CHAPTER 3
LITERATURE SURVEY

3.1 WORKS RELATED WITH CLUSTERING

Robust classification of multinomial observations with possible outliers was studied by Kharin and Zhuk (2000) they investigated the problem of cluster analysis of discrete (multinomial) random observations, assuming the presence of outliers in the sample. They further proposed a robust decision rule based on the truncation principle and demonstrated that the robust algorithm essentially improves the clustering performance approximately twofold.

A linear algebra measure of cluster quality was developed by Mather (2000) an application of linear algebra to text clustering, a metric for measuring cluster quality was described. The metric was based on the theory that cluster quality is proportional to the number of terms that are disjoint across the clusters.

The metric compared the singular values of the matrices for each of the clusters to determine the amount of overlap of the terms across clusters. As the metric can be difficult to interpret, a standardization of the metric was defined, which specified the number of standard deviations a clustering of a document set is from an average, random clustering of that document set. Empirical evidence showed that the standardized cluster metric correlated with clustered retrieval performance when comparing clustering algorithms or multiple parameters for the same clustering algorithm.
Clustering of spatial data using random walks was done by Koren and Harel (2001). They argued that discovering significant patterns that exist implicitly in huge spatial databases is an important computational task, a common approach to this problem is to use cluster analysis. However, traditional clustering methods have several shortcomings when addressing spatial data.

A novel approach was proposed to cluster spatial data, based on the deterministic analysis of random walks on a weighted graph generated from the data. There are many advantages of this approach, first, it decomposes the data into arbitrarily shaped clusters of different sizes and densities, second, it can overcome noise and outliers that blur the natural decomposition of the data, third, the method requires only $O(n \times \log n)$ time, and one of its variants needs only constant space.

A three-stage approach was suggested by Drobics et al (2002) for mining clusters and corresponding interpretable descriptions. The suggested approach put special emphasis on the visualization and interpretability of the results. In the first stage, the input data were represented by a self-organizing map in order to allow visualization and to reduce the amount of data while removing noise, outliers and missing values.

Then this pre-processed information was used to identify and display fuzzy clusters of similarity. Finally, descriptions close to natural language were computed for these clusters in order to provide the analyst with qualitative information. This was accomplished by generating fuzzy rules using an inductive learning method. The proposed approach was applied to three case studies, including image data and real-world data sets. The results illustrated the robustness, intuitiveness and wide applicability of the method.
Local clustering in breast, lung and colorectal cancer in Long Island, New York was discovered by Acqueez et al (2003). Analysis of spatial disease patterns usually employs a univariate approach that uses one technique to identify disease clusters. As different methods are sensitive to different aspects of spatial pattern, an approach employing a battery of techniques was expected to describe geographic variation in human health more fully.

Their study employed a multi-method approach to elucidate geographic variation in cancer incidence in Long Island, New York, and evaluated spatial association with air-borne toxics. The local Moran statistic was used to identify cancer hotspots and spatial outliers then, the geographic distributions of breast cancer in females and colorectal and lung cancer in males and females in Nassau, Queens, and Suffolk counties, New York, USA were evaluated. Then, standardized morbidity ratios (SMR values) from New York State Department of Health (NYSDOH) data were compared.

As a result, significant local clusters of high and low SMR and significant spatial outliers for each cancer-gender combination were identified. Finally the results were compared with the study conducted by NYSDOH using Kulldorff’s spatial scan statistic. They identified patterns on a smaller spatial scale with different cluster shapes than the NYSDOH analysis did, a consequence of different statistical methods and analysis scale. This methodological and comparative study evaluated whether there is substantial benefit added by using a variety of techniques for geographic pattern detection at different spatial scales.

They recommended that future studies of geographic patterns should use a concordance of evidence from a multiscalar integrative geographic approach to assure that:
1. Different aspects of spatial pattern are fully identified.

