CHAPTER 2

LITERATURE SURVEY

In order to meet the objective specified in chapter 1, the review of literature had been carried out with the following aspects:

- Clustering
- Dimensionality reduction
- Cluster quality
- Structural and Attribute Similarity
- Clustering Process
- Metrics

2.1 Clustering

Clustering is a vital process to distinguish the objects according to either their shapes or characters. Clustering is used as unsupervised learning process and its goal is to discover a new set of categories. Farley and Raftery (1998) have suggested two broad groups of clustering methods as Hierarchical and Partitioning methods.

Jain et al (1999) have proposed density-based clustering which employ non-parametric methods such as searching for bins with large counts in a multidimensional histogram of the input instance space.

Soft computing method including fuzzy clustering and Evolutionary process for clustering have been presented by Estivill-Castro and Yang (2000). Han and Kamber
(2001) have proposed additional three methods such as density based methods, model based methods and grid based methods.

Xu and Wunsch (2005) have studied Partitional clustering which is considered as a statistical way of clustering. A classical technique in this category is k-means clustering, which directly divides a data set into several number of clusters. Given a data set consisting of n observations and an integer k, the number of expected clusters, together with an initial set of k clusters, the standard k-means clustering algorithm calculates the means (the gravity centers) of these clusters and reassign each observation to the cluster with the closest mean.

Sujatha and Iyakutty (2010) have suggested refinement of web usage data clustering through k-means with genetic algorithm.

Tajunisha and Saravanan (2011) have proposed a method to find initial centroid for k-means and they have used similarity measure to find the informative genes. The goal of their clustering approach is to perform better cluster discovery on sample with informative gene.

Mukhopadhyay and Bandopadhyay (2011) have used multi-objective approaches for clustering and lend themselves well to problems where an individual best solution may be difficult to identify.

2.2 Dimensionality reduction

Costa et al (2004) have analyzed the landmarks-based Round Trip Time prediction schemes to determine coordinates of few fixed reference points (so called
landmarks) and use them to determine the positions of all other hosts. Some of them are based on positioning landmarks hosts using function Minimization.

Dabek et al (2004) have studied Landmarks-less Round Trip Time prediction scheme VIVALDI which is based on distributed simulation of physical systems to iteratively reduce the overall error of the host embedding in a Euclidean space.

Shi et al (2005) have proposed an incremental nonlinear dimensionality reduction algorithm based on ISOMAP for new samples. The geodesic distances between new data and some landmark points with self-organizing map algorithm selected from the training set experiments proved that the proposed algorithm is efficient and accurate.

Milic and Braun (2007) have proposed an algorithm based on the simplex inequality to extract the optimal number of dimensions based on distance measurements between landmarks. The methods to compute a good starting point for the minimization problem and to reduce the number of variables involved in the minimization have been provided. The experimental results showed that enhancements to Global Networking landmark positioning are able to find the optimal number of dimensions for embedding the landmarks. These enhancements also accelerate the function minimization.

Tsai and Chan (2007) have evaluated the applicability of dimensionality reduction techniques for data exploration. A summary of some current linear and nonlinear dimensionality reduction techniques has been presented. Principal component analysis (PCA) is useful in identifying significant coordinates and linear correlations in high dimensional data. PCA as well as classical MDS are unsuitable if the data set contains nonlinear relationship among the variables. General MDS techniques are appropriate when the data is highly nonmetric or sparse. If the original
high-dimensional data set contains nonlinear relationship, then nonlinear dimensionality reduction techniques are more appropriate.

Classical MDS techniques are not able to fully capture the nonlinearities in the data. However, MDS algorithms generally performs better than the nonlinear algorithms when faced with a sparse data set. In order to apply nonlinear dimensionality reduction techniques effectively, the neighbourhood, the density, and noise levels need to be taken into account. The techniques described fall under the general category of local linear techniques, which apply MDS-based techniques on a set of local neighbourhoods. This assumption of local linearity is valid when the original data set constitutes a high dimensional manifold, which appears in many applications where there is a time-varying component, such as a sequence of images in a video or microarray gene expression data taken under time-varying conditions.

Deegalla and Bostrom (2009) have proposed a new way of fusing features and classifiers based on searching for the optimal number of dimensions for each considered dimensionality reduction method. An empirical evaluation on microarray classification has been presented, comparing classifier and feature fusion with and without the proposed method, in conjunction with three dimensionality reduction methods namely Principal Component Analysis (PCA), Partial Least Squares (PLS) and Information Gain (IG). The new classifier fusion method outperforms the previous in 4 out of 8 cases, and is on par with the best single dimensionality reduction method. The novel feature fusion method is however outperformed by the previous method, which selects the same number of features from each dimensionality reduction method. Hence, it is concluded that the idea of optimizing the number of
features separately for each dimensionality reduction method can only be recommended for classifier fusion.

