CHAPTER 1

INTRODUCTION

1.1 Data Clustering

Data clustering is the most important problem, it has been extensively considered in the data mining and machine learning classifiers , because of its several applications such as summarization, education, biomedical, segmentation, and target marketing (Jiang *et al.*, 2004; Jain, 2010; Kaufman and Rousseau, 2005). In the lack of exact labeled information, clustering can be measured as a brief model of the data which can be understudied in the sense of a generative model. The major problem of clustering might be defined as follows: Given set of datapoints are divided into a set of clusters which are as similar as possible. This is considered as very rough definition, and the variations in the problem definition are important, depending on the used specific model. For example, a generative model follows the procedure of probabilistic generative method to define similarity, whereas a distance-based approach follows the procedure of traditional distance function to define similarity among data objects. Additionally, the specific data type used also has most important drawback on the problem definition.

Some of the common application areas in which the data clustering problem occurs are described as follows:

Intermediate Step for some other basic data mining problems: Clustering have been considered as the most important step in data summarization. They offer a several key intermediate steps for several basic data mining problems such as feature selection, classification or prediction and outlier analysis.

The detailed summary of the data frequently used for different types of specific application are described as follows:

Collaborative Filtering (CF): In CF methods, the clustering presents the summarization results of concurring users. The ratings presented by the different users are used to perform the CF, which have been used to give recommendations in different types of applications.

Customer Segmentation: This is similar to CF, because it creates clusters of similar users or customers in the data. The major difference between these two applications is that instead of using rating information, arbitrary attributes related to the data objects might be used for the clustering process.

Data Summarization: Several clustering methods used in the recent works are directly related to feature reduction methods. Those methods can be measured as a form of data summarization. It has been very useful in creating compact data representations, which are easier to process and applied to variety of applications.

Dynamic Trend Detection: Several forms of dynamic and streaming algorithms have been used to carry out trend detection in a different type of social networking applications. In, those applications, the data is clustered or grouped in dynamic manner and streaming algorithms are used in order to find important patterns for changes. Examples for such streaming data's are multidimensional data, text streams and time-series data. Key trends and events in the information have been discovered by the use of traditional clustering methods.

Multimedia Data Analysis: Different kinds of documents such as images, audio or video, fall in the common type of multimedia data. For those types of multimedia data, determination of similar segments has been applied to numerous applications, such as the finding of similar snippets of music or photographs. In many of the cases multimedia data might be multimodal and might fall under different types. In those cases, the problem becomes even more difficult.

Biological Data Analysis: Biological data have become persistent in the recent years, since the success of the human genome attempt and the rising capability to gather various kinds of gene expression data. Biological data is commonly represented as either as series or as networks. A traditional clustering

algorithm gives optimal ideas about the key trends in the data and also the unusual sequences.

Social Network Analysis: In social networking applications, the important communities in the network are determined by the formation of a social network. Community detection has become the most important application in social network analysis, because it provides significant information about the community structure in the social network. Clustering algorithm has also been applied to social network summarization, which is helpful in a number of applications.

The above mentioned list of applications is not complete in any way; since it represents wide variety of problems which have been solved by using clustering algorithms. The work in the data clustering area is normally divided into a many categories.

Technique-centered: Data clustering is considered as the most important problem. Clustering uses many numerical methods to solve this problem, such as probabilistic, distance spectral, density, and dimensionality-reduction. Each of these methods has its own advantages and disadvantages; it provides good results in different problem domains. Some special types of data such as multi dimensional, big data and heterogeneous data have some challenges and these data might require some specialized clustering methods.

Data-Type Centered: Several applications create different categories of data types and its different properties. Some examples are explained here; an electrocardiogram (ECG) machine produces highly correlated time series data points which are related to one another, while a social network will create a combination of document and structural data. Some of the most general examples are category data, streaming data, distinct sequences, network data and probabilistic data. Clearly, the methodology used for the clustering process is varied depending on the type of the data. Furthermore, a number of data types are more complex than others, because of the variation among the attributes such as behavior or contextual attributes.

Additional approaches from Clustering Variations: A number of approaches have been introduced for different kinds of clustering variation. Some examples like visual analysis, supervised analysis, ensemble analysis or multi-view data analyses are used to get more approaches. Additionally, the problem of cluster validation is also essential from the viewpoint of gaining specific insights to increase the performance of clustering process.

