CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Data clustering is one of the significant data mining techniques that become a fundamental and widely used method to generate useful information from the given set of data. The purpose of clustering is to group together data points which share appropriate similar traits to one another. While having a closer look on the purpose of clustering problem, clustering methods clearly suffer from trapping in a local optimum and cannot further progress to find global clusters. Most of the classical clustering algorithms operated by any single measure to make partitions and such methods may perform well only on some data sets. However many of the datasets lack robustness in the result by using single measures. To solve this problem, a technique has been proposed for multi-objective clustering based on hybrid optimization algorithm.

1.2 PROBLEM DEFINITION

Clustering is a way of representing large number of related data in a group. So as to represent similarities among the data. Therefore, a cluster is a collection of objects which are similar and dissimilar to the objects belonging to other clusters. So the goal of clustering is to determine the intrinsic grouping in a group of unlabeled data (Rizwan 2012). The major clustering problems arise are stated as the number of clusters is unknown, clusters may contain intra cluster density variations and inter cluster density differences, and clusters may have arbitrary shapes. This research work prioritizes the clustering problems such as maximization problem, and exploitation and exploration problem.
The swarm concept promotes self-organization and the rise of collective behavior. The properties oriented feature of Swarm Intelligence (SI) is considered as a promising approach to deal with clustering problems. Therefore, the applicability of Ant Colony Optimization (ACO) becomes significant as one of the SI algorithms to the clustering problem (Inkaya 2011). But in this ACO, the theoretical analysis is very difficult, and the probability distribution is modified by iteration. Since the existing algorithm is experimental rather than the theoretical result it is supposed to have uncertain time convergence. Another important SI approach namely, Cuckoo Search Algorithm (CSA) is used to avoid the above mentioned drawback of the existing algorithm like ACO.

Several data mining approaches are proposed by using stochastic population-based algorithms such as Particle Swarm Optimization (PSO), Immune Algorithm (IA), Artificial Bee Colony (ABC) and the ACO. However, there are various issues in real time application that are NP-hard and combinatorial (Wahid 2011). Moreover, the evolutionary algorithm such as PSO is well-suitable for solving these difficult problems due to its stochastic nature. The PSO is a biologically inspired computational search and optimization algorithm depends on social behaviors of bird flocks or schools of fish. This PSO is successfully applied in various fields because of its easier implementation especially in optimization applications and data mining.

1.3 OBJECTIVES

This research discusses various kinds of existing clustering algorithms and discusses about their merits, demerits and application areas. This research to address a more suitable optimal solution, an efficient hybrid evolutionary optimization algorithm is proposed based on the hybridization of Cuckoo Search Algorithm (CSA) and Particle Swarm Optimization (PSO).
The final product is an efficient and accurate clustering results, the Multi Objective – Cuckoo Search – Particle Swarm Optimization (MO-CS-PSO) algorithm is developed.

1.4 CONTRIBUTIONS

The main contributions of the research are as follows:

- A CSA algorithm based on a general random walk is implemented to avoid trapping in local minima and deals with large scale problem to solve the traditional clustering algorithm problems.

- A hybrid clustering algorithm by integrating the CS and PSO named CS-PSO for optimal clustering of objects. This method addresses the problem of optimal partition of data by using the features of cuckoo and swarm optimization. It also solves the exploration and exploitation problem that occur in existing optimization algorithms.

- Multi-objective Clustering with CS-PSO algorithm improves the robustness of the algorithm, which uses single measure for partitioning the data. It also attempts to solve the stability problem of the current techniques. MO-CS-PSO algorithm effectively finds the number of clusters and stable partition present in the dataset.

1.5 SCOPE OF THE RESEARCH

In the past, the researchers have mainly focused on optimization problem and also on dynamic single-objective optimization problems in the field of multi objective optimization. Some of the clustering problems that
might occur while clustering the data using various clustering methods become a major issue. This Multi-Objective Optimization Problem (MOOP) with conflicting objectives do not contain proper single solution. Therefore, the main objectives of the multi objective optimization algorithm are used to obtain a diverse set of non-dominated solutions, i.e. solutions. A Multi-Objective Optimization Problems (MOPs) are obtained everywhere in nature, to generate a single element of the set of solutions, while solving MOPs with traditional mathematical programming techniques. Moreover, these traditional methods are vulnerable to the shape or continuity of the group of solutions.