Moon and Qi (2010) have discussed an effective dimensionality reduction method based on Support Vector Machine (SVM) to incorporate generalization capability into dimensionality reduction process. The redundancy removal process using the asymmetric decorrelation metric to reduce the number of mapping vectors with minimum loss of essential information has been presented. Experimental results showed that SVM presents better robustness than other supervised methods like Linear Discriminant Analysis and kernel Discriminant Analysis and higher classification accuracy with less dimensionality needed than the unsupervised approaches such as Principal Component Analysis(PCA) and kernel PCA resulting in the best tradeoff among number of dimensions, classification accuracy and robustness.

Magdalinos et al (2011) have discussed some high dimensionality measures for analyzing quality clusters. They have provided a theoretic and experimental study of the family of landmark–based dimensionality reduction algorithms with computational resources and applications. Also they have presented Fast and Efficient Dimensionality Reduction Algorithm (FEDRA) which follows the principles of landmark dimensionality with two extensions such as an effective landmark selection heuristic and a heuristic for choosing the best embedding data object. Finally, the behavior of FEDRA has been validated.

Liu et al (2013) have proposed a novel clustering algorithm called landmark fuzzy neighborhood DBSCAN which, is used to represent a subset of the input data set which makes the algorithm efficient on large scale data sets. A theoretical analysis
on time complexity and space complexity, which shows both of them are linear to the size of the data set has been presented. The experiments show that the landmark FN-DBSCAN is much faster than FN-DBSCAN and a very good quality of clustering is provided.

Gonen (2013) has proposed a simple and novel Bayesian supervised dimensionality reduction (BSDR) method where the linear projection matrix and the supervised learning parameters of a linear model are learned together to improve prediction performance in the projected subspace. The probabilistic model for multiclass classification and the Gibbs sampling approach for inference on this model have been given. The formulation to find the intrinsic dimensionality using automatic relevance determination (ARD) and discussed the key properties of the algorithm. The given algorithm with seven baseline linear dimensionality reduction algorithms on three benchmark data sets and one image recognition data have been compared.

2.3 Cluster quality

Halkidi and Vazirgiannis (2001) have created five data sets and provided solutions with varied k using three algorithms. Four cluster quality measures have been assessed by determining if they have selected the correct k for each of the data sets.

Strehl and Ghosh (2002) have studied Ensemble analysis which improves classification accuracy and the general quality of cluster solution. They have also discussed the availability of multiple segmentation solutions within an ensemble and the method is Meta clustering algorithm and is based on the notion of “clustering on clusters”
Kannan et al (2004) have analyzed conductance metric which is used to estimate cluster quality and compared the number of edges. They have discussed inter cluster and intra cluster conductance and obtained internal density of clusters.

Tobert et al (2005) have discussed the applications of cluster quality in extra cellular recordings and utilized L ratio and Isolation distance measures for separating clusters.

Legany et al (2006) have assessed the quality of four measures by calculating their values for a correct and incorrect clustering solution on two data sets.

Chen and Moh (2007) have introduced a new approach, Hierarchical Clustering Search Engine with Correlation (HICSEC), to have a more efficient and more understandable clustering search engine. There are two key modules in HICSEC such as the search engine module and the clustering module. HICSEC divides the clustering engine module into three phases namely data collection, document clustering, and labeling. The architecture and technology used in the HICSEC are described.

Hang et al (2009) have proposed a Hierarchical Clustering Algorithm based on K-means with Constraints (HCAKC). In HCAKC, in order to improve the clustering efficiency, Improved Silhouette is defined to determine the optimal number of clusters. In addition, to improve the hierarchical clustering quality, the existing pairwise must-link and cannot-link constraints are adopted to update the cohesion matrix between clusters. Penalty factor is introduced to modify the similarity metric to address the constraint violation. The experimental results showed that HCAKC has lower computational complexity and better clustering quality compared with the existing algorithm.
Raj and Singh (2010) have summarized the concept of clustering, six types of clusters and four important clustering techniques with their respective algorithms which lead to validate the quality clusters.

Faraz Zaidi et al (2010) have analyzed the evaluation of quality cluster by using a new metric in terms of the path length of elements of a cluster. The proposed metric is composed of positive score to the cluster and negative score to edges between clusters.