In this research work, all of these categories mainly focus on technique centered methods only but this research work reflects the information regarding various important applications, such us text mining, image retrieval, and bioinformatics. In general, clustering algorithms are divided into two groups, they are partitioning and hierarchical clustering respectively. Partitioning methods divides the given data directly into number of groups. Some of the examples for partitioning methods are k-means (Duda et al., 2012) and probabilistic clustering using the Naive Bayes (NB) or Gaussian Mixture Model (GMM) (Baker and McCallum, 1988) (Liu et al., 2002). Hierarchical Clustering (HC) forms hierarchy of clusters by constructing a tree structure which shows the association among the clusters. The results of clustering might be obtained via cutting a tree at a desired level (Willett 1988).

In recent times, spectral clustering is developed and implemented in various domains (Jeribi *et al.,* 2015) , in which similarity among two data objects are represented by data objects and modeled as nodes of a weighted graph by edge weights. Then, clustering is formed by finding a solution of an eigen value problem and by dividing the graph nodes into different partitions.

1.2 Importance of Spectral Clustering

Spectral methods (Luxburg, 2007) (Zhang et al., 2011) are analyzed as linear programming methods to the integer programs which characterize the optimization of graph cuts. Different types of objective functions have been created and applied to different cuts such as unnormalized cut, ratio cut, and normalized cut. Thus the solution from linear programs is used to build a

multidimensional embedding tree for the nodes, and here traditional k-means algorithms are applied. These linear programs are used to create a suitable form, however the optimization problem solutions are found by using eigenvectors of the graph Laplacian.

Spectral methods (Santos et al., 2008) (Ng *et al.,* 2002) are applied to dimensionality reduction problem. Instead of working with the actual points and dimensions, the similarity matrix between the data is considered for clustering process. Spectral methods have their own advantages and disadvantages which are described below. The major advantage of this spectral clustering algorithm is that it is now possible to work with random objects for feature reduction with multi-dimensional data space. In addition, SC methods perform two tasks, measurement of Euclidean space and dimensionality reduction simultaneously. So, SC methods are particularly popular for performing clustering on random data objects that are nodes in graph.

However, SC suffers from the problem computation complexity. The similarity measure and graph cutting are used in SC algorithms. In order to perform graph partitioning (Chen *et al.,* 2011), the properties of eigenvectors of Laplacian matrix is used in spectral clustering algorithms (Verma and Meila, 2003).

1.2.1 Issues of spectral clustering

Some of the issues of SC methods are described here. First, the SC methods work with an n×n similarity matrix, so it requires more time in calculating similarity matrix which is proportional to the square of number of the data points. Second, it is more expensive method in determining the eigenvectors of data points, even though it requires only few eigenvectors. Third, it is hard to create the representation for data points in lower dimensional, except the original sample of data points which is used to create the similarity matrix. For multidimensional data, similarity matrix creation is very difficult since the data is extremely high dimensional and noisy.

For high-dimensional data space clustering problems, recently matrix-factorization based clustering method has been used and thus reduces the issues in clustering problems. In (Dhillon *et al.*, 2003), Non-negative Matrix Factorization (NMF) is applied for document clustering to achieve higher accuracy and efficiency when compared to spectral methods. So in this research work, NMF clustering algorithm is used for gene clustering.

1.3 Motivation of Co-Clustering Methods

With the rapid development of Internet and computational technologies in the recent years, several data mining applications have difficult to cluster quickly switched over from the traditional clustering of one data type to the co-clustering of multiple data types, generally concerning high heterogeneity principles. For example, in text corpus, the interrelations of words, documents, and categories in text corpus, Web pages, search queries etc, and Web users in a Web search system; papers, keywords, authors, and conferences in a scientific publication domain can be identified through clustering with several related data types. It is a very difficult task in traditional clustering methods. Initially, heterogeneous data consists of different types of relations. Processing and interpreting of heterogeneous data in an integrated way becomes a very difficult task. Integration of normalization methods does not work well.

Second, a variety of data types are correlated to each other. Dealing with each data type independently will lose this association, which is essential to increase understanding of the data. So co-clustering is developed and introduced in the data mining literature, for both two types of data (pairwise co-clustering), (Dhillon et al., 2003) and multiple data types (high-order co-clustering) (Bekkerman and Jeon, 2007) (Gao et al., 2006). During co-clustering process, hidden global structure is determined in the heterogeneous data, which effortlessly combines multiple data types to give a better understanding of the fundamental data sharing, which are extremely helpful in several real world applications.