2. The results from the suite of analyses are logically consistent.

Robust classification of multinomial observations with possible outliers was studied by Kharin et al (2000) they investigated the problem of cluster analysis of discrete (multinomial) random observations, assuming the presence of outliers in the sample. They further proposed a robust decision rule based on the truncation principle and demonstrated that the robust algorithm essentially improves the clustering performance approximately twofold. A linear algebra measure of cluster quality was developed by Mather (2000) an application of linear algebra to text clustering, a metric for measuring cluster quality was described. The metric was based on the theory that cluster quality is proportional to the number of terms that are disjoint across the clusters. The metric compared the singular values of the matrices for each of the clusters to determine the amount of overlap of the terms across clusters. As the metric can be difficult to interpret, a standardization of the metric was defined, which specified the number of standard deviations, a clustering of a document set is from an average, random clustering of that document set. Empirical evidence showed that the standardized cluster metric correlated with clustered retrieval performance when comparing clustering algorithms or multiple parameters for the same clustering algorithm.

Shivakumar et al (2000) developed Model-Based Hierarchical Clustering; they presented an approach to model-based hierarchical clustering by formulating an objective function based on a Bayesian analysis. The model organized the data into a cluster hierarchy while specifying a complex feature-set partitioning, a key component of our model; features were having either a unique distribution in every cluster or a common distribution over some (or even all) of the clusters. The cluster subsets over which these features had such a common distribution corresponded to the nodes (clusters) of the tree
representing the hierarchy. This general model was applied to the problem of document clustering for which a multinomial likelihood function and Dirichlet priors were used. The algorithm consisted of a two-stage process wherein first a flat clustering is performed followed by a modified hierarchical agglomerative merging process that included determining the features that will have common distributions over the merged clusters.

Polanco et al (2001) argued in favour of artificial neural networks for exploratory data analysis, clustering and mapping. They suggested using artificial neural networks for mapping of science and technology as a multi-self-organizing maps approach. They proposed the Kohonen self-organizing map (SOM) for clustering and mapping according to a multi maps extension. After presenting the Kohonen SOM algorithm, following improvements were detailed i.e. the way of naming the clusters, the map division into logical areas, and the map generalization mechanism. An Interactive Visual Framework for Detecting Clusters of a Multidimensional Dataset was developed by Bhadra et al (2001). A key problem in data mining applications is that of detecting clusters, i.e., groups of closely related data items, of a multidimensional dataset. Users are adept at detecting clusters in visually presented information. They presented an interactive framework that allowed a user to visualize a multidimensional dataset, and detect clusters in it by repeatedly changing its visualization using interactive operations such as parameter re-adjustment, zooming, subset selection, and cluster probing. The experiments also suggested that animation provided important clues in detecting clusters.

Clustering of spatial data using random walks was done by Koren et al (2001). They argued that discovering significant patterns that exist implicitly in huge spatial databases is an important computational task, a common approach to this problem is to use cluster analysis. However,
traditional clustering methods have several shortcomings when addressing spatial data. A novel approach was proposed to cluster spatial data, based on the deterministic analysis of random walks on a weighted graph generated from the data. There are many advantages of this approach, first, it decomposes the data into arbitrarily shaped clusters of different sizes and densities, second, it can overcome noise and outliers that blur the natural decomposition of the data, third, the method requires only $O(n \log n)$ time, and one of its variants needs only constant space.

Low-Complexity Fuzzy Relational Clustering was studied at The Pennsylvania State University. New algorithms (Fuzzy c-Medoids or FCMdd and Robust Fuzzy c-Medoids or RFCMdd) were proposed for fuzzy clustering of relational data. The objective functions were based on selecting $c$ representative objects (medoids) from the data set in such a way that the total fuzzy dissimilarity within each cluster was minimized. A comparison of FCMdd with the well-known Relational Fuzzy c-Means algorithm (RFCM) showed that FCMdd was more efficient. Several applications of these algorithms to Web mining, including Web document clustering, snippet clustering, and Web access log analysis were also presented.

Supervised harvesting of expression trees was proposed by Hastie et al (2001). They proposed a new method for supervised learning from gene expression data called 'tree harvesting'. This technique started with a hierarchical clustering of genes, and then modeled the outcome variable as a sum of the average expression profiles of chosen clusters and their products. It can be applied to many different kinds of outcome measures such as censored survival times, or a response falling in two or more classes (for example, cancer classes). The method discovered genes that had strong effects on their own, and genes that interacted with other genes. The method on data from a lymphoma study and on a dataset containing samples from eight
different cancers was illustrated. It identified some potentially interesting gene clusters. In simulation studies, it was found that the procedure may require a large number of experimental samples to successfully discover interactions.

