CHAPTER 2

REVIEW OF LITERATURE

Data analysis motivates many applications in the field of design and their operations. Data analysis procedures are differentiated based on the data source availability as either exploratory or confirmatory. Key elements in hypothesis formation or decision making procedures are grouping or classification of measurements respectively. This key element is based on either (i) fitness to an advanced model or (ii) natural clustering. It is clear that clustering analysis is the association of a group of patterns usually represented clusters based on similarity. Naturally patterns belonging to same cluster are more similar to each other compared to patterns belonging to a different cluster.

In this chapter a brief study on the details of conventional clustering methods is given. Clustering plays a significant role in different types of applications and involves many disciplines. Clustering deals with the applications with large dataset which is having different types of attributes. Study of such type of dataset is called as data mining. In order to solve different types of problem in data mining, clustering techniques is differentiated into partition based clustering, density based clustering, hierarchical methods and grid-based methods etc. This analysis focuses on clustering algorithms from a data mining angle. The main goal of this chapter is to learn the basic concepts and techniques by using large cluster subset for analysis with statistics and decision theory.

2.1. Traditional Clustering Methods

Various approaches in data clustering can be described in particular order because there is a difference between hierarchical and partition approaches hierarchical methods that produce a nested series of partitions while partition methods produce only one.

Agglomerative vs divisive: This method mainly concentrates on algorithmic structure and process. An agglomerative approach starts with individual cluster with different pattern and consecutively combines clusters together until an exact criterion is satisfied. Divisive method begins with different patterns in a single cluster and performs division until a stopping criterion is met.

Hard vs fuzzy: A hard clustering algorithm deals with separate pattern that is related to a single cluster during its operation and output. A fuzzy clustering method applies degrees of membership in each input pattern of several clusters. A fuzzy clustering can be transformed to a hard clustering by allocating each pattern with the highest membership to the cluster.

Deterministic vs stochastic: Partitioned approaches are designed to optimize a squared error function. Many traditional techniques like random search of the state space are used to achieve this optimization.

Incremental vs non incremental: When each clusters has large pattern sets and constraints on execution time or memory space that will affect the design of the algorithm, this issue will occur. At the staring stage of clustering methodology, many clustering algorithms are not designed for large data sets. But the arrival of data mining has adopted the growth of clustering algorithms that reduces the number of scans through the pattern set or decreases the data structure size used in the operation of algorithms.

2.1.1. Hierarchical Clustering Methods

Hierarchical Clustering (HC) is one type of clustering technique which concentrates mainly on the similarities among the nearby objects than the farer objects. These clustering methods groups the similar "objects" based on their distance to form "clusters". Each cluster majorly requires a minimum distance to group similar datapoints. Based on this way different clusters will be formed, that is named as a dendrogram. From the above procedure it is clear that the name "hierarchical clustering" came based on the clusters which provide an extensive hierarchy of clusters that merge with each other at certain distances instead of single partitioning of the data set.

In HC, data is represented by point-by-attribute representation sometimes that is considered as secondary importance. Sometimes hierarchical clustering is called as connectivity matrix which mainly deals with the N × N matrix of distances (dissimilarities) or similarities between training points. Linkage metrics are measured from matrix elements.

Connectivity matrix occupies more memory which is impractical. In order to overcome this limitation sparsification is introduced in the connectivity matrix. This is implemented by eliminating data which are smaller than a certain threshold value. This elimination is done by using only a particular subset of data representatives, or by using each point and particular number of its nearest neighbors (for nearest neighbor chains see Olson, 1995). The process of the constructing original dissimilarity matrix and a linkage metric reflects our a priori ideas about the data model.

