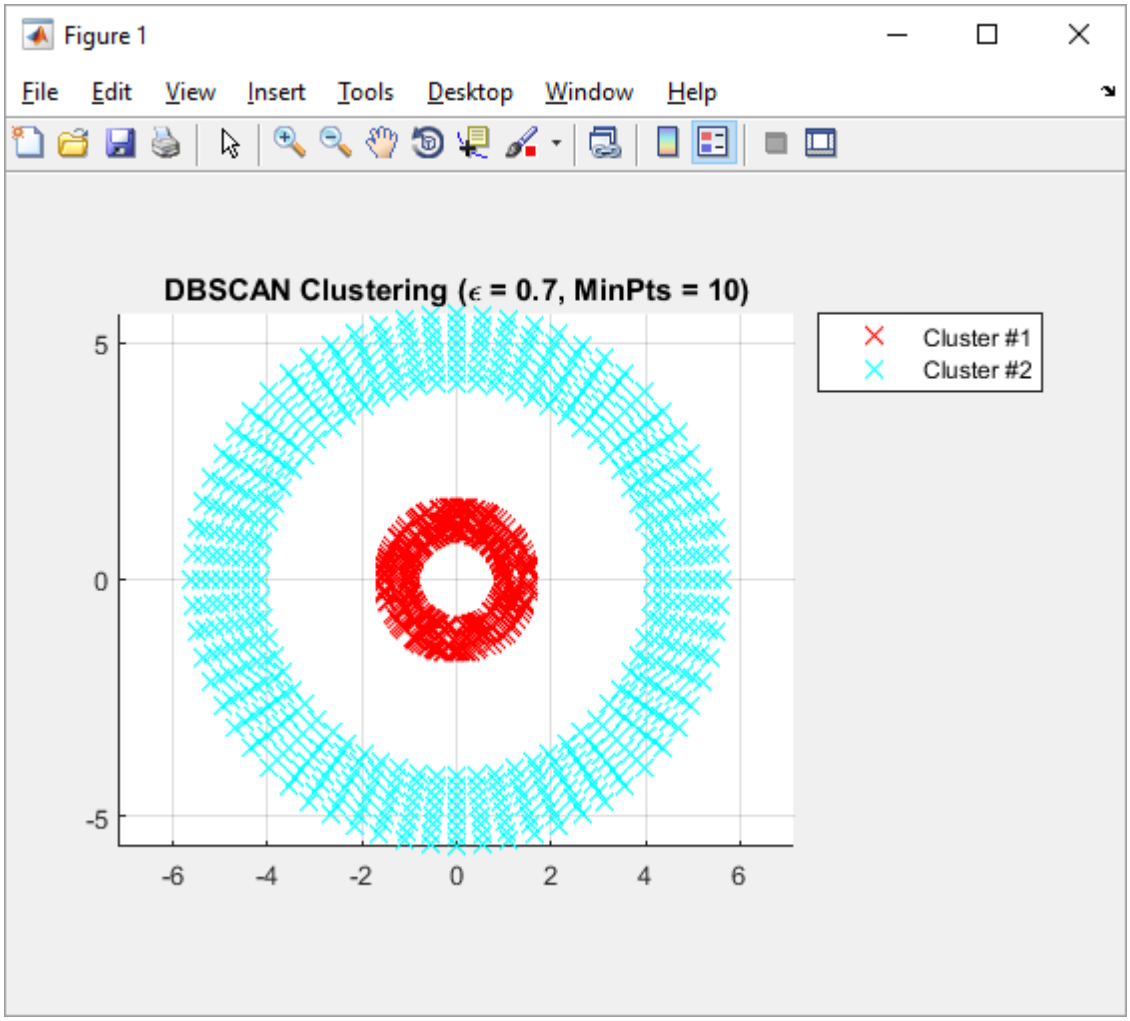


Figure 17: Clusters of dataset 2

CHAPTER 6

CONCLUSION AND FUTURE WORK

The bio-inspired algorithms are very flexible and they can change according to the changes occurring in the environment. These algorithms perform very efficiently even if the problem is not clear and they are very good in optimization. Recently, handling big data is the major issue and researchers are very focused in this field. The proposed work optimized the density based clustering algorithm DBSCAN with the help of bio-inspired bat algorithm. Bat algorithm is based on the echolocation behavior of bats. Bat changes their frequency and wavelengths according to the position of the prey and get the whole details of the background by the eco waves. Many other algorithms are also applied in the fields of the clustering. Bat algorithm can also be applied on various types of application. DBSCAN forms the arbitrary type of clusters and with the help of bat algorithm we reduced the intra-cluster of the clusters to improve cluster's quality. Intra-cluster distance is the very important measure to judge any cluster's quality. Very few researchers have implemented the bat algorithm in different problem statement. Bat algorithm is very good in exploring global search and it may misbehave in local search if its parameters are not set properly. Further improvement of DBSCAN's performance can also be done by using different kind of parameters and other bio- inspired algorithms.

REFERENCES:

1. *A Density-Based Algorithm for Discovering Clusters.* **Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu.** Germany : AAAI, 1996. KDD-96 proceedings. p. 6.

2. *A New Metaheuristic Bat-Inspired Algorithm, in: Nature Inspired Cooperative Strategies for Optimization.* **Yang, Xin-She.** 2010, p. 10.

3. *DBCURE-MR:An efficient density-based clustering algorithm.* **Younghoon Kim, Kyuseok Shim ,n, Min-Soeng Kim , June Sup Lee.** 2013, ELSEVIER, p. 21.

