# CHAPTER 2

# LITERATURE SURVEY

## 2.1 INTRODUCTION

This chapter reviews the existing data mining techniques available in the literature related to data preprocessing, feature selection and classification. This chapter categorizes the existing techniques into preprocessing, feature selection and classification.

## 2.2 VARIOUS APPROACHES BASED ON FEATURE SELECTION

TSVM is an iterative algorithm which comprises of unlabeled samples in training phase. Ujjwalet al (2013) introduced a transductive modus operandi in which transductive sample is preferred in the course of a filtering process of unlabeled data and an algorithm is proposed.

Shuichi etal (2012) introduced a non linear semi-supervised logistic discriminant process in which it corresponds to Gaussian basis expansions through regularization based on graph. Graph laplacian employed in regularization expression is one of the most important techniques in graph based regularization method. It is based on degree matrix and weighted adjacency matrix. Weighted matrix M is resolved to be an n x n matrix. To select values of more than a few tuning parameters they develop a model for the selection norm from Bayesian and information theoretic approach.

Laplacian graph are functional with no trouble to examine high dimensional or multifaceted dataset in both labeled and unlabeled data set. This method also decreases the error rate of prediction.

Some datasets are composed of numerous given labels. In particular conditions, data samples are being actually labeled by certain domain specialist. These labels, or hypotheses, are very supportive in machine learning and data mining responsibilities. In point of fact, they produce to be very functional in feature selection. Feature selection methods make employ of the specified labels to direct the feature search. In particular, with the accessibility of the class label, the author can explain the relevance measure. The feature is considerable to the class if it is associated with the class. Class labels can be used to investigate the statistical relations and features of samples that combine to the same class as discussed in (Leung et al 2006). This type of feature selection is called supervised, given that the space search and feature evaluation is supervised by the definite labels. There is enormous amount of accessible supervised feature selection in the field so far. Some of them will be temporarily mentioned when the author talks about the feature selection model.

Dimensionality reduction can be either feature selection or feature extraction as discussed in (Saeys et al 2007; Van der et al 2009). The latter, such as PCA, reduces data dimensionality by prognostic data into lower dimensional space. While feature extraction has been successfully used to reduce dimensionality and develop learning performance, the novel feature space does not correspond to the original one as mentioned in (Thangavel&Pethalakshmi 2009). In other words, the new features are not actually liked to the original features, therefore, they are meaningless. Therefore, they neither can be used to justify the reduction nor for further domain analysis. On the other hand, Feature selection does not suffer from

this limitation. It selects a subset of the original features with not including any kind of transformation as discussed in (Bolon-Canedo et al 2012;Alelyani et al 2013). For that reason, the selected features maintain their physical meaning, therefore, justification and domain analysis is possible further.

Feature selection algorithms of filter model are self-governing of several classifiers. The result totally depends on the characteristics of the dataset itself with regard to the class label. For instance, Fisher score evaluates each feature separately by means of fisher criterion as discussed in (Guet al 2012). Other approaches make use of different criterion to evaluate features' relevancy. Spectral feature selection SPEC (Zhao &Liu 2007) and Laplacian score (He et al 2006) select features based on the examination of the eigen system. Another form is lasso and it is given by Tibshirani et al (1996). It is paying attention on large number of researchers and shows significant results in feature selection lately by Liu et al (2009); Meier et al (2008); Zhou et al (2012). Lasso penalizes the estimator with 1 norm. As a result, a sparse weight will be formed where most of the features will be specified as zero weight.

Group Lasso by Meier et al (2008) Overlapping Group Lasso by Jacob et al (2009); Yuan et al (2011) proposed Graph Lasso, and so on. A recent analysis related to lasso and its variations may be found here (Ye & Liu 2012). Filter model is known to be very resourceful and generally scalable and generalizable as it is independent of any classifier. Owing to these advantages most of the presented methods fit in to this model. On the other hand, it may not be as precise as wrapper model in particular if the classifier is known in advance. Well-known filter algorithms include: Information Gain, ReliefF, Chi Square, Gini Index, t-test, FCBF, CFS, MRMR, and so forth.