(Wang *et al.,* 2000) developed a new Gaussian Mixture Modeling (GMM), and PCA for hierarchy of model based clusters. Akaike Information Criterion (AIC) selects the best cluster model. Top level Gaussian Mixture Model (GMM) with (spatially aware) which is Pseudo-Likelihood Information Criterion (PLIC) (Murtagh *et al*., 2005) is used for cluster chosen and identifiability. In Hierarchical Clustering (HC) the GMM is used with the marginal distributions inside each cluster and it is also performed based on the Bayesian Information Criterion (BIC). The cluster results are termed as a model-based cluster tree (Murtagh *et al*., 2008) which is a contentious HC algorithm (Agglomerative algorithms) is discussed in this article. It is often sufficient to implement a divisive algorithm compared to graph cut (for example) which is important for application concerned.

First application (Contreras *et al.*, 2010) is used to reduce the quantity of data that is to be clustered hierarchically. This hybrid approach is performed based on hashing and Ward minimum variance agglomerative principle. Second application, forms a HC from relationships among the sets of observations, compared to the traditional approaches. This is performed based on the use of common prefix ultrametric space. The longest common prefix ultrametric space (Contreras *et al.*, 2010) construct a hierarchy multiway tree with the purpose of being created very efficiently. For high dimensional data, the Baire distance is the base for hierarchy clustering construction on random projections. It is also addressing high quality regression with lower quality spectroscopic collection and photometric redshifts. Nonlinear regression is used for mapping photometric and astrometric redshifts.

Koga *et al.*, 2007 proposed fast agglomerative HC algorithm using Locality-Sensitive Hashing (LSH). The major advantage of this algorithm is the less time complexity with the purpose of getting reduced by O(nB), where B is practically a constant factor and represents number of information points. However, it simply relies on vector data and performed based on single linkage only so clustering accuracy is not enhanced. Additionally, it is also not applied for huge information.

Advantages of HC include:

- Tractability in the granularity level
- Easy for handling the similarity or distance in any form
- Applicability to any attributes types

Disadvantages of HC are related to:

- Ambiguity of end criteria.
- Constructed clusters are not revisited by most of the hierarchical algorithms.

Partitioning clustering algorithms slowly improve clusters that are unlike traditional hierarchical methods, in which clusters are not revisited after being constructed. This clustering method provides most appropriate and high quality clusters.

2.1.2. *Partitioning Clustering Methods*

Data partitioning algorithms such as K-means, K-medoids and probabilistic clustering divide the dataset into several subsets. Partition based algorithms find Non convex shapes clusters.

K-means Clustering

The k-means algorithm results higher clustering accuracy for many applications (Fahim *et al.,* 2006) in a successful manner. The main advantages of this algorithm are easy for implementation, scalability, speed of convergence and adaptability to sparse data (Fahim *et al.,* 2006). K-means is simple and can be easily used for practical applications as well as the time complexity is O(nkt), n is the number of objects, k is the number of clusters, t is the number of iterations. Based on the above discussion it is considered as very fast. But the computation complexity of this algorithm are complex and the result of the clusters majorly varied based on the selected centroids (Urmila and pol, 2014) initially in each iteration.

From the above mentioned initial centroid selection, new improved k-means clustering algorithm (Shunye, 2013) is proposed in recent work. This algorithm consists of three major steps: The initial step is the dissimilarity matrix measurement. Second step is the creation of Huffman tree depends on dissimilarity matrix. The Huffman tree provides the initial centroid as output then the k-means algorithm is applied to initial centroids to get k clusters. The proposed algorithm provides enhanced accuracy rates and results compared to traditional k-means.

Cominetti et al., 2010 proposed three categories of fuzzy clustering. The first category is fuzzy relation. The second one is single objective function. Third category is a nonparametric classifier. Here fuzzy clustering uses universal K Nearest Neighbor (KNN) rule for data clustering. Higher clustering results are obtained in different functional areas.

New empirical method is introduced to calculate perfect initial centroids. The newly presented algorithm (Mahmud *et al.,* 2012) provides very exact clusters with less computational time. In this algorithm average score of data points are computed from multiple attributes and weight factor. The output is sorted by using Merge sort method. The final data points are divided into k clusters which is our desired number of cluster. Finally mean of the nearest possible data point is considered as initial centroid value. From the experimental results it is clear that this algorithm decreases the number of iterations to allocate data into a cluster. But the main problem of the algorithm is assigning number of desired cluster as input.