1.4 Non Negative Matrix Factorization (NMF)

In dimensionality reduction methods, frequently used classes are matrix factorization and co-clustering methods. These methods are normally applied to data it is represented as sparse nonnegative matrices and it is possible to simplify these methods into other kinds of matrices. However, the real fact of this method is the extra interpretability inherent in Nonnegative Matrix Factorization (NMF) methods (Bishop, 2006), here the data can be represented as a nonnegative linear matrices with their concepts in the original data.

NMF methods are similar to co-clustering, which groups the rows and columns of a matrix consecutively. Let A be n × d nonnegative matrix, where n denotes the number of data entries and d represents the dimensionality of the data. For example, in text data applications, the matrix A denotes small nonnegative quantities of word frequencies and sparse data. Then, the matrix A can be divided into two low rank nonnegative matrices U and V with their sizes n × k and k × d, respectively. From these matrices, exactly k clusters are formed to represent both rows and columns respectively.

Clearly, it is desirable to determine the factorized matrices U and V, such that the sum of the squares of the residuals in R is minimized. This is similar to determining nonnegative matrices U and V by minimizing Froebinius norm of $A-U\cdot V$ matrix. It is considered as constrained optimization problem, in which the constraints related to the non-negativity entries in U and V matrices respectively. Here the parameters of this optimization problem are learned using a Lagrangian method. The detailed description of the iterative approach is discussed in (Xu *et al.*, 2003), here U represents nonnegative matrix with size n × k where n represents the data coordinate points and k represents the recently created dimensions. Highest data value of the entry (i, j) denotes that coordinates of data point i is closely related to the newly created data dimension j. So, a trivial way to do the clustering would be to assign every data point to the recently created data dimension for which it has the largest component in U.

On the other hand, if the data points belong to more than one clusters, and every k columns within V, whose values are greater than a predefined threshold those data points corresponds to document clusters. Therefore, the recently created data dimensions are to be identical with the clustering of the data points. This can be performed by using conventional k-means clustering which gives much better results. Important point of the NMF method is the size of an entry in U explains the correlation between a data point and newly created data dimension. The recently created data dimension sets are not orthogonal to each another. This principle is applied to sparse nonnegative matrices and is also named as co-clustering. Obviously, NMF is only possible way to perform the co-clustering. In order to perform co-clustering (Lee and Seung, 1999) (Dhillon *et al.,* 2003) (Rege *et al.,* 2006) , variety of graph-based spectral methods and other information theoretic methods are used in recent works.

1.5 Need for the Study

There has been a common augument in the demonic data publicly available in wide-reaching manner predominantly in the field of Bioinformatics, where huge amount of data have been collected in the form of DNA, RNA, MIRNA, etc. So far the above mentioned biological data, clustering or grouping of data was typically carried out by the labor intensive experimental work or by the semi-automatic manner. The proposed work will carry out the clustering by making use of sequential manner.

1.6 Problem Statement

Existing co-clustering methods are depends on the graphic model, which needs to be solved by eigen-problem. Computationally, these methods are less efficient and unsuitable to large-scale data sets. Furthermore, these methods are completely unsupervised. However the prediction of exact data points is a difficult task in a supervised manner by using distance function.

Conversely, exactly co-clustering heterogeneous data without considering domain-dependent background data is still becomes a difficult task. At this time, existing Swarm Intelligence (SI) based clustering methods uses preprocessed data and solves computational complexity problem.

In this research work the data point is represented in a matrix format. Therefore, the missing values are found in records or columns.

1.7 Objective of the Study

In this research work robust ensemble co-clustering is introduced in order to analyze how the integration of various data sources in the form of constraints affects the success of heterogeneous data clustering. The proposed clustering methods which might lead to important information about the functions and structures of the various dataset samples, as well as functional diversification acquired throughout family evolution and to improve the performance for the same.

The knowledge of whether to add or not, information from external sources to the database is able to improve the clustering quality for this application; Improves the efficiency of co-clustering methods and solves heterogeneous data clustering problem.

More recently, Ensemble clustering has emerged as an effective approach for clustering problems in high-dimensional spaces. Experimentation result shows that it outperforms spectral methods in document clustering, achieving higher accuracy and efficiency.

1.8 Methodology of the Study

The remarkable methodology of this research is:

- Perform Improved Ant Colony Optimization (IACO) for Robust Ensemble Co-Clustering Algorithm (IACO-RECCA) for heterogeneous data clustering.

- Perform Improved Cuckoo Search (ICS) for Robust Ensemble Co-Clustering Algorithm (ICS-RECCA) for heterogeneous data clustering.

To perform swarm intelligence based co-clustering in order to improve the performance by reducing the computation time and increasing average accuracy value.