Unlike filter model, wrapper model uses a classifier to assess the quality of the selected features. It starts by selecting a subset of features, typically by means of greedy search approach. Then, the given classifier assesses the quality of the selected subset. If the quality is acceptable, the selection stops. Or else, it searches for another, perchance, improved subset. This is extremely expensive and time consuming approach comparing to the filter model. However, the selected features by means of wrapper model are more accurate with regard to the given classifier than the filter model. Different search approaches could be combined with any classifier and generate one probable wrapper feature selection method. For instance, Recursive Feature Elimination Support Vector Machine (RFE-SVM) is extensively used wrapper approach (Saeys et al 2007).

In order to triumph over the limitations of the earlier models, a hybrid model was presented to bridge the gap and to afford reasonably competent and accurate selection. It tags along filter model in the search step, where it selects small number of candidate subsets of features. Therefore, unlike wrapper approach, hybrid approach estimates the quality of small number of candidate subsets, which make possible of less complex model. The selected subset provides better classification accuracy. In view of that, the hybrid model is more well-organized than filter and less expensive than wrapper based approach. Related to wrapper, different grouping of filter criterion and classifiers may possibly produce novel hybrid techniques. For instance, Improved F-score and Sequential Forward Floating Search (IFSFFS) combine F-score with Sequential Forward Floating Search and SVM to achieve high accuracy in selection as described by Xie et al (2010). Likewise, Correlation-based Feature Selection with Taguchi-genetic algorithm (CFSTGA) achieved very high classification accuracy with KNN by Chuang et al (2011). An additional hybrid technique may be found in (D'Alessandro et al 2003; Oh et al 2004).

Various methods have been introduced to handle feature selection problem in the nonappearance form of class label. One of the familiar approaches is to generate labels without any human intervention for the given samples before selecting features. These generated labels, after that, will be used to direct the feature search alike as the supervised feature selection. Some other methods use k-means clustering to produce the labels (Boutsidis et al 2009; Jing et al 2007). At the same time other methods also employ more complicated approaches such as harnessing spectral analysis to haul out the underlying clusters (Cai et al 2010;Li et al 2012).

Spectral Feature Selection (Zhao &Liu 2007) is a model yet, it can handle supervised data in addition to unsupervised. Thus, it is a unified feature selection method. Entropy Weighting K-Means (EWKM) was presented for subspace clustering.

It concurrently minimizes the within-cluster dispersion and maximizes the negative weight entropy in the clustering process (Jing et al 2007). It uses k-means to find the clusters before doing feature selection. This step is repeated for a number of times until convergence. Cai et al (2010) proposed a Multi-Cluster Feature Selection (MCFS) that makes use of spectral analysis to determine the correlation between different features without label information required. With the top eigenvectors of graph Laplacian, spectral clustering can cluster data samples not including of utilizing label information. Some other methods are presented to evaluate feature's weight separately of any clustering techniques. Term Frequency (TF), Inverse Document Frequency (IDF) and TF-IDF are among the most well-liked feature weighting, (aka term selection) techniques particularly in text mining domain.

Analogous to feature selection for supervised learning, methods of feature selection for clustering are classified into filter wrapper (Roth &Lange

2003) and hybrid models (Dy 2008). A wrapper model measures the candidate feature subsets by the quality of clustering whereas filter model is self-determining of clustering algorithm. Therefore, the filter model is still preferable in terms of computational time and unbiased toward any clustering method, at the same time as the wrapper model produces better clustering if the author knows the clustering method in advance. To improve the computational cost in the wrapper model, filtering criterion is employed to select the candidate feature subsets in the hybrid model (Dy 2008).

Tin Kam Ho (1998) presented an approach known as Random Subspace Method (RSM). In this method, a subset of features was randomly chosen for each prediction model. The number of features selected for each prediction model was half the total number of features. The experimental results were commenced to compare the RSM to bagging, boosting and single tree prediction models employing all features.