Purohit and Joshi (2013) developed a new improved K-means clustering algorithm. This algorithm increases the performance and cluster quality by solving the problem occurred in conventional clustering methods previously. The developed algorithm uses a systematic manner for selecting the initial centroid. Initially, the closest data points are calculated by using Euclidian distance between each data points. Closest data points are found by repeating the above step on new set. The developed algorithm provides more accurate results and also reduces the computational complexity for sparse dataset.

Wang and Su (2011) proposed a k-means clustering algorithm to solve the outlier detection problem. The proposed algorithm outliers are removed by using density based clustering algorithm. The motivation of this density based clustering algorithm is that the initial cluster does not involve the outliers in its computation, applied to various datasets such as Iris, Wine, and Abalone. The main parameters that are used to test the performance are clustering accuracy and clustering time. The main drawback of this algorithm is that it will consume more time for large data sets.

A common clustering algorithm which uses greedy approach to create K-clusters with associated center of mass is called as K-Means, which measures convergence by using a squared error distortion. The efficiency of K-Means has been improved by using methods in two main directions. An effective data structure is used to reduce the computation complexity, particularly in order to store centroids or data points a multi-dimensional binary search tree (Pettinger and Di Fatta, 2009). In other way, parallel processing is applied to distribute data and computation loads over many processing nodes. However, parallel processing has been done using efficient sequential techniques based on binary search tree with some modification. The disadvantages of such techniques are irregular distribution of computation load and load imbalance. This type of efficient K-Means techniques in parallel computational environments has only limited application.

Hajiee (2010) introduced a distributed k means clustering technique in which a set of homogenous data is characterized into a set of different categories (clusters) based on the similarities between groups of parameters. The major objective of this clustering technique is the distribution of the data over a set of clusters. In order to reduce high computational cost of clustering algorithms, parallel and distributed algorithms are proposed in this work. This predicts a distributed clustering algorithm with scalability, high degree of independency for each site and cooperation.

Many clustering techniques do not concentrate on the different size or levels of the objects. Some other cases, clustering concentrates mainly on grouping similar objects or samples together without taking into account its similarity and sizes. These problems are solved by divide and conquer technique (Khalilian *et al.*, 2009) to improve the performance of the k-means clustering method. The major disadvantages of k-means clustering methods are,

Severe degradation of spaces in high dimensional data which means all pairs of points are far away from the average point. Simply to say the concept of distance between points in high dimensional spaces is ill-defined.

- Dependence on the user to identify the number of clusters in earlier
- High sensitivity to initialization phase, noise and outliers
- Repeated setups into local optima

- Non-convex clusters are difficult to deal because of its varying size and density.

K-Medoids

The Partitioning Around Medoids (PAM) (Han *et al.*, 2011) (Berkhin , 2006) method is used to represent a cluster by its medoid that is the centrally located object in the cluster which is different from traditional K means clustering. Medoids are considered as an important outliers and noise compared to centroids. PAM selects an object randomly as medoid for each of the k clusters. Then, medoid groups each of the non-selected objects that are most similar. Because of medoid selection the PAM is considered as an expensive algorithm, because it compares each medoid with the whole dataset at each iteration of the algorithm.

2.1.3. Probabilistic Clustering

Probabilistic clustering is a mixture model that is grouped into cluster for heterogeneous data. This clustering is used for the data object that has multivariate static data (demographics) in combination with variable length dynamic data (customer profile) (Smyth, 1999). The dynamic data consists of fixed sequences that are related to first-order Markov model with a transition matrix that depends on a cluster. This clustering considers the data objects with several sequences in which the sequences count per object mainly depends on geometric distribution.