1.6 SWARM INTELLIGENCE

A Swarm Intelligence (SI) is a recent technique that emerges by observing the natural social insects and artificial systems. This system consists of multiple individual agents depending on the collective behavior in decentralized and self-organized systems. Here, the ant colonies, bees and bird flocking can effectively be applied to computational intelligent systems. These systems becomes the basic agents of SI.

1.6.1 Fundamental of SI

This SI has two fundamental notations such as self-organization and division of agents. These basic notations contain the required and sufficient properties to obtain swarm agent behavior. The flexibility and robustness of the physical environment is influenced by self-organization with SI for various tasks.

**Self-organization** - It is defined as a group of dynamical mechanism that results in structures at the global level via a system. This interaction is obtained among its low-level elements. The behavior in the group arises from observing the collective interactions of all the individual
agents. The dynamical mechanism describes the basic interaction guidelines among the various components of the system. The experimental insects can only have two key priorities in the life time to find food and defend against enemies. It has simple life when compared to the human beings.

**Division of agents** - There are different tasks simultaneously performed by specialized individuals inside a swarm. This kind of phenomenon is named as division of workers for given tasks. When this task is performed simultaneously, it has better efficiency while comparing with the serial task executing by unspecialized individuals. It also enables the SI to modify the constraints that are considering in the search space.

The above mentioned fundamental concepts of the swarm are grateful inspired and acquired to model and to obtain efficient environment behavior of the SI agent accordingly. The optimal solution for certain problems by representing the collective behavior of natural swarms such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) algorithm, artificial bee colony algorithm, bacterial foraging algorithm and so on. The ACO is simulated by the behavior of ants and the PSO is a simulation of simplified social system. Initially, the choreography of a bird flock or fish school is simulated by an original intent. Finally, the PSO is concluded as a better optimizer when compared with ASO. These algorithms are used to solve the problems of clustering more effectively.

### 1.7 CLUSTERING

Clustering is a type of unsupervised classification technique of data mining used to segregate the data into different groups or clusters. This segregation is performed based on the similarities among the data objects (Saida et al. 2014). Here, similar objects are located in the same cluster and different objects are in different clusters. The similarities and dissimilarities
are assessed based on the attribute values describing the objects. The basic requirements of the clustering in the data mining are listed below:

- Scalability and ability to deal with various types of attributes.
- Ability to handle dynamic data and discovery of clusters arbitrary shape.
- Ability to deal with noise, high dimensionality, interpretability and usability

### 1.7.1 Cluster Analysis

Cluster analysis is a major tool in data analysis to find the similarities between data according to their characteristics tails the inter-cluster similarity is low and the intra-cluster similarity is high for high quality clusters. The cluster analysis is used as a standalone data mining tool. In general, the clustering technique is used to find the cluster centre. Which represents each cluster with an input vector. A similarity metric between the input vector and all the cluster center is estimated and determined to find the nearest or the most similar one. This clustering technique is used in different applications such as marketing, land use, insurance, city-planning and earthquake studies.

An user is always capable of understanding the natural clusters or structures underlying a dataset. The typical clustering activities are listed below:

- Characterization
- Association and Correlation Analysis
- Classification
• Prediction
• Cluster analysis
• Data abstraction.
• Output assessment.

The cluster analysis has two major aspects such as clustering and cluster validation. It is an iterative process of clustering and cluster verification by the user that is facilitated with clustering algorithms such as, cluster validation techniques, visualization and domain knowledge to databases.

1.7.2 Clustering Techniques

Since the clustering is generally an unsupervised classification process, the problem of clustering is to divide a dataset into multiple groups or clusters. Therefore, the data elements in the cluster are similar to each other when compared with the data elements in different clusters. A number of clustering algorithms has been proposed for various purposes. Some more efforts is done on the combining of various clustering methods for dealing with some specific applications. Various clustering techniques are categorized into independent classes.

The major classification of the clustering algorithm is that a partitional clustering and a hierarchical clustering. In the partitional clustering, division data objects are divided into non-overlapping subnets. Here, each data object is exactly present in one subnet. The hierarchical clustering is a set of nested clusters that are organized as a hierarchical tree. Clustering algorithms are used to organize and categorize the data for model construction. In general, the clustering algorithm is further classified into
three types such as density-based methods, model-based clustering and grid-based methods.