Faud& Kittler (2000) introduced a technique that makes use of traditional feature selection algorithms with the intention of maximize the overall ensemble performance. The author presented about three different variations for structuring the ensemble: the parallel system, the serial system, and the optimized conventional system. In the parallel system, each expert (the term they used for a prediction model) is allowed, in turn, to take one feature such that the overall ensemble performance is optimized on a validation set. In the serial system, in contrast, the first expert is allowed to take all the features that achieve the maximum ensemble accuracy on the validation set.

In Lei Yu &Huan Liu (2003) Feature selection, as a preprocessing step to machine learning, has been effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving comprehensibility. However, the recent increase of dimensionality of data

poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. In this work, we introduce a novel concept, predominant correlation, and propose a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. The efficiency and effectiveness of our method is demonstrated through extensive comparisons with other methods using real-world data of high dimensionality.

Hsu et al (2002) studied the behavior and relationship between rank combinations by introducing a concept called rank/score graph. They showed that under certain condition rank combination outperformed score combination.

Chuang et al (2004) applied rank combination to combine different feature selection methods. The ranks of features are combined by using a weighted sum (or average) from each of the component rankings obtained from individual feature selection method. It is shown that the combination approach performs better than each individual feature selection method in many cases.

Existing algorithms are traditionally categorized as wrapper or filter methods, with respect to the criteria used to search for relevant features (Kohavi&John 1997). In wrapper methods, a classification algorithm is employed to evaluate the goodness of a selected feature subset, whereas in filter methods criterion functions evaluate feature subsets by their information content, typically interclass distance (e.g., Fisher score) or statistical measures (e.g., p-value of t-test), instead of optimizing the performance of any specific learning algorithm directly. Hence, filter methods are computationally more efficient, but usually they do not perform like wrapper methods.

The classification of high dimensional data with kernel methods is considered in this work (Fauvel et al 2013). Exploiting the emptiness property of high dimensional spaces, a kernel based on the Mahalanobis distance is proposed. The computation of the Mahalanobis distance requires the inversion of a covariance matrix. In high dimensional spaces, the estimated covariance matrix is ill-conditioned and its inversion is unstable or impossible. Using a parsimonious statistical model, namely the High Dimensional Discriminant Analysis model, the specific signal and noise subspaces are estimated for each considered class making the inverse of the class specific covariance matrix explicit and stable, leading to the definition of a parsimonious Mahalanobis kernel. SVM based framework is used for selecting the hyperparameters of the parsimonious Mahalanobis kernel by optimizing the so-called radius-margin bound. Experimental results on three high dimensional data sets show that the proposed kernel is suitable for classifying high dimensional data, providing better classification accuracies than that of the conventional Gaussian kernel.

Simon &Horst (2002) introduced an ensemble formation approach based on feature selection algorithms. They experimented their approach in the context of handwritten word recognition, using Hidden Markov Model (HMM) recognizer as the fundamental formation of prediction model. In this approach, each prediction model is specified as a well functioning set of features by means of any existing feature selection algorithm.

Oliveira et al (2003) presented a GA-based ensemble formation technique. The proposed technique also employed GA algorithm to identify the best feature subsets. On the other hand they used a hierarchical two-phase approach to ensemble creation. In the initial phase, a set of high-quality prediction models are formed by means of Multi-Objective Genetic Algorithm (MOGA) search. The pedestal prediction models used here were

Artificial Neural Networks (ANN), but any type of prediction model can be used. The second phase searches through the space fashioned by the different combinations of these high-quality prediction models, again using MOGA, to find the best possible combination i.e. the best ensemble

FRM approach considers the information given by numerous or even all the fuzzy rules in the system. To do so, we should consider the use of the Choquet integral as the aggregation operator in the FRM. The Choquet integral is associated to a fuzzy measure (David 1999), which models the interaction between the elements to be aggregated (the information given by the rules of the system in this case). Therefore, a key point is the choice of an appropriate fuzzy measure for each problem we want to deal with.