"Sessions with different lengths are dealt with finite-state Markov model which is augmented with a distinct end state". New mixture model (Cadez *et al.,* 2000) for customer profiling depends on transactional information. The main task of this clustering is the determination of the number of clusters k due to the perfect probabilistic foundation of the mixture model. From a data mining viewpoint, over fitting may occur due to the excess dataset, but in probabilistic viewpoint Bayesian framework is used to address the number of parameters.

In (Cades *et al.,* 2001), an approximation of probability density is proposed for efficient random variables. This method is mainly based on the truncation of the Karhunen-Loeve expansion, and uses the density of principal components which are output of Fast Principal Component Analysis (FPCA) method of the curves. The mean and the mode of some functional dataset are found by using non-parametric kernel-based density estimation that depends on independence assumption on the principal components.

Even though probability density sometimes gives better results (to estimate the class label), the main problem is the spline coefficients or the FPCA scores which is difficult to choose between the discrete data. For instance PPCA performs well on the Growth dataset because of the FPCA scores, but it is difficult to apply into the discrete data or the spline coefficients. The problems of hierarchical clustering and partitioning clustering methods that have mentioned above are solved by spectral clustering methods which have been focused in the future work.

2.2. Co-clustering and Spectral Clustering Methods

Nowadays spectral clustering becomes more popular than other clustering algorithms. Compared to the traditional clustering techniques, spectral clustering has many advantages and it is applied to different types of data set. Though spectral clustering algorithms are simple and efficient, their performance is highly dependent on the similarity matrix that is constructed from the dataset. From the last years, the spectral clustering research has received great attentions (Luxburg, 2007).

2.2.1. Spectral Clustering Methods

Spectral clustering method is performed based on distance is introduced (Chin *et al.,* 2010) which do not assume any value regarding suitable similarity measure and the cluster number. The kernel value of the distance-based SC method is a symmetric weighted matrix that is measured from the Laplacian of the pairwise distance matrix. Additionally, the inter-cluster structure is surrounded and the relationships between intra-cluster structures are maximized in order to increase the discrimination capability on finding clusters. Different types of test datasets are used for experimentation and the results prove the correctness of the extracted clusters. Moreover, it is clear that the proposed method is robust to noisy datasets.

Stochastic perturbation theory (Huang *et al.,* 2009) is presented to perform data perturbation effects on the spectral clustering performance. The error of spectral clustering under perturbation depends on the perturbation of the eigenvectors of the Laplacian matrix. It is clear that approximate upper bounds are derived from this method on the clustering error. This bound is fitted empirically across a wide range of problems, proposing the amount of data reduction is determined by using the practical settings in order to meet permitted loss specification in performance of clustering.

The range of SC (Yan *et al.,* 2009) is extended by creating a common framework for good SC where a local transformation is initially applied to the data. This framework uses a theoretical analysis where a statistical characterization of local distortion effect on the misclustering rate is provided. Two solid instances of general framework are developed, one is local k-means clustering and another one is random projection trees. Widespread experiments explain that these algorithms can accomplish important speedups but small deviation in clustering accuracy. This algorithm is faster than approximate Nystrom method based SC, with accuracy and less memory. Unusually, these algorithms perform clustering for a single machine to cluster data sets with many observations with minimum time period.

In (Von Luxburg *et al.,* 2008) examine reliability of the SC algorithms, which uses eigenvectors of graph laplacian matrices to cluster the data. New methods are developed to establish convergence of eigenvectors to certain limit operators with increasing sample size. Finally, it can be proved that one of the two major classes of SC are converges under very general conditions, while the other is only dependable under solid additional assumptions, without satisfying real data. Analysis delivers strong proof for the preeminence of normalized spectral clustering.

A Spectral Clustering (SC) algorithm (Kumar and Daumé, 2011) is proposed for the multi-view setting where there are many accesses to multiple views of the data, which are independently used for clustering. SC algorithm uses some of the co-training steps, which is used idea in semi-supervised learning. This algorithm does not have any hyper parameters, which is a major advantage in unsupervised learning. The efficiency of the proposed multi-view based SC algorithm is better empirically compared with a traditional method on synthetic and real-world datasets.