Different types of existing clustering algorithms such as k-means, Fuzzy C-Means (FCM), Fuzzy PSO (FPSO), genetic k-means clustering, Genetic Algorithm (GA) are compared with the proposed clustering techniques such as Cuckoo Search Algorithm (CSA), efficient hybrid evolutionary optimization algorithm based on combining PSO and Cuckoo Search (CS-PSO), and Multi-Objective clustering based on hybrid optimization algorithm (MO-CS-PSO). These clustering techniques are used to solve the problems of existing clustering algorithms. The problem associated with the existing clustering algorithm is stated below:

**K-means Clustering** - The partitioning clustering algorithms, such as k-means clustering assign some specific objects into k clusters, where k denotes the predefined cluster number. This k-means is one of the popular and easy understandable clustering algorithms (Sridhar & Sowndarya 2010). However, this k-means clustering algorithm is more sensitive while selecting the initial centroids. Another major drawback of this clustering algorithm is that there is no theoretical solution to estimate the optimal number of clusters for certain dataset. The obtained result is completely dependent on the choice of initial position of the cluster centers and thus the cluster becomes linearly separable. Therefore, this algorithm does not provide a guarantee for optimal solution. In addition, the user of this clustering needs to mention the total number of cluster prior to the clustering.

Fuzzy C-Means (FCM) is a process of clustering that allows a piece of data corresponding to the multiple clusters. This method is most frequently used in pattern recognition (Xess & Agnes 2012). The FCM presents fuzziness for objects and retain large amount of data set information.
However, this clustering has high computational time and it is sensitive to the initial condition of cluster number and cluster center.

Fuzzy Particle Swarm Optimization (FPSO) clustering algorithm is used to solve the fuzzy clustering problem, especially for larger sizes. This algorithm is primarily used to enhance the performance of the PSO clustering algorithm (Khan & Engelbrecht 2012). It considers seeding of the initial swarm that is integrated with the result of FCM algorithm. This algorithm is an integration of both FCM and PCA. Therefore, it contains the drawback of the fuzzy clustering algorithm which is mentioned in the above FCM clustering.

**Genetic K-means Clustering** - This clustering algorithm is mainly used to cluster high dimensional data in sub spaces. The data sparsity problem arises in dealing with the high dimensional data. It is used to estimate the weight for each dimension in each cluster. This estimated weight is to identify the subsets of major dimensions to segregate the clusters. Here, the objective function is reduced by means of k-means clustering process. The purpose of GA with k-means is to choose the best k-value with the help of the entropy weighting, since the GA explores the space thoroughly than k-means algorithm. Furthermore, the coverage is ensured to the global optimum by using the GA algorithm. However, the accuracy and robustness of this clustering algorithm is low (Rana et al. 2010). And also this clustering algorithm had higher time complexity while performing some tasks in particular time period.

**Genetic algorithm** - In a Genetic algorithm, the basic principle of natural genetics is used to create search and optimization procedures (Izakian & Abraham 2011). The main drawbacks of GA are listed below:
• **No guarantee of finding global maxima.** But then again, apart from brute force, there is rarely any guarantee for non-trivial problems. But the likelihood of getting stuck in a local maxima early on is something you might have to deal with, for example, with some kind of simulated annealing mutation rate decay.

• **Time taken for convergence.** You usually need a decent sized population and a lot of generations before you see good results. And with a heavy simulation, you can often wait days for a solution.

• **It's a black art.** Fine tuning all the parameters for the GA, like mutation rate, elitism percentage, crossover parameters, fitness normalisation/selection parameters, etc, is often just trial and error.

• **Other complex aspects.** Apart from the genetic parameters of the GA, other things like the fitness function, choice of genetic encoding, genotype to phenotype mapping, etc, are also important in the efficacy of the system.

• **Incomprehensible solutions.** The way you communicate your desires to the system is through the fitness function. But GAs will take it literally, with absolutely no common sense. The result could be totally crazy, inefficient or incomprehensible from an engineering point of view.