Cluster-Rasch models for microarray gene expression data were developed by Hongzhe et al (2001). They proposed two different formulations of the Rasch statistical models to the problem of relating gene expression profiles to the phenotypes. One formulation allowed investigating whether a cluster of genes with similar expression profiles is related to the observed phenotypes; this model can also be used for future prediction. The other formulation provided an alternative way of identifying genes that are over- or underexpressed from their expression levels in tissue or cell samples of a given tissue or cell type. The results illustrated the methods on available datasets of a classification of acute leukemias and of 60 cancer cell lines. For tumor classification, the results were comparable to those previously obtained while for the cancer cell lines dataset, four clusters of genes were found that were related to drug response for many of the 90 drugs that were considered. In addition, for each type of cell line, genes that are over- or underexpressed relative to other genes were identified. The cluster-Rasch model provided a probabilistic model for describing gene expression patterns across samples and can be used to relate gene expression profiles to phenotypes.

Dudoit et al (2002) devised a prediction-based resampling method for estimating the number of clusters in a dataset. Microarray technology is increasingly being applied in biological and medical research to address a wide range of problems, such as the classification of tumors. An important statistical problem associated with tumor classification is the identification of new tumor classes using gene-expression profiles. Two essential aspects of this clustering problem are: to estimate the number of clusters, if any, in a
dataset; and to allocate tumor samples to these clusters, and assess the confidence of cluster assignments for individual samples. The first of these problems was addressed by them. They developed a new prediction-based resampling method, Clest, to estimate the number of clusters in a dataset. The performances of the new and existing methods were compared using simulated data and gene-expression data from four recently published cancer microarray studies. Clest was generally found to be more accurate and robust than the six existing methods considered at the time of the study. Validation of visual clusters in large datasets fixed point clusters of spectral features was done by Hennig et al (2002).

The clustering property of corner transformation for spatial database applications was designed by Song et al (2002). Spatial Access Methods (SAMs) are often used as clustering indexes in spatial database systems. Hence, a SAM should have the clustering property both in the index and in the data file. They argued that corner transformation preserves the clustering property such that objects having similar sizes and positions in the original space tend to be placed in the same region in the transformation space.

Vishwanathan et al (2002) worked on Kernel enabled K-means algorithm. They presented a novel method to learn arbitrary cluster boundaries by extending the k-means algorithm to use Mercer kernels. Each cluster centroid was interpreted as a linear combination of the cluster points in the higher dimensional space and this formulation was used to kernel enable the k-means algorithm. The advantage of this formulation was that one could work in the higher dimensional kernel space where it is easier to find smooth surfaces which separate points belonging to different clusters.

Chen et al (2002) conducted two case studies for visualizing and tracking the growth of competing paradigms. They demonstrated the use of an
integrative approach to visualize and track the development of scientific paradigms. The approach was designed to reveal the long-term process of competing scientific paradigms. A cluster of highly cited and cocited scientific publications in a cocitation network represented the core of a predominant scientific paradigm. They depicted and animated the growth of a paradigm through the rise of citation rates and the movement of its core cluster towards the centre of the cocitation network. Two cases of competing scientific paradigms in the real world were studied.

A three-stage approach was suggested by Drobics et al (2002) for mining clusters and corresponding interpretable descriptions. The suggested approach put special emphasis on the visualization and interpretability of the results. In the first stage, the input data were represented by a self-organizing map in order to allow visualization and to reduce the amount of data while removing noise, outliers and missing values. Then this preprocessed information was used to identify and display fuzzy clusters of similarity. Finally, descriptions close to natural language were computed for these clusters in order to provide the analyst with qualitative information. This was accomplished by generating fuzzy rules using an inductive learning method. The proposed approach was applied to three case studies, including image data and real-world data sets. The results illustrated the robustness, intuitiveness and wide applicability of the method.