Though, first find the similarity between data objects is important, and depends on the neighborhood graphs spectral clustering becomes unstable under different parameters. So spectral clustering cannot work as a "black box algorithm" which distinguishes the correct clusters in any given data set. But it can be converted into a powerful tool if it is applied with care that will produce good results. The commonly used classes of dimensionality reduction methods are matrix factorization and co-clustering methods. These methods use sparse nonnegative data, though it can provide good results to other kinds of matrices as well.

2.2.2. *Co-clustering Methods*

Co-clustering approaches are divided into three categories: Probability-based models, information theory- based models, and graph theoretic approaches.

In the probability-based models, for co-occurrence data new Probabilistic Latent Semantic Analysis (PLSA) co-clustering model is proposed by Hoffman and Puzicha (1999) and it is applied for collaborative filtering. Singular Value Decomposition (SVD) is used to embed the data objects into a low-dimensional space in PLSA, to get efficient pairwise co-clustering. At last, PLSA was additionally developed into a more complete generative Latent

Dirichlet Allocation (LDA) model, to cluster rows and columns of data simultaneously.

In LDA framework, many pairwise co-clustering approaches, such as Infinite Relational Model (IRM) (Kemp *et al.*, 2006), Mixed Membership Block model (Airoldi *et al.*, 2008), and Bayesian co-clustering (Domeniconi, and Laskey, 2015), are presented newly using different inference engines. A nonparametric Bayesian model (Kemp *et al.*, 2006) is presented to discover related concepts systems. Data used here contains several sets of entities, this model determines the kinds of entities in each set and the relations between different kinds possible or likely entities. Four problems are considered in this approach: clustering objects and features, learning ontology's, discovering kinship systems, and discovering structure in political data.

Data is analyzed with probabilistic models provides delicate results because the simple exchangeability assumptions having many boilerplate models. A class of latent variable models of such data is described is called as Mixed Membership Stochastic Block models (Airoldi *et al.*, 2008). This model spreads block models for relational data that capture relational structure of mixed membership, therefore an object-specific low-dimensional representation is provided. A general variation inference algorithm is developed for fast approximate posterior inference. Applications related to social networks and a protein interaction network is proposed.

Understanding the relations between clusters within different kinds of objects provides causes of a particular problem. The Bayesian approach (Domeniconi and Laskey, 2015) is initiated by developing a data generating process model, and then that model is inverted through Bayesian inference to gather cluster membership, learn cluster's characteristics. Here a basic Bayesian clustering model and several extensions to the model are considered. Experimental calculations and evaluations among the clustering methods are presented.

Recently, high-order co-clustering framework is proposed namely Mixed Membership Relational Clustering (MMRC) (Long *et al.*, 2007), where soft clustering results are performed by Expectation Maximization (EM) for large distributions. MMRC process each data type to provide interactive patterns between different datatypes and find multiple cluster structures.

Information-theory-based models is proposed by Long *et al.,(* 2006) to perform pairwise co-clustering algorithm which increase the mutual information among the clustered random variables under different conditions on size of the clusters. Later, Gao et al (2006) expanded theoretic models of pairwise information to high-order data.

In recent times, Bekkerman and Jeon (2007) suggested the Combinatorial Markov Random Field (CMRF) algorithm for high-order data, where modality of each data is combined with Markov Random Field(MRF). There is no any Theoretical proof of the effectiveness and correctness of information-theory-based models. Graph approaches are having a precise objective function for data co-clustering.

Bipartite Spectral Graph Partitioning (BSGP) (Dhillon, 2001) , is a Spectral learning that was explained and applied to co-cluster documents and words. In BSGP the data matrix is expressed as a bipartite graph and found the graph optimal normalized cut.

For high-order data, Consistent Bipartite Graph Co-partitioning (CBGC) co-clustering model is proposed which uses a semi definite programming and it is applied to hierarchical text taxonomy preparation (Gao *et al.,* 2005). The nature of graph partitioning theory provides some restriction to this algorithm that clusters from different objects must have single associations.