1.9 Limitations of the Study

Existing ensemble clustering has lesser clustering efficiency for heterogeneous data clustering problem. Ensemble clustering has taken higher execution time to perform clustering task, due to the dataset consisting of irrelevant and missing data.

1.10 Preprocessing Methods for Missing Data

Missing data is the most important problem in statistical practice. Certainly, they are never welcome, since many statistical methods should not be directly applied to incomplete dataset. One of the general approaches to solve this problem is imputing missing values to plausible values. It converts the incomplete dataset into complete dataset and it is evaluated by using statistical methods. Conversely, the examination of the results must be done with caution, since there is essentially uncertainty related with the prediction of values. If missing data is higher in a record, then this record will be deleted. If missing data is higher in a column, then this column will be deleted. However if the missing values are not too much, these missing values are replaced with

1. The average value of all column data in which the missing occurred and then perform before running clustering algorithms.

2. By building a pre clustering algorithm.

Here clusters are formed based on the variables, and their missing values are replaced by average value of column data. The variable (column) which is more similar to the missing values is determined by computing mean

and standard deviation values of each data. Then the missing values are replaced by the average value of similar variable. Principal Component Analysis (PCA) is proposed in the recent work to solve missing data imputation problem.

PCA is a famous linear technique and it is performed based on the second-order statistics data which is represented by covariance matrix of the data. PCA is used extensively as a preprocessing method for solving missing data problem and it optimally compresses the mean-square error sense before classification methods. However, applying PCA is adequate technique for receiving sufficient results, if the problem measured is simple enough. PCA is developed from various applications for example by "Bro and Smilde (2014)". In the chapter 3, consider the mathematical formulation of PCA, the optimization problems related to PCA and methods for computing it.

1.11 Swarm Intelligence Methods

Swarm Intelligence (SI) extended depending on social insect collective behavior which shows many interesting properties are flexibility, strength, devolution and self-organization. SI (Bharne *et al.,* 2011) is an Artificial Intelligence (AI) technique motivated by nature, depending on the study of collective behavior in centralized and self-organized systems.

SI systems mainly depend on number of simple agents which are working together locally with one another and their environment. These systems usually have no centralized control structure regularizing how individual agents must perform; local interactions among such agents often lead to the emergence of global behavior. Some of the biological creatures are fish schools and bird flocks which visibly show structural order by means of behavior of the organisms accordingly integrated that they may change shape and direction. They appear to move as an only logical entity. The major properties of the collective behavior are pointed out as follows and are summarized in Figure 1.1.

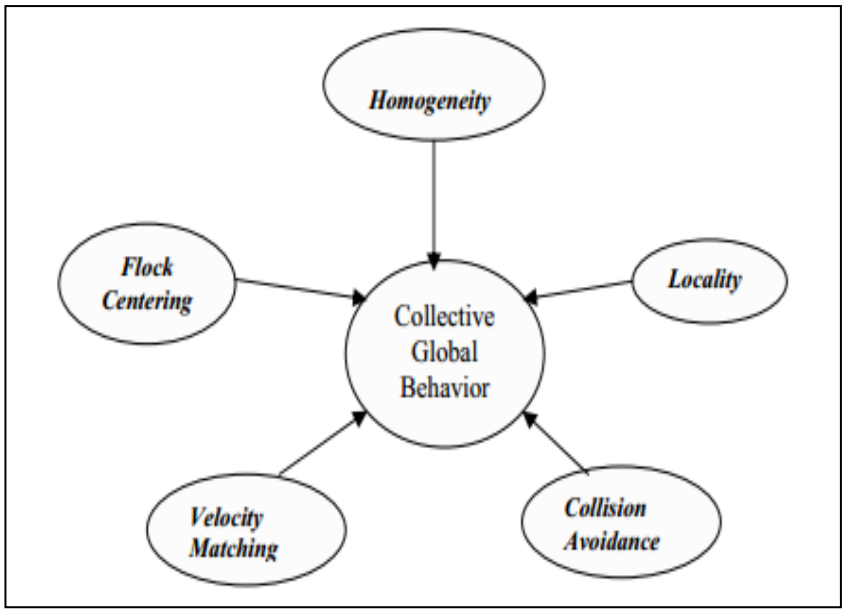

Figure 1.1. Main traits of collective behavior

Homogeneity: Every bird in the flock has similar behavioral model. The flock can move without a leader, although temporary leaders look to appear.

Roy Locality: The motions of each bird are influenced by its nearest flock-mates. Vision is the most important senses for flock association.

Collision Avoidance: Collision Avoidance should be avoided by nearby flock mates.

