CHAPTER 2

LITERATURE REVIEW
2.1 Introduction

Data mining is the process of extracting hidden patterns from the given data. With the amount of data doubling every three years, Data Mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as, marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size. The term Data Mining is often used to apply to the two separate processes of Knowledge Discovery and Prediction. Knowledge Discovery provides explicit information about the characteristics of the collected data, using a number of techniques like association rule mining. Forecasting and predictive modeling provide predictions of future events, and the processes may range from the transparent rule-based approaches to the opaque neural networks.

Data mining is usually performed on real world data. Such data are vulnerable to collinearity because of unknown and possibly unobserved interrelations. An unavoidable fact of Data Mining is that the (sub) set of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships that exist across other parts of the domain.
The process of Data Mining mainly involves three steps. They are preprocessing raw data, mine the data and interpret the results.

2.1.1 Preprocessing

Once the objective for the KDD process is known, a target data set must be assembled. As Data Mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable timeframe. A common source for data is a data mart or data warehouse.

The target set is then cleaned. Cleaning removes the observations with noise and missing data. The cleaned data is reduced into feature vectors. A feature vector is a summarized version of the raw data observation.

The feature vectors are divided into two sets, the training set and the test set. The training set is used to train the Data Mining algorithm(s), while the test set is used to verify the accuracy of any patterns found.

2.1.2 Data mining

Data mining commonly involves four classes of task:

- **Classification**

It arranges the data into predefined groups. For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include Nearest neighbour, Naive Bayes classifier and Neural network.
Clustering

Clustering is an unsupervised learning in which objects with similar properties are grouped together. Clusters exhibit two main properties: high intracluster similarity and low intercluster similarity.

Regression

It attempts to find a function which models the data with the least error. A common method is to use Genetic Programming.

Association rule learning

This searches for relationships between variables. For example a supermarket might gather data of what each customer buys. Using association rule learning, the supermarket can work out what products are frequently bought together, which is useful for marketing purposes. This is sometimes referred to as market basket analysis.

2.1.3 Interpreting the results

The final step of knowledge discovery from data is to evaluate the patterns produced by the Data Mining algorithms. The evaluation uses a test set of data which the Data Mining algorithm was not trained on. The learnt patterns are applied to this test set and the resulting output is compared to the desired output.

The massive explosion of data makes it essential to classify or group it into a set of clusters for further analysis and management. Basically, classification systems are either supervised or unsupervised, depending on whether they assign
new inputs to one of the finite number of discrete supervised classes or unsupervised categories, respectively [6].

2.2 Literature Review of Data Clustering

In clustering, also called exploratory data analysis, no labeled data is available [2][5]. The primary objective of clustering is, to separate a finite unlabeled data set into a discrete set of natural, hidden data structures. According to Jain [5] “in cluster analysis, a group of objects are split into a number of more or less homogeneous subgroups on the basis of, subjectively chosen measure of similarity (i.e., chosen based on its ability to create interesting clusters) such that the similarity between the objects within a subgroup is larger than the similarity between the objects belonging to different subgroups”.

Clustering is a well studied problem by various researchers. General references of clustering techniques have been cited by various researches [6-9]. A comprehensive review of clustering till the year 2005 has been done by Rui Xu, et al. [10] Jain, Murty, and Flynn have given a detailed review on clustering based on pattern recognition viewpoint [11]. Clustering problems under mathematically programmed viewpoints have been elaborately described by Hansen and Jaumard [12]. Many useful surveys on clustering are available in literature [13-14]. Apart from review papers, comparative research on clustering algorithms is also available [15]. Fsulo compared several clustering techniques to colour image
quantization with emphasis on computational time and the possibility of obtaining global optima [16].

Clustering algorithms can be categorized based on different starting points and criteria [9][11][13]. Basically, clustering techniques are categorized as hierarchical clustering, partitional clustering, density based clustering, graph theory-based clustering, combinatorial search techniques-based clustering, fuzzy clustering, neural network based clustering, and kernel based clustering.

Partitional clustering is primarily studied in this thesis. Therefore literature review is confined only to partitional approaches.

Partitional based clustering algorithms are also called square error-based clustering algorithms. The critical factor in the clustering algorithm is the criterion function [17]. The sum of squared error function is one of the most widely used criterions. The $k$-means algorithm is the best-known partitional clustering algorithm [18]. Although the $k$-means algorithm is very simple, easy to implement and works very well with compact and hyper spherical cluster, it is slow in nature. Parallel technique for $k$-means has been developed, that can largely accelerate the algorithm [19]. Despite its simplicity and wide acceptance, it has few disadvantages.