Clustering gene-expression data with repeated measurements was done by Yeung et al (2003). They evaluated several clustering algorithms that incorporated repeated measurements, and showed that algorithms that take advantage of repeated measurements yielded more accurate and more stable clusters. It was shown in particular that the infinite mixture model-based approach with a built-in error model produces superior results.
Similarity-Driven cluster merging method for Unsupervised Fuzzy Clustering was developed by Xiong et al (2004). The method was used to resolve the problem of cluster validation. Starting with an overspecified number of clusters in the data, pairs of similar clusters were merged based on the proposed similarity-driven cluster merging criterion. The similarity between clusters was calculated by a fuzzy cluster similarity matrix, while an adaptive threshold was used for merging, in addition, a modified generalized objective function was used for prototype-based fuzzy clustering. The function included the p-norm distance measure as well as principal components of the clusters, the number of the principal components was determined automatically from the data being clustered.

The performance of this unsupervised fuzzy clustering algorithm was evaluated by several experiments of an artificial data set and a gene expression data set.

Kaymak et al (2000) worked on extended fuzzy clustering algorithms. Fuzzy clustering has been applied successfully in various fields including finance and marketing. Despite the successful applications, there are a number of issues that need to be dealt with in practical applications of fuzzy clustering algorithms. He proposed two extensions to the objective function based fuzzy clustering for dealing with these issues. First, the (point) prototypes should be extended to hyper volumes whose size will be determined automatically from the data being clustered. These prototypes are less sensitive to a bias in the distribution of the data. Second, cluster merging should be done by assessing the similarity among the clusters during optimization. Starting with an over-estimated number of clusters in the data, similar clusters were merged during clustering in order to obtain a suitable partitioning of the data. An adaptive threshold for merging was introduced.
The proposed extensions were applied to Gustafson-Kessel and fuzzy c-means algorithms, and the resulting extended algorithms were given. The properties of the new algorithms were illustrated in various examples.

Gordon et al (1991) conducted User-based document clustering by re-describing subject descriptors with genetic algorithms. Information retrieval systems use clustering of documents and queries to improve both retrieval efficiency and retrieval effectiveness. Normally, clustering involves grouping together static descriptions of documents by their similarity to each other, though user-based clustering suggested that usage patterns concerning co-relevance can form a basis for clustering. They reported that clusters of co-relevant documents obtain increasingly similar descriptions so that documents become more effective in matching relevant queries. As a result of the increased similarity, clustering algorithms can more accurately group documents into useful clusters.

Johnson et al (1996) tried adaptive indexing in very large databases; they compared the use of 2 adaptive algorithms (genetic algorithms, and neural networks) in clustering the tables of a very large database. These clusters allowed the user to index into this overwhelming number of tables and find the needed information quickly. As the tables were based on the user's queries, and not on the content of the tables, thus the clustering reflected the unique relationships each user saw among the tables. In the process the original database remains untouched; however each user has a personalized index into the database.

hypermedia documents. The clusters allowed the user to index into the nodes and find information quickly. It focused on the user's paths through the hypermedia document and not on the content of the nodes or the structure of the links in the document, thus the clustering reflected the unique relationships each user sees among the nodes. The original hypermedia document remains untouched, and each user has a personalized index into this document.

3.2 WORKS RELATED WITH CTG DATA CLASSIFICATION

Onisko and Druzdel (2011) have analyzed the work of Bayesian Network parameters on the accuracy of medical diagnostic systems. They have performed a series of experiments studying the influence of precision in parameters, on model performance, in the context of a practical medical diagnostic model, Hepar II and three additional models based on real medical data from the Irvine Machine Learning Repository. Their study revealed that the performance of all four models was sensitive to noise in numerical parameters, i.e., the diagnostic accuracy of the models decreased after introduction of noise into their numerical parameters. They also analyzed the influence of bias in parameters on model performance. Diagnostic accuracy of Bayesian network models was found sensitive to imprecision in probabilities, provided they were rounded up. However, the main source of the sensitivity appeared to be in rounding small probabilities to zero.

Warrick and Hamilton (2011) modeled the Cardiotocography data as an input-output system to estimate its dynamics in terms of an impulse response function (IRF). They identified the models using subspace methods that incorporated noise-suppression and permitted the use of non-contiguous
data. Using contiguous data, the subspace method performed better than linear regression. When they included non-contiguous data, even more pathological records were modeled with their approach. Compared to linear regression, the IRF gain showed statistically significant differences more often between normal and pathological records (in 15/18 vs. 10/18 epochs) over the final three hours of labor.