## 2.3 VARIOUS APPROACHES BASED ON CLASSIFICATION

The most popular class of clustering algorithms is k - means algorithm, a center based, simple, and fast algorithm, aims to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean (MacQueen 1967). However, in real applications there are no sharp boundaries within the clusters so that data objects might partially belong to multiple clusters. In fuzzy clustering, the data points can belong to more than one cluster and membership degrees between zero and one are used instead of crisp assignments of the data to clusters (Jain et al 1999). The degree of membership in the fuzzy clusters depends on the closeness of the data object to the cluster centers.

A Self Organizing Map (SOM) (Tamayo et al 1999) is more robust than K-means for clustering noisy data. Due to the noisy data there would be some miscalculation in the accuracy. The input required is the number of clusters and the grid layout of the neuron map. Prior identification of the number of clusters is tough for the gene expression data. Furthermore,

partitioning approaches are restricted to data of lower dimensionality, with intrinsic well-separated clusters of high density. Thus partitioning approaches do not perform well on high dimensional gene expression data sets with intersecting and embedded clusters. A hierarchical structure can also be built based on SOM such as Self-Organizing Tree Algorithm (SOTA) (Dopazo&Carazo 1997). Fuzzy Adaptive Resonance Theory (Fuzzy ART) (Tomida et al 2002) is another form of SOM which measures the coherence of a neuron (e.g., vigilance criterion). The output map is accustomed by splitting the existing neurons or adding new neurons into the map, until the coherence of each neuron in the map satisfies a user specified threshold.

Fuzzy c-means (FCM) which is introduced by Bezdek(1981) is the most popular fuzzy clustering algorithm. However, FCM is an effective algorithm; the random selection in center points makes iterative process falling into the local optimal solution easily. To tackle this problem, evolutionary algorithms such as genetic algorithm (GA), differential evolution (DE), ant colony optimization (ACO), and particle swarm optimization (PSO) have been successfully applied (Gulgezen et al 2009;Zhao 2007).

Semi-supervised learning methods construct classifiers using both labeled and unlabeled training data samples. While unlabeled data samples can help to improve the accuracy of trained models to certain extent, existing methods still face difficulties when labeled data is not sufficient and biased against the underlying data distribution. In this paper, we present a clustering based classification (CBC) approach. Using this approach, training data, including both the labeled and unlabeled data, is first clustered with the guidance of the labeled data. Some of unlabeled data samples are then labeled based on the clusters obtained. Discriminative classifiers can subsequently be trained with the expanded labeled dataset. The effectiveness of the proposed method is justified analytically. Related issues such as expanding labeled

dataset and interacting clustering with classification are discussed (HuaJun Zeng et al 2003).

Genetic algorithm is widely used for mining classification rules. If the data set is of three or four years old, the Artificial Bee Colony (ABC) optimization algorithm, which is described by Karaboga based on the foraging behavior of honey bees for numerical optimization problems (Karaboga 2005), is applied to classification benchmark problems (13 typical test databases). The performance of the ABC algorithm on clustering is compared with the results of the Particle Swarm Optimization (PSO) algorithm on the same data sets that are presented in (De et al 2007). ABC and PSO algorithms drop in the same class of artificial intelligence optimization algorithms, population-based algorithms and they are proposed by inspiration of swarm intelligence. Besides comparing the ABC algorithm and PSO algorithm, the performance of ABC algorithm is also compared with a wide set of classification techniques that are also given in (De et al 2007).

Decision trees are of influential classification technique. Mainly, two well-known algorithms called Classification and Regression Trees which are used for building decision trees are examined by Breiman et al (1984). ID3/C4.5 is introduced by Quinlan (1993). The tree attempt is to deduce a separation of the training data based on the values of the accessible features to produce a good generalization. The separation at each node is based on the feature that provides the maximum information gain. Each leaf node is based on corresponding class label. A novel example is classified by the subsequent path from the root node to a leaf node, where at each node a test is performed on some feature of that example. The leaf node reached is measured as the class label. The algorithm can obviously tackle over binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned.