1. The first being the unavailability of an efficient and universal method for identifying the initial partitions and the number of clusters $K$. This results in final variation of outputs with different initializations. Several techniques have been
adopted to choose starting points which lead to solutions with minimal sum of squared distances [20]. Likas, Vlassis, and Verbeek put forth a global $k$-means algorithm consisting of a series of $k$-means clustering procedures, with the number of clusters varying from 1 to $K$ [21]. The authors claim that the algorithm is independent of initial partitions and provides strategies to accelerate it. However, their suggested algorithm is computationally complex due to the requirement of executing $k$-means; $N$ times for each value of $K$. Ball and Hall have developed ISODATA which dynamically adjusts the number of clusters by merging and splitting clusters, according to a predefined threshold [22]. Krishna and Murty proposed a novel hybrid genetic algorithm (GA) for optimal partitions of given dataset [23]. Maulik et al., have used GA for the purpose of appropriately determining fixed number of $K$ values [24]. A method of initializing $k$-means clustering using kd-trees is also suggested [25].

2. Convergence to the global optimum happens to be another important disadvantage of $k$-means clustering. The evolutionary optimization techniques, like genetic algorithm [26], particle swarm optimization (PSO) [27] can find the global optimum at the price of expensive computation. New operators were designed to achieve global optimum with fast convergence [28-29]. GAs are very useful for improving the performance of $k$-means. GA was used by Babu and Murty to find good initial partitions [29]. Laszlo M et al., used exchange of neighbouring centers for $k$-means clustering [30]. Dw Van der Merwe and AP
Engelbrecht proposed new approaches to use PSO for clustering data with better convergence [31].

3. *k*-means is its sensitivity towards outliers and noise. Both ISODATA [22] and PAM [2] consider the effect of outliers on clustering procedures. ISODATA gets rid of clusters with few objects and PAM utilizes real data points as cluster prototypes and avoids the effect of outliers.

4. *k*-means algorithm is suitable for use only to numerical data sets in its original form as there is a requirement for computing “means” of clusters. Hung [32] and Gupta et al [33] defined different dissimilarity measures to extend *k*-means to categorical variables. Ahmad. A et al., developed a *k*-means algorithm for mixed numeric and categorical data [34].

5. The last demerit is that to perform *k*-means we must know the exact value of *k* apriori. This is not always possible while dealing with real world dataset wherein the exact value of the number of cluster is not known beforehand. Several researchers [35-36] have attempted to alleviate this problem, however as per our information collected from literature not much have been achieved in this area and it remains as a challenging task to solve.

Several advances on *k*-means and other squared-error based clustering methods are also available in the literature [37-42]. Recently a hierarchical initialization approach for *k*-means has been suggested by Lu, J.F et al [43].
Newer techniques of clustering called ensemble clustering has been reported [44-45]. Recently a honey-bee mating optimization technique has been used to effectively cluster the data [46].

Most of the approaches discussed in various papers are to enhance the effectiveness, efficiency and convergence characteristics of partitional clustering algorithms. In this thesis we have investigated few newer approaches of hybridizing well established techniques like, PSO and DE with traditional $k$-means, to cluster data which produce good cluster results with faster convergence. Efforts have also been made in this thesis to propose new algorithm to find the optimal number of clusters.

2.3 Problem Identification

Studies of above cited references have revealed that there are number of challenges in clustering problems. Although many scientists, over the years, have suggested various approaches to resolve them, still there seems to be a requirement of further improvement of these techniques.

In this thesis, we have limited our study to partitional clustering algorithm, in general, and, $k$-means technique, in particular. We have identified few major challenges to resolve the popular $k$-means and they are given below:

- Initial starting point of $k$-means
- Finding of number of $k$ values apriori
- Improving the convergence rate
• Overcoming the trapping nature of k-means in local optima, and,
• Finally improving the clustering results.

2.4 Proposed Approaches

Clustering of a dataset is viewed as an optimization problem. The objective of a clustering algorithm is to group data points into predefined label having the goal of minimizing the intra-cluster distance and maximizing the inter-cluster distance. Tremendous research has gone in the past few years to evolve the clusters in complex datasets through evolutionary computation techniques like GA [26][47], and Simulated annealing (SA) [48] etc. However, the key issues of the $k$-means algorithm such as being trapped in local optima, determining the optimal number of clusters at the beginning, and convergence rate of algorithm is still to be resolved to a greater level of acceptance by research community.

In this thesis, we have proposed a class of Swarm Intelligence [49-50] techniques such as PSO and DE to answer the above issues. Swarm Intelligence algorithms are like the evolutionary techniques that are modeled after the nature. In this paradigm, the techniques are PSO [27], ant colony optimization [51], and bacteria foraging [52-53]. In this work we have given emphasis to PSO and DE [54] which are evolutionary-like computational techniques (a search method based on natural system) for solving problems which are identified in this thesis. Suitable hybridization of $k$-means with PSO is also proposed to overcome problems of initial starting point. A new technique to adopt the Scaling factor for
DE is proposed and is known as Adaptive Differential Evolution. Integration of PSO and DE [54] is also investigated so that improved clustering results can be obtained without being trapped in local optima. Regarding the convergence rate of $k$-means algorithm, the local search capability of Nelder-Mead [55] directed search is combined with $k$-means and DE. The comparative performance analysis is presented for establishing the effectiveness of these hybridizations. In this work, we have proposed a novel approach of dynamic clustering with a new variant of DE to optimally find the number of clusters apriori.