Practically, there is a numerous amount of unlabeled data with limited amount of labeled data, because generation of large amounts of labeled data will be expensive. Therefore, semi supervised learning, is a learning method with combination of both labeled and unlabeled data that has significant recent interest topic.

Conversely Semi-supervised clustering uses class labels or pairwise constraints to help unsupervised clustering. Data are grouped by the labeled data to modify the existing categories set and the entire regularities of the data are reflected. Existing semi-supervised clustering methods based on source information are usually divided into two categories: constraint-based and distance-based methods.

Nowadays in many researches the constraint-based and distance-based approaches are combined to perform semi-supervised clustering. Some of the semi supervised clustering are Semi Supervised Kernel K-means (SS-KK) (Kulis *et al.,* 2005), Semi-supervised Spectral Normalize Cuts (SS-SNC) (Ji and Xu, 2006), and SS-NMF (Chen *et al.,* 2007). In semi-supervised clustering of data SS-KK converts the measurement of clustering distance by weighted kernel k-means with some constraints such as reward and penalty. These clustering are represented either as vectors or as a graph. SS-SNC uses supervision to modify the clustering distance measure with pairwise information by using spectral methods.

In (Chen *et al.,* 2008), SNMF is explained to provide an integrated background for semi supervised clustering. Existing algorithms, such us SS-KK and SS-SNC are combined to form SS-NMF. From experiment results it is that SS-NMF is able to provide efficient clustering results through quick learning method with few constraints.

Existing co-clustering uses graph based clustering method, which considers mainly eigen-problem. They are ineffective and unsuitable for large-scale data sets because of the complex computation. Furthermore, these methods are completely unsupervised methods. Exactly co-clustering for domain-independent heterogeneous data is a challenging task.

2.3. Optimization Based Co-Clustering and Clustering Methods

Qiu (2004) presented a method which simultaneously models and clusters large set of images and their low-level visual features. A computational energy function suited for co-clustering images and their features was constructed and a Hopfield model based stochastic algorithm is then developed for its optimization. Apply the method to cluster digital color photographs and present results to demonstrate its usefulness and effectiveness

Yang *et al* (2009) proposed an idea of analyzing both queries and advertisements which occur with queries at the same time. It was a co-clustering algorithm that suggests queries by co-clustering advertisements and queries. It poses the co-clustering problem as an optimization problem in information theory.

De França (2016) proposed a new co-clustering technique, named Hash-based Co-Clustering (HBLCoClust) Algorithm with the aim of extracting a set of co-clusters collected from a categorical data set, however with the cooperation of scalability. The scalability issue is solved by using a probabilistic clustering algorithm, called Locality Sensitive Hashing, integrated with enumerative algorithm known as InClose. The results are compared to existing work by experimenting using categorical data sets and text corpora.

Liu et al (2015) proposed a new Fuzzy Triclustering (FTC) algorithm designed for automatic classification of three-dimensional data collections. FTC specifies membership function designed for each dimension and it is able to create fuzzy clusters concurrently on three dimensions. The results demonstrated that the proposed FTC provides higher accuracy results when compared to other fuzzy clustering and coclustering approaches on MovieLens dataset.

Wang *et al* (2011) presented a general HOCC framework, named as Orthogonal Nonnegative Matrix Tri-factorization (O-NMTF), for simultaneous clustering of multi-type relational data. The proposed O-NMTF approach employs Nonnegative Matrix Tri-Factorization (NMTF) to simultaneously cluster different types of data using the inter-type relationships, and incorporate intra-type information through manifold regularization. Where, different from existing works, we emphasize the importance of the orthogonal ties of the factor matrices of NMTF.

Kanzawa and Endo (2012) some types of fuzzy co-clustering algorithms are proposed. It was shown that the common base of the objective functions for quadratic-regularized fuzzy co-clustering and entropy-regularized fuzzy co-clustering is very similar to the base for quadratic-regularized fuzzy non-metric model and entropy-regularized fuzzy non-metric model, respectively.