Velocity Matching: This effort to match velocity by means of nearby flock mates.

Flock Centering: This attempts to continue close to nearest flock mates.

Clustering models and algorithms based on SI, motivated by co-operative brood sorting of ants, are also put forward, although they are still in an opening, proof-of-model stage. The major advantages of SI clustering models and algorithms are, there is no need of background information and self-association. Conversely, the number of resultant clusters is often too high and the less convergence speed since of the ant's ineffective behaviors with the purpose of randomly picking up items and dropping down items.

An artificial Ant Colony System (ACS) is a system that creates the normal behavior of ants and increases mechanisms of cooperation and

knowledge learning mechanisms. ACS was proposed by Abraham et al., 2008 to solve combinatorial optimization problems. This novel heuristic schema is named as Ant Colony Optimization (ACO), is strong and flexible in handling extensive range of combinatorial optimization problems. The major aim of ACO is to form a problem in a graph model that finds an optimum cost path. Artificial ants move on this graph, searching for optimal paths. Each ant has their own behavior and has capability for finding moderately optimal paths. Those optimal paths are considered as the global results between ants in their colony. The function of artificial ants is stimulated from real ants: they place pheromone trails on the graph edges and choose their path related to the probabilities that relying on pheromone trails. These pheromone trails gradually decrease by evaporation. Furthermore, these artificial ants have some additional features not seen in their equivalent in real ants. In particular, they live in a discrete graph and their moves consist of transitions from nodes to nodes.

Cuckoo Search Algorithm (CSA) was proposed by recent work (Yang and Suash, 2009; Yang and Suash, 2010). CSA is fully performed based on the interesting breeding behavior like offspring parasitism of certain species of cuckoos and usual distinctiveness of Lévy flight's. The CSA is widespread and hard for many optimization problems. It is a population based algorithm and overcomes the problem of local optimum to global one.

1. Each cuckoo lays one egg at a time, and dumps its egg in randomly chosen nest;

2. The best nests with high quality of eggs will carry over to the next generations;

3. The number of available host nests is fixed and the egg laid by a cuckoo is discovered by the host bird with a probability $p_a \in [0, 1]$. In this case, the host bird can either throw the egg away or abandon the nest, and build a completely new nest. For simplicity, this last assumption can be simulated by the fraction (p_a) of the n worse nests that are replaced by new random nests. In this research work we use above mentioned both algorithms for co-clustering heterogeneous data.

1.12 Organization of the Thesis

The overall summarization of thesis is described as follows:

Clustering is unsupervised classification of data or features into similar clusters. The problem of clustering has been explained by researchers in many applications; this reveals broad usefulness of clustering in data analysis. Though clustering is a very difficult problem, transfer of its useful generic concepts and methodologies becomes slow to various applications. Nowadays it is still one of the most active research areas in data mining. Chapter 2 does a survey on several clustering algorithms by highlighting in brief state of the art, current issues, challenge and limitations and some suggestions. It is expected that, the state of the art of clustering algorithms will help the interested researchers to put forward in proposing more robust and but it has some major issues in the near future.

Chapter 3 discusses the details of proposed mechanism by Enhanced weighted version of Principal Component Analysis (EPCA) technique for preprocessing of data clustering. Then the objective function for the co-clustering ensembles towards application to data clustering is presented. In order to calculate the numerical measurements shared between two co-clustering, mutual information is considered as a symmetric measure in this work. Simulation results proved that the proposed mechanism RECCA performs better in terms of accuracy and computation time.

Chapter 4 discusses the details of Improved Ant Colony Optimization (IACO) based on clustering schema for data clustering. The overall search space is divided into two parts; “Class Hierarchy sub-graph” and “Antecedent Construction sub-graph”. In this way, powerful optimization system is proposed in this research work, which initially deals with the EPCA for preprocessing. Simulation results proved that the proposed mechanism IACO-RECCA performs better in terms of accuracy and computation time.

Chapter 5 intends to propose an Improved Cuckoo Search based Robust Ensemble Co-Clustering Algorithm (ICS - RECCA) schema for data clustering. The proposed ICS-RECCA algorithm is capable enough to perform co-clustering with the objective function as the primary component. Simulation results proved that the proposed mechanism ICS-RECCA performs better in terms of accuracy and computation time.

Finally chapter 6 gives the conclusions of the entire work and the scope of the future work. The performances of the proposed approaches are compared to existing methods. The improvements of the proposed approaches are measured in terms of precision, recall, accuracy, and time comparison. The future scope of the work and major issues of the existing works are also discussed at end of the chapter.