KNN (Stephen 1998) is well thought-out among the oldest non-parametric classification algorithms. To categorize an unknown sample, the distance (by some distance measure e.g. Eculidean) from that sample to every other training sample is measured. The k smallest distances are identified, and the large amount are represented by class in these k classes is considered to be an output class label. The value of k is in general determined through a validation set or using cross-validation.

Chen et al (2008) employed an approach of clustering the classes into a binary tree called Hierarchical SVM (HSVM). Nevertheless, the clustering is performed by means of arranging classes into an undirected graph, with edge weights in place of the Kullback-Leibler distances between the classes, and employing a max-cut algorithm to divide the classes into two sub-clusters that are most far-away from each other. SVMs are used as the binary classifier at each node of the tree. The description is about improved performance versus bagged classifiers using remote sensing data.

Santi et al (2011) presented an algorithm based on clustering approach which is called as K-Mode RSVM (KMO-RSVM) for dealing with large definite dataset. The KMO-RSVM algorithm reduces support vectors by building k-mode clustering technique with RSVM. Applying k-mode clustering algorithm to each class can create cluster centroids of each class and exploit them to outline the reduced set which is used in RSVM.

Although research prolonged on the common classifier ensemble algorithms, efforts have been made by researchers to combine only one exact classifier in a classifier ensemble. Zheng (1998) explained about naive bayesian classifier ensembles. Various neural network ensembles were studied and examined thoroughly in (Amanda 1999).

Recursive Partitioning algorithms for classifier ensembles bring into play a divide-and-conquer strategy to division a space into regions that comprises instances of only one class. Utgoff (1989) made available of scheme examples for recursive partition ensemble algorithms. The work was about perception tree algorithms which combines a univariate decision tree with linear threshold units. It first determines if a subspace is linearly separable by using a heuristic measure. If the subspaces are linearly separable, then a linear threshold unit is applied. If not, the space is divided using an information theoretic measure. Brodley's model class selection system creates a recursive, tree-structured hybrid classifier which combines decision trees, linear discriminant functions and instance-based classifiers.

In recent times two ensemble classifiers, boosting (Schapire 1990) and bagging (Breiman 1996), have been broadly used. Both these methods make use of a resampled learning set to build a base classifier. Boosting modifies the distribution of the learning set based on earlier classifiers and combines weak classifiers by means of weighted voting. The author said that boosting sometimes fails, and the class distributions across the weight vectors turn out to be skewed. Bagging uses a bootstrap sample to construct each base classifier. Each sample is selected randomly with replacement and the final decision is prepared by equal weight voting of these base classifiers. Random Forest is based on this algorithm and in advance recognition. It combines classification trees which are constructed by bagging and random subspace of the predictors. Owing to their resampling algorithms, both bagging and boosting reason overlap of predictor variables along with classifiers, and consequently high correlation among the base classifiers.

By expanding the modern large scale classifier Power Mean SVM (PmSVM) proposed by Thanh-Nghi et al (2013) Doan in three ways: (1) An incremental learning for PmSVM, (2) A balanced bagging algorithm for

training binary classifiers, (3) Parallelize the training process of classifiers with numerous multicore computers. The proposed approach is evaluated on 1K classes of ImageNet (ILSVRC 1000). The evaluation shows that our approach can save up to 82.01% memory usage and the training process is 255 times faster than the original implementation and 1276 times faster than the state-of-the-art linear classifier (LIBLINEAR).

Linda & Manic (2009) proposed a solution to the problem of classifying large datasets through learning of the data structure. The algorithm is described as one which combines the GNG algorithm through the SVM solver into a detailed algorithm for classification of large datasets. The input dataset is initially preprocessed with the GNG algorithm. A new-fangled reduced training dataset is formed from the extracted topological knowledge. For the reason that the size of the dataset is considerably reduced, the training process of the SVM solver becomes significantly less memory challenging. The experimental result shows that the proposed GNG-SVM approach is tested on both synthetic and benchmark real world datasets.