Li *et al.* (2015) proposed two formulations for evolutionary co-clustering and feature selection based on the fused Lasso regularization. The evolutionary co-clustering formulation is able to identify smoothly varying hidden block structures embedded into the matrices along the temporal dimension. It was very flexible and allows for imposing smoothness constraints over only one dimension of the data matrices.

Rahman and Islam (2016) proposed a new Fuzzy Expectation Maximization and Fuzzy Clustering-based Missing Value Imputation Framework (FEMI) for solving missing data imputation problem. It imputes numerical and categorical missing values by means of use an educated guess based on records with the purpose of are similar to the record having a missing value. While recognizing a group of related records and building a guess depending on the group, it introduces a fuzzy clustering approach and presents FEM algorithm.

Ji et al (2015) proposed a new Radial Basis Function (RBF) network to find the missing values of the user-item rating matrix. In this work the co-clustering algorithm, cluster the rows and columns of the user-item rating matrix. It is able to cluster the matrix into several small matrixes by means of high similarity. Regard as the benefit of the similarity of a cluster and then predict the values via the use of RBF network. This RBF network is able to complement the sparse user-item rating matrix and increase the prediction accuracy of news recommendation system by means of collaborative filtering algorithm.

Takeuchi (2008) described that a graph-based co-clustering approach is suitable for extraction of verb synonyms from large scale texts. Their proposed bipartite graph algorithm can produce clusters of verb synonyms as well as noun synonyms taking into account word co-occurrence between verb and its argument.

In Honda *et al.* (2014) a new fuzzy co-clustering model was proposed, which is a fuzzy variant of multinomial mixture density estimation. Multinomial mixtures are a probabilistic model for co-clustering of co occurrence matrices and the proposed method extends multinomial mixtures so that the degree of fuzziness can be tuned in a similar manner to K-L information-based Fuzzy C Means (FCM).

Bao *et al.* (2015) proposed a co-clustering method, called co-clustering via local and global consistency, not only make use of the relationship between word and document, but also jointly explore the local and global consistency on both word and document spaces, respectively. That method has the following characteristics: 1) the word-document relationships is modeled by following Information-Theoretic Co-Clustering (ITCC); 2) the local consistency on both interword and interdocument relationships is revealed by a local predictor; and 3) the global consistency on both interword and interdocument relationships is explored by a global smoothness regularization.

Hierarchical agglomerative clustering through Hierarchical Particle Swarm Optimization (HPSO) based clustering is inspired by the collective intelligent behavior of swarms. HPSO-clustering (Alam *et al.*, 2010) has the properties of both partitioned based data clustering and hierarchical data clustering. Experimental results verify the performance of HPSO against PSO, traditional hierarchical agglomerative clustering and K-means clustering.

Cobos *et al.* (2014) proposed a new description-centric algorithm for the clustering of web results, called WDC-CSK, which is based on the cuckoo search meta-heuristic algorithm, k-means algorithm, Balanced Bayesian Information Criterion, split and merge methods on clusters, and frequent phrases approach for cluster labeling.

The objective of Elkeran (2013) was to minimize the length of the sheet while having all polygons inside the sheet without overlap. Their methodology hybridizes cuckoo search and guided local search optimization techniques is proposed.

Elyasigomari *et al.* (2015) developed a new hybrid optimization algorithm, COA-GA synergizing recently invented “Cuckoo Optimization Algorithm (COA)” with a more traditional Genetic Algorithm (GA) for data clustering to select the most dominant genes using shuffling. For gene classification, Support Vector Machine (SVM) and Multilayer Perceptron (MLP) Artificial Neural Networks (ANN) are used.

Mohapatra *et al.* (2015) proposed an Improved Cuckoo Search based Extreme Learning Machine (ICSELM) classify binary medical datasets. Extreme Learning Machine (ELM) is widely used as a learning algorithm for training Single Layer Feed forward Neural networks (SLFN) in the field of classification.