• **High computational demand**

• **Different adjacent of parameters**

• **Heuristic principle**

• **There is no guarantee for optimal solution at certain time period.**

• **Parameter tuning is a major issue.**
In fact, some of the numerous SI algorithms become visible, and also applied in real world problems. The PSO and CSA are major examples of well-established SI algorithms. The CSA is considered to be one of the latest SI algorithms based on breeding and Levy-flight. This Levy-flight is based on the aging behavior of the Cuckoo birds. Therefore, in this research, three different types of clustering algorithm such as Cuckoo Search Algorithm (CSA), Cuckoo Search - Particle Swarm Optimization (CS-PSO) and Multi Objective-Cuckoo Search-Particle Swarm Optimization (MO-CS-PSO) are proposed.

1.8 CUCKOO SEARCH ALGORITHM

Generally, the Cuckoo Search Algorithm (CSA) is a novel SI algorithm based on breeding behavior of Cuckoo bird. This research work provides a brief insight of the advancement of the Cuckoo Search Algorithm (Shair et al. 2014). Generally, a Cuckoo Search (CS) depends on the brood parasitism of some specific cuckoo species. This CS algorithm is enhanced by levy flights instead of simple isotropic random walks. An aggressive reproduction strategy of some of the cuckoo species such as Ani and Guira cuckoos had deeply inspired this CS algorithm. These cuckoos lay their eggs in the communal nests. Therefore, it may remove other species eggs to improve the hatching probability of their own eggs. Few species is engaged in the obligate brood parasitism by laying their eggs in the nests of other host birds.

The standard cuckoo search depends on three idealized rules:

1. Each cuckoo lays an egg at a time, and places it in a chosen nest randomly.
2. The best nests with high quality of eggs are carried out to next generation.
3. The total numbers of host’s nests are fixed, and the egg laid by the cuckoo is exposed by the host bird with certain probability. In this situation, the host bird gets rid of the egg or abandons the nest and creates a new nest completely.

1.9 PARTICLE SWARM OPTIMIZATION

A kind of social optimization is simulated by the particle swarm. Assigning neighbors for each of the individual is defined by the underlying communication structure. It is mainly used to have interaction with the population and is also defined as the random guesses. This random guess initializes the solutions of the problem. Here, the individuals are the candidate solutions, i.e., particles. And so it is named as the particle swarm optimization. An iterative process is mainly used to enhance the solutions of the candidates by improving the fitness value of the candidate solution.

It has a capability to remember the location of the candidate solution, where it has the best solution. The best solutions of the individuals are named as the local best or the particle best. Each of the particles in the swarm creates this information in their neighbors. The movements through the search space are directed by the results of previously chosen node and operation. The solution of the problem is better in this particle swarm when compared with the non-swarm approach.

The major objective of this discovering pattern is to govern the ability of the birds to fly synchronously. It suddenly changes the direction and could also regrouping in optimal formation to achieve efficiency. The position of the particles in the search space is modified based on the social-psychological tendency of the individuals. It is mainly used to emulate the success rate of the other individuals in the swarm. The changes made in the particle were influenced by the knowledge acquired of its own or experience.
of its neighbors. It is also notable that the other particles in the swarm may also affect the search behavior of the particle. Because of this reason, the PSO is classified under symbiotic cooperative algorithm.

The basic consequence of modeling using this kind of social behavior in the search process is to apply the neighborhood principle and its social network structure to enhance the output. The swarm contains group of particles, where each particle in the swarm represents a potential solution. Here the velocity vector drives the optimization process and it also reflects the socially exchanged information. In order to form the initial PSO algorithm, three different phases are needed such as individual best, global best, and the local best, that are discussed later in the upcoming sections. The major advantage of PSO is as follows,

- PSO depends on the intelligence. It is applicable to both scientific research and Engineering.
- PSO have no overlapping and mutation calculation, and it is simple when compared with other developing calculations.

1.10 HYBRIDIZATION OF CSA AND PSO

In the recent years, the optimization algorithms find its application in the field of data mining greatly to cluster the data. The issue has to be accepted is that, most of the optimization algorithms are getting affected due to exploitation and exploration problem. Since, data accuracy and precision is entirely based on the initial solution of specific optimization algorithm, several hybridization of CSA with other clustering algorithms have also been proposed to achieve better efficiency. During this hybridization, it brings major advancement in terms of performance.
It provides better fitness value or rapid convergence rate. Therefore, this hybridization of CSA outperforms than that of the other existing algorithms such as k-means, FCM, FPSO, genetic k-means clustering, and GA.