In Ungurean et al (2012) propose an efficient algorithm to apply computing resources given by a CBEA-based cluster through parallelization and optimization of an algorithm for classification of a large dataset. In order to examine the proposed approach of parallelization, the algorithm is carried out on a CBEA-based cluster, which comprises of 96 PowerXCell 8i processors (among theoretical peak performance of 9.83TFlops). The author examines the execution time on a processor and on all 96 processors, with and without utilization of the SPE cores.

The Supervised Self-Organizing Map (SSOM) approach is based on the neural network classification, which is an efficient approach and improved image classification accuracy, with the synthetic dataset is used to evaluate the classification uncertainty (Lawawirojwong et al 2013). Monte

Carlo simulation technique is applied to assess the reliability of the classification output by focusing on the uncertainty associated with the input data, training data, and the classifier. The results indicate that increasing the levels of noise have an extensive influence on the classification accuracy. SSOM with different sequences of training data produces the variation of classification accuracy.

### 2.3.1 Ensemble Classification

Al-Khateeb et al (2012) propose a novel class-based Micro-Classifier Ensemble classification technique (MCE) for classifying data streams. Traditional ensemble-based data stream classification techniques build a classification model from each data chunk and keep an ensemble of such models. Due to the fixed length of the ensemble, when a new model is trained, one existing model is discarded. This creates several problems. First, if a class disappears from the stream and reappears after a long time, it would be misclassified if a majority of the classifiers in the ensemble does not contain any model of that class. Second, discarding a model means discarding the corresponding data chunk completely. However, knowledge obtained from some classes might be still useful and if they are discarded, the overall error rate would increase.

A novel ensemble classification method is proposed named as SREC (Simple Rule-based Ensemble Classifiers) (Hualong Yu &SenXu, 2011). Firstly, the classification contribution of each gene is evaluated by a novel strategy and the corresponding classification rule is extracted. Then we rank all genes to select some important ones. At last, the rules of the selected genes are assembled by weighted-voting to make decision for testing samples. It has been demonstrated that the proposed method may improve classification accuracy with lower time-complexity than that of the traditional classification methods.

An ensemble text classification model is proposed by combining classification rules and N-gram language model (Jinhong Liu &Yuliang Lu 2007). In order to generate strong classification rules, we propose an exhaustive noun-phrase extraction algorithm and a new optimized rule induction algorithm, called SCA (strong covering algorithms). We also introduce an improved good-Turing (GT) smoothing method for N-gram model. Experimental results show that our ensemble classifier achieves an approximately 8% improvement as compared to bi-gram with word-based classifier and 15% improvement as compared to traditional rules-based classifier.

In A novel ensemble classification method task is described by Vannucci et al (2012). The proposed approach is based on the use of a set of classifiers, each of which is trained by exploiting a different subset of the available training data, which are created by partitioning the input space by means of a Self Organizing Map (SOM) based clustering algorithm. Subsequently, the reliability of each classifier belonging to the ensemble is measured according to the classification accuracy on whole dataset and each classifier is associated to a feed forward neural network, which is able to self-estimate the reliability of single classifiers when coping with a new data. The estimated reliabilities are used in the ensemble aggregation phase in order to provide the final classification of new patterns. The method, tested on literature datasets coming from the UCI repository, achieved satisfactory results improving the classification accuracy with respect to other popular ensemble techniques.