Ameryan *et al.* (2014) presented a four novel clustering methods based on a recent powerful evolutionary algorithm called Cuckoo Optimization Algorithm (COA) inspired by nesting behavior and immigration of cuckoo birds. To take advantage of COA in clustering, here, an individual cuckoo represents a candidate solution consisting of clusters centroids.

In Tang *et al.* (2012), the constructs of the integration of bio-inspired optimization methods into K-means clustering were presented. The extended versions of clustering algorithms integrated with bio-inspired optimization methods produce improved results.

Nature inspired, unsupervised classification method, based on the most recent metaheuristic algorithm, stirred by the breeding strategy of the parasitic bird, the cuckoo, and was introduced in Goel *et al.* (2011). The proposed Cuckoo Search Clustering Algorithm (CSCA) yields good results on benchmark dataset. Inspired by the results, the proposed algorithm is validated on two real time remote sensing satellite- image datasets for extraction of the water body, which itself is a quite complex problem. The CSCA makes use of Davies-Bouldin index (DBI) as fitness function. Also a method for generation of new cuckoos used in this algorithm is introduced. The resulting algorithm is conceptually simpler, takes less parameter than other nature inspired algorithms, and, after some parameter tuning, yields very good results.

Emary *et al*. (2014) presented an approach to automatic vessel segmentation in retinal images that utilizes Possibilistic Fuzzy C-Means (PFCM) clustering to overcome the problems of the conventional FCM objective function. In order to obtain optimized clustering results using PFCM, a cuckoo search method was used. In cuckoo search method optimized clustering results are obtained using PFCM. The main concept of CSA is the combination of offspring sponging behavior of some cuckoos and the Levy flight behavior of some birds and fruit flies. These concepts are used to optimize the fuzzy clustering method. Two benchmark databases such as the Digital Retinal Images for Vessel Extraction (DRIVE) and Structured Analysis of the Retina (STARE) datasets are used to analyze the performance of our algorithm and encouraging observed segmentation behavior.

Chifu *et al*. (2014) presented a method for clustering food offers based on the cuckoo search algorithm. The correlation between two data points is evaluated by using the Sorensen-Dice coefficient. The proposed clustering method is tested by developing a set of 800 food offers. By using a set of food recipes and a database having features of nutritional information, the food offers are generated. This database maintains the nutritional features for separate food type, based on the information provided by the Agricultural Research Service of the United States Department of Agriculture.

2.4. Summary

Clustering ensures the similarities between data objects and based on the similarity it groups the objects and data into clusters. Normally, there are two categories in clustering which is discussed above. Nowadays most of the recent active research areas in data mining use this concept. In this chapter a survey on several clustering algorithms are briefly discussed, current issues, challenge and limitations and some suggestions. It is estimated that, the state of the art of clustering algorithms will motivate the interested researchers to suggest more robust algorithms and it has some most important issues in future work. In the next chapter the quality of the co-clustering algorithms is improved by new clustering algorithm is introduced. Clustering process and the algorithms provides efficient results in grouping high quality of clusters. A suitable algorithm is found for clustering the datasets that is very important in creating a high quality of clusters and that should fit for particular applications. Due to the application of heterogeneous data, the algorithms must be very robust in managing enormous amount of data and high-dimensional structures with timely manner. The algorithms should be integrated with a feature that can accurately categorize and remove all the probable outliers and noises as an approach to decrease low quality clusters. The necessity of users-dependent parameters should be reduced because the users might tentative in number of suitable obtained clusters and others things. Since the parameters that specified wrongly might affect the overall computational performance as well as the cluster's quality. At last, algorithms are dealing with both the categorical attributes and numerical attributes separately and combination of both types of attributes. Therefore, the main goal of this research work is to develop an efficient ensemble co-clustering technique. The methodologies adopted by the present research work are explained in the following chapters.