Mesfin and Bjorn (2011) have proposed an evaluation of four clustering algorithms: k-means, average linkage, complete linkage, and Ward’s method with the latter three being different hierarchical methods. The quality of the clusters created by the algorithms has been measured in terms of cluster cohesiveness, semantic cohesiveness and both quantitative and predicate-based similarity criteria.

Bhatia and Dixit (2012) have studied to refine the original clusters generated by k-means and standard Self organizing map (SOM) algorithms. A Knockout refinement Algorithm (KRA) has been proposed for this purpose by using contingency table. By the help of two algorithms, Davies Bouldin index, Dunn index and precision, recall and F-measure index have been computed for the original and refined clusters and comparison is also carried out.

Jayabrabu, Saravanan and Vivekanandan (2012) have formulated cluster quality based on quality parameters by using Data mining agents. Clustering algorithms produce clusters based on given input data. But, it is noted that all clusters are not good clusters.

Kirkland and Iglesia (2013) have performed a comparative experimental study on some of the internal cluster quality measures and focused on the following
measures: the Variance Ratio Criterion (VRC), the Davies-Bouldin index (DB), the SD validity index (SD), the SDbw validity index (SDbw), the CDbw validity index (CDbw), Root-Mean-Square Standard Total Deviation (RMSSTD), R-Squared (RS), Silhouette Width Criterion (SWC), Overall Deviation (Dev), Connectivity and Disconnectivity. It is observed that SDbw is very erratic for varying the number of dimensions whereas all the other cluster quality measures have steady behaviour. The performances of CDbw, SDbw and SWC decrease as the number of clusters increases. Also maximum and minimum observed correlations get closer as the number of clusters increases, showing that performance of the cluster quality measures tends to converge as the number of clusters increases. Connectivity, Disconectivity, RS, RMSSTD and Dev show the best maximum and minimum correlations for varying the size of cluster.

2.4 Structural and Attribute Similarities

The algorithms for attributed graph clustering can be mainly categorized into two types, distance-based and model-based. Most existing works on attributed graph clustering fall into the category of distance-based approaches. The main idea is to design a distance/similarity measure for vertex pairs that combines both structural and attribute information of the vertices. Based on this measure, standard clustering algorithms like k-medoids and spectral clustering are then applied to cluster the vertices.

The concept of random walk has been widely used to measure vertex distances and similarities. Jeh and Widom (2002) have designed a measure called SimRank,
which defines the similarity between two vertices in a graph by their neighbourhood similarity.

Neville et al. (2003) have proposed a weighted adjacency matrix as the similarity measure. The weight of each edge is defined as the number of attribute values shared by the two end vertices. They then applied three existing graph clustering algorithms on the weighted adjacency matrix to perform clustering.

Kriegel and Schonauer (2003) have presented a similarity measure for attributed graphs, called the edge matching distance, which can be evaluated efficiently. A filter refinement architecture for efficient query processing and set of filter methods for the edge matching distance.

Pons and Latapy (2006) have proposed to use short random walks of length 1 to measure the similarity between two vertices in a graph for community detection. Sun et al (2007) have proposed Graph scope which is able to discover communities in large and dynamic graphs, as well as to detect the changing time of communities.

Many graph clustering techniques have been proposed which mainly focused on the topological structures based on various criteria including normalized cut, modularity or structural density. The clustering results usually contain densely connected components within clusters. On the other hand, Tian et al (2008) have proposed OLAP-style aggregation approaches to summarize large graphs by grouping nodes based on user-selected attributes and relationships. Vertices in one group share the same attribute values and relate to vertices in another group through the same type of relationship. This method achieves homogeneous attribute values within clusters, but ignores the intracluster topological structures.
Sun et al (2009) have proposed an algorithm, RankClus, which integrates clustering with ranking in large-scale information network analysis. The final results contain a set of clusters with a ranking of objects within each cluster.

Jiang and Pei (2009) have introduced and analyzed the problem of mining cross graph quasi-cliques with applications.

Zhou et al (2009) have proposed graph clustering algorithm based on both structural and attribute similarities and estimated the effectiveness of SA cluster as compared with other clusters through experimental analysis. A random walk based distance metric in an augmented graph where vertices from original graph which are connected to new vertices that represent vertex attributes.

In the distance measure, different weights are assigned to structure and attributes, which can be tuned automatically by their algorithm. The K-medoids algorithm is then applied to find the clustering. In order to efficiently compute the distance measure, an approximate distance computation in SA-Cluster-Opt and an incremental distance computation in Inc-Cluster (Zhou et al 2010) have been proposed.

Although the distance-based approach has been extensively studied, little has been done on the model-based approach to attributed graph clustering. Zanghi et al (2010) have adopted a similar generative process and also proposed a probabilistic model to cluster attributed graphs.