Recently, many classifiers were developed exploring various fields with the help of computer science (Bharathi&Natarajan 2010). In fact, most of the research work found in the literature related to disease classification either makes use of statistical models or artificial neural networks. Statistical

methods such as linear discriminate analysis, generalized linear regression such as logistic regression, and nearest neighbor classification are widely used. There are many methods and algorithms used to mine biomedical datasets for hidden information. They include NNs, Decision Trees (DT), FL Systems, Naive Bayes, SVM, cauterization, logistic regression and so on. On studying the literature it turns out that the most frequent choices for the medical decision support systems are the DT (C4.5 algorithm), NNs and the Naive Bayes. These algorithms are very useful in medicine because they can reduce the time spent for processing symptoms and producing diagnoses, making them more precise at the same time. Also, many of the research assessed the algorithms on a narrow set of medical databases. However and to the best of our knowledge ensembles of these techniques have not been used in the bioinformatics dataset classification.

NNs are networks of units, called neurons that exchange information in the form of numerical values with each other via synaptic interconnections, inspired by the biological neural networks of the human brain. They become very powerful and flexible approaches to function approximation. NNs mainly refer to the feed forward networks such as multilayer perceptrons and radial basis function neural networks, which have been widely used to develop diagnostic models. In order to improve the costs benefit ratio of breast cancer screenings, authors of (Tarek et al 2009) evaluated the performance of a back-propagation NN to predict an outcome (cancer/not cancer) to be used as classifier. NNs were trained on data from family history of cancer, and socio demographic, gyneco obstetric and dietary variables. Research is going on in capitalizing the use of NNs in medical diagnosis of breast cancer. This work indicates that statistical NNs can be effectively used for breast cancer diagnosis to help oncologists in which classification is based on a feed forward NN rule extraction algorithm. General regression NN, or probabilistic NN was used in order to get the

suitable result. The problem with NNs is that they usually adopt gradient-based learning methods which are susceptible to local minima and long training times especially when the number of classes/categories is high.

The authors Van et al (2007) introduce artificial NNs with back propagation for classification of heart disease cases. This solution is implemented in a medical system to support the classification of the Doppler signals in cardiology. The predictions yielded by the method were more accurate than similar presented in Turkoglu et al(2002). The NNs' major disadvantage is complexity, which makes classification process difficult to interpret. Nevertheless, the authors prove that they produce effective classifications in case of medical data. As far as NN is concerned, the influence of the noisy inputs on the output variable together with the transfer functions, implicit in the values of the weights. Hence an unattractive feature of such networks is that the number of weights and complexity increase greatly as the network grows. Also the weights may not always be easy to interpret if the data is imprecise and uncertain which leads to the problem of under fitting or over fitting and the problem becomes difficult to visualize from an examination of the weights.

SVM has been proposed as a very effective method for pattern recognition, machine learning and data mining (Cortes &Vapnik 1995). The general idea is to map non-linearly D-dimensional input space into a high dimensional feature space. A linear classifier (separating hyper plane) is constructed in this high dimensional space to classify the data. The use of the kernel trick allows constructing the classifier without explicitly knowing the feature space. It is considered to be a good candidate because of its high generalization performance.

Intuitively given a set of points which belong to either one of the two classes, a SVM can find a hyper plane having the largest possible fraction

of points of the same class on the same plane. This hyper plane called the optimal separating hyper plane (OSH) can minimize the risk of misclassifying examples of the test set. SVM, when using One-Versus-All (OVA) approach to make binary classifiers applicable to multi category problems, it can be seen that, when the number of classes increases, the complexity of the overall classifier also increases. So the system becomes more complex and requires extra computations. In SVM classifiers, problems with corrupted inputs are more difficult than problems with no input uncertainty. Even if there is a large margin separator for the original uncorrupted inputs, the observed noisy data may become non-separable. For example by using a kernel function in SVM, the input vector is mapped to in a usually high dimensional feature space and the uncertainty in the input data introduces uncertainties in the feature space. To overcome this problem, researchers used total least square regression methods with SVM but could not achieve promising results. FNs are extensions of NN which consist of different layers of neurons connected by links. Each computing unit or neuron performs a simple calculation: a scalar typically monotone function f of a weighted sum of inputs. The function f, associated with the neurons, is fixed and the weights are learned from data using some well-known algorithms such as the least-square fitting. A FN consists of a layer of containing the input data; a layer of output units containing the output data; one or several layers of neurons or computing units which evaluate a set of input values, coming from the input units, and which give a set of output values to the output units. The computing units are connected to each other, in the sense that the output from one unit can serve as part of the input to another neuron. Once the input values are given, the output is determined by the neuron type, which can be defined by a function (Tarek et al 2010;Castillo 1998).