Chen et al (2009) developed a Lab view based FHR and uterine contraction (UC) pattern analysis software. This software had great potential for home-care use. The signal processing methods utilized in their study were median filter and peak/valley detection method. The analysis performance was verified from nineteen pregnant women's data. The accuracy of FHR baseline, baseline variability, early Deceleration, UC frequency and Non Stress Test (NST) reached 100%. The accuracy of acceleration frequency was 90%. The accuracy of late and variable decelerations attained 95%.

Vullings et al (2009) presented a new technique for Maternal ECG (mECG) removal in antenatal abdominal recordings. This technique operated by linear prediction of each separate wave in the mECG. Its performance in mECG removal and FHR detection was evaluated by comparison with spatial filtering, adaptive filtering, template subtraction and independent component analysis techniques. The new technique outperformed the other techniques in both mECG removal and FHR detection (by more than 3%).

Baluz et al (2011) assessed the performance of different machine learning algorithms for Cardiotocography exams classification. Two classification tasks were considered: the first was to characterize the exam according to the fetal state; the second intended to classify the exam according to the morphological pattern. Their experimental results concluded
that the tested algorithms had a promising performance for both tasks. Among the analyzed algorithms, Random Forest algorithm classified with the best results. They have reported an accuracy of 94.9% for the FHR classification task. It also produced an accuracy of 87.3% for the morphological pattern dataset.

Czabanski et al (2010) described an application of the artificial neural network based on logical interpretation of fuzzy if-then rules neuro-fuzzy system to evaluate the risk of low-fetal birth weight using the quantitative description of CTG signals. They applied different learning procedures integrating least squares method, deterministic annealing (DA) algorithm, and \( \varepsilon \)-insensitive learning, as well as various methods of input dataset modification. The performance was evaluated with the number of correctly classified cases (CC) expressed as the percentage of the testing set size, and with overall index (OI) being the function of predictive indices. The best classification efficiency (CC = 97.5% and OI = 82.7%), was achieved for integrated DA with \( \varepsilon \)-insensitive learning and dataset comprising of the CTG traces recorded as earliest for a given patient.

Zhang et al (2012) presented a new dynamic method for incrementally updating approximations of a concept under neighborhood rough sets to deal with numerical data. A comparison of the proposed incremental method with a non-incremental method of dynamic maintenance of rough set approximations was conducted by an extensive experimental evaluation on different data sets from UCI. Experimental results revealed that the proposed method effectively updated approximations of a concept in practice.
3.3 WORKS RELATED WITH OUTLIER DEDUCTION

Outlier detection techniques can be classified along several dimensions. The most extensive effort in this direction has been done by Hodge and Austin (2004). But they have only focused on outlier detection techniques developed in machine learning and statistical domains. Most of the other reviews on outlier detection techniques have chosen to focus on a particular sub-area of the existing research.


A review of selected outlier detection techniques, used for network intrusion detection, was presented by Lazarevic et al (2003).

Outlier detection techniques developed especially for system call intrusion detection has been reviewed by Forrest et al (1999) and later by Snyder (2001) and Dasgupta et al (2000). A substantial amount of research on outlier detection has been done in statistics and has been reviewed in several books (Rousseeuw et al 1987, Barnett et al 1994) as well as other reviews Hawkins (1980). Tang et al (2006) provide a unification of several distance based outlier detection techniques. These related efforts have either provided a coarser classification of research done in this area or have focused on a subset of the gamut of existing techniques. To the extent of our knowledge, our survey is the first attempt to provide a structured and a comprehensive overview of outlier detection techniques. Outlier deduction methodologies and techniques were presented Hodge et al (2004). These techniques overcome the intruder deduction in input layer as well as output layer.
3.4 SUMMARY

In this section, literature survey was done. Different approaches related to cardiotocogram were analyzed and how do overcome challenging problems were discussed. Various clustering and classification algorithms have surveyed and also discussed the advantages and disadvantages for the same. Recent works with CTG data classification have been noted. Several techniques related to outlier deduction have been observed. Different classification methods were discussed and how do overcome the existing problems.