1.11 MULTI OBJECTIVE OPTIMIZATION

Multi-objective optimization (also known as multi-objective programming, multi criteria optimization, multi attribute optimization) is an area of multiple criteria decision making which concerning with mathematical optimization problems involving more than one objective function to be optimized simultaneously. Multi-objective optimization has been applied in many fields such as science, engineering, economics and logistics where optimal decisions need to be taken in the presence of trade-offs between two or more conflicting objectives.

1.12 MULTI OBJECTIVE CUCKOO SEARCH AND PARTICLE SWARM OPTIMIZATION (MO-CS-PSO)

In current standards, a single validity does not operate satisfactory for different types of datasets. Moreover, the combination of various measures is more difficult, due to this modality for the hybridization. It is not feasible and sometimes those measures are not suitable in general. In this research work, the enhancement of CS-PSO is proposed in the form of MO-CS-PSO. Generally, a Multi-Objective (MO) optimization is expected to have faced various real-life optimization problems. This process will become tedious due to the inherent conflicting nature of the MO.

The available computational intelligence based approaches, are evolutionary computation, swarm intelligence and artificial immune systems. These approaches are used to solve the issues encountered in the MO
problems. Here, the proposed PSO method makes use of computational intelligence of swarm intelligence. The population based nature of evolutionary techniques simultaneously captures various compromising solutions in the population at each iteration. These multi objective optimization techniques play a vital role in various fields such as engineering design and resources optimization.

The main objective of multi objective is to find a group of better solutions. These solutions are obtained from a designer or a decision-maker to derive maximum benefit from the available resources. The several applications of a Multi-Objective Optimization Problem (MOOP) are more complex. The complex relationships in the decision space, namely non-convex and disconnected Pareto-optimal fronts are also involved. While making multi-criterion decision, no single solution is considered as an optimum solution to several conflicting objectives.

The MOOP with a number of trade-off optimal solutions. Therefore, the MOOP creates a Pareto front that is a group of non-dominated solutions for certain problems with multiple objectives. The major aim of the MOOP algorithm is to generate a well-distributed true Pareto-optimal front. In order to solve the MO problem, three main goals are required to be achieved:

1. Enhance the number of elements of the Pareto optimal set found.

2. Reduce the distance of the Pareto front produced in this MO algorithm regarding the true or global Pareto front.

3. Enhance the solution to be found, therefore a distribution of vectors is considered as smooth and uniform.
The major advantages of this MO are briefly explained below:

- The positive feedback leads to discoverer better solution rapidly depends on the application.
- It can be better suited in both the fields of scientific and engineering research problems.
- The premature convergence could be avoided using the distribution of computation.

Finally, these proposed clustering approaches are applied in the medical data such as well as other informatics related datasets. Here, better clustering technique is chosen in terms of average fitness value, median fitness value, standard deviation of fitness, minimum convergence and maximum convergence.

1.13 ORGANIZATION OF THE THESIS

This thesis is organized as follows:

Chapter 2 delivers the literature survey of the data clustering using various techniques. This chapter explains various clustering techniques that are already evolved in the recent past.

Chapter 3 presents data clustering using Cuckoo Search Algorithm (CSA). The cuckoo breeding behavior and the basic description of the proposed CSA is briefly explained in this chapter. The other stochastic algorithms viz. K-Means, Fuzzy C-Means, Fuzzy PSO and Genetic K-Means clustering algorithms on several well-known data sets are compared with each other to estimate the performance of the CSA.
Chapter 4 describes a hybrid optimization algorithm based on Cuckoo Search and Particle Swarm Optimization (CS-PSO) for data clustering. It is optimally clustering N object into K clusters. The novel CS-PSO algorithm is tested on various data sets, and its performance is compared with those of GA, FCM, Fuzzy-PSO and K-means clustering.

Chapter 5 introduces a multi-objective clustering based on hybrid optimization algorithm (MO-CS-PSO). Two objectives such as cluster validity index (I-index) and stability are explained. The multi-objectives are incorporated in the fitness function of the hybrid optimization algorithm.

Chapter 6 concludes the research outcome of this thesis, and provides suggestions for some possible research work that can be done in the future to extend the outcome of the present research.