4. *A novel meta-heuristic algorithm: Dynamic Virtual Bats Algorithm.* **Ali Osman Topal, Oguz Altun.** 2016, ELSEVIER, p. 14.

5. *Directed Artificial Bat Algorithm (DABA).* **Rekaby, Amr.** 2013, IEEE, p. 6.

6. *Bat Algorithm: Literature Review and Applications.* **Yang, Xin-She.** 2013, IJBIC, p. 10.

7. *A Survey of Bio inspired Optimization Algorithms.* **Binitha S, S Siva Sathya.** 2012, IJSCE, p. 15.

8. *An Exhaustive Survey on Nature Inspired Optimization Algorithms.* **Manish Dixit, Nikita Upadhyay, Sanjay Silakari.** 2015, IJSEIA, p. 14.

9. *A Clustering Algorithm Based On Swarm Intelligence.* **zhongzhi, Wu bin Shi.** china : IEEE, 2001, IEEE, p. 9.

10. *BBA: A Binary Bat Algorithm for Feature Selection.* **R. Y. M. Nakamura, L. A. M. Pereira, K. A. Costa, D. Rodrigues, J. P. Papa.** Brazil : s.n.

11. *C-DBSCAN: Density-Based Clustering with Constraints.* **Carlos Ruiz, Myra Spiliopoulou, and Ernestina Menasalvas.** Germany : springer, 2007.

12. *An Improved Chaotic Bat Algorithm for Solving Integer Programming Problems* . **Osama Abdel-Raouf, Mohamed Abdel-Baset,Ibrahim El-henawy.** Egypt : MECS, 2014.

13. *Comparative Study of Density based.* **Pooja Batra Nagpal, Priyanka Ahlawat Mann.** Kurukshetra : International Journal of Computer Applications, 2011.

14. *A Novel Bat Algorithm Based on Differential Operator and Lévy Flights Trajectory.* **Jian Xie, Yongquan Zhou and Huan Chen.** china : Hindawi Publishing Corporation, 2013.

15. **Kucuksille, Selim Yilmaz and Ecir U.** Improved Bat Algorithm (IBA) on Continuous Optimization Problems. *Lecture Notes.* 2013.

16. *Bat Algorithm for Multi-objective Optimisation.* **Yang, Xin-She.** 2011.

17. *P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos.* **Slava Kisilevich, Florian Mansmann, Daniel Keim.** Washington : ACM, 2010.