Cheng et al (2011) have studied graph clustering using unified random walk distance measures. A comparative analysis of clusters and their efficiencies have been carried out.

Arlei Silva et al (2012) have defined several definitions relating to structural correlation of attributed graph. Structural correlation pattern mining algorithm has
been formulated. By using DBLP dataset, last Fm data set and Cite seer dataset, different structural correlations have been computed.

Nawaz et al (2012) have studied intra graph clustering which can handle both structural and contextual similarity in an efficient manner, by considering weighted, undirected and multi-labeled vertex graphs. The proposed method and SA cluster method have been implemented using open source matrix multiplication. The comparative analysis has been processed through IGC-CSM, SA cluster, S-cluster, W-cluster and K-SNAP methods. Quality of clusters has been estimated based on density and entropy measures.

2.5 Clustering process

The problem of estimating the number of clusters in a large dataset with high dimensionality is arguably a difficult one. Over the past years, several approaches to the problem have been suggested. Among them advanced techniques have been proposed by Smyth (1996) on cross validation, Tibshirani et al (2000) on gap Statistics, Hansen and Yu (2001) on penalized likelihood estimation, Halkidi et al (2001) on validity indices and Roth et al (2002) on resampling.

Li et al (2007) have developed an efficient method to determine the number of clusters through the performance of spectral analysis by using the eigen values. The relation between data set and its underlying spectra with empirical results have been presented.

Tepwalkul and Maneewongwattana (2010) have studied the problem of clustering uncertain objects whose locations are described by discrete probability density function. A new algorithm U-DBSCAN that extends the existing DBSCAN
algorithm to make use of the derived vector deviation function has been proposed. The U-DBSCAN can extended the epsilon value in a way that reflects the directional probability density function of objects. The experiments to evaluate the quality of U-DBSCAN compared with ED-DBSCAN have been performed. The results showed that the proposed algorithm outperform ED-DBSCAN over the varied experimental parameters.

2.6 Metrics

Halkidi et al (2002) have summarized a collection of cluster validity approaches based on both internal and external criteria. They have used Monte Carlo technique for the computation of the probability density function of the validity indices. Also they have framed algorithms to estimate cluster validity indices.

Marx et al (2002) have proposed a new paradigm and a computational framework for revealing equivalencies (analogies) between sub-structures of distinct composite systems that are initially represented by unstructured data sets. This method coupled clustering extends the standard clustering methods for a setting consisting of a pair of distinct data sets. The problem of partitioning these given two sets into corresponding subsets, so that every subset is matched with a counterpart in the other data set. Each pair of matched subsets forms jointly a coupled cluster.

Newman and Girvan (2004) have proposed a quality metric namely modularity for evaluating internal density of clusters.

Handl et al (2005) have attempted to familiarize researches using postgenomic measurements with the multitude of validation techniques available for cluster analysis. For this purpose, the seven different types of validation measures have been
reviewed and specific weaknesses of individual measures have been addressed. It is assured that the analysis has been demonstrated not only the importance, but also the intricacy of cluster validation. It is fundamental to comprehend that the use of analytical validation techniques on their own is not sufficient, but that an understanding of the working principles of clustering algorithms, validation measures and their interactions are crucial to enable fair and objective cluster validation.

Prakash Kumar and Ramaswami (2011) have studied cluster validation based on institutional quality assessment. They have discussed institutional parameters and decision tree. The quality validity index has been obtained by applying fuzzy K-means and compared with other traditional K family clusters.

Almeida et al (2012) have proposed a new method to evaluate internal density of a cluster. They have discussed the concepts and methodologies of modularity and conductance metrics with other indices and justified the suitable metrics for the evaluation of cluster density.

2.7 Chapter summary

This chapter is devoted to the past contributions of the techniques and technologies adopted in this thesis.

On clustering studies, various researchers have utilized some methods like Hierarchical, partitioning, non-parametric, density, based and Grid based methods. Similarly, different contributions have been listed for analysing dimensionality reduction algorithm by using Landmark, ISOMAP, Global network, Principal component analysis, FEDRA, DBSCAN, BSDR and other techniques. The most
valuable methods have been periodically listed to measure cluster quality and to separate cluster. The techniques and tools like random walk, Graph scope, DBLP data set and etc have been summarized for structural and attribute similarities.

The concepts of validation, gap statistics, likelihood estimation, validity indices and U-DBSCAN have been presented to show the clustering process.

Finally, clustering metrics have been explained to measure validity indices through Monte Carlo, coupling, modularity and fuzzy k-means methods.

These developments will grow with several branches of techniques and technologies in clustering analysis.