Type-2 FLS was introduced as an extension of the concept of Type-1 FLS (Mendel et al 1999). Type-2 FLS has membership grades that are

themselves fuzzy. For each value of a primary variable (e.g., pressure and temperature), the membership is a function (not just a point value). The secondary Membership Function (MF) has its domain in the interval (0, 1), and its range may also be in (0, 1). Hence, the MF of a Type-2 FLS is three dimensional, and it is the newly introduced third dimension that provides new degrees of design freedom for handling uncertainties. Type-2 FLS does not obtain good performance when the number of training data is small, but it can perform better when the number of training prototypes is large.

Jerzy et al (2002) proposed an experimental study of using the rule induction algorithm MODLEM in the multiple classifier schemes called combiner. The main aim of this proposed method is to improve the classification accuracy. Experiments are done over various benchmark datasets and found that the combiner classifier is having higher classification accuracy than the single classifiers.

Jerzy et al (2001) proposed another combiner method. In this method two rough sets based filtering approaches combined with rule based classifiers. It is mainly suited for handling imbalance datasets. This work shows a higher improvement in sensitivity and gain. Based on the various characteristics of input data, Sohn et al (2007) proposed a method to compare the performance of classifier methods using logistic regression. The combination method includes modified random subspace method, bagging, parametric fusion, classifier selection. Taguchi design has been used for typically unknown combination function among input variables. Monte carlosimulation is used to improve the classification accuracy of classifier combination methods based on various data characteristics.

Bikash et al (2011) present a hybrid classifier called DTGA (Decision Tree and Genetic Algorithm) which is a combination of C4.5 and GA as GA is more suitable for getting more optimized solutions. The

experiments are done on UCI repository datasets. In DTGA, the dataset is first passed to C4.5 to generate rules. After discretizing the rules, GA is applied to refine the rules. This proposed model increases the classification accuracy and is able to classify imbalance datasets.

Deborahet al (2005) proposed a hybrid decision tree/ GA method. In this hybrid approach, two specifically designed GA algorithms are used for discovering rules of examples belonging to small disjuncts and conventional decision tree algorithm are used for producing rules of examples belonging to large disjuncts. The advantage of this hybrid method is that they are flexible and robust. This hybrid method is having two versions, C4.5/GA-Small and C4.5/GA-Large-SN. The performance of both the versions are compared with C4.5 and "double C4.5". And the better results are shown by the C4.5/GA-Large-SN and it has been considered as the best solution for small disjunct problems.

Deborah et al (2007) derived rules to predict the class based on the value of the attributes. These rules decrease the classification accuracy and they are error prone. For overcoming all the limitations, a hybrid decision-tree/ GA approach is proposed. And the performance result of this hybrid model has been compared with three versions of C4.5 over eight domain data sets. In all the comparisons the hybrid model achieves better predictive accuracy.

For discovering small-disjunct rules in the form "If P Then D", a classification algorithm based on Evolutionary Algorithm (EA) has been proposed by Basheer et al (2012). The experiments are carried out over several UCI dataset repositories. From large datasets, the small disjunct rules are generated using this successful application of GA. It is having appropriate crossover and mutation operators, flexible chromosome encoding, and

suitable fitness function. The results show that the proposed algorithm is much more efficient for better rule extraction.

Zhang & Zhang (2008) proposed a novel ensemble classifier generation method RotBoost through combining Rotation Forest and AdaBoost. In this new ensemble method, the base classifier in Rotation Forest algorithm is replaced with AdaBoost. The experimental results show that RotBoost performs better than either Rotation Forest or AdaBoost when using some non-microarray gene-related data sets from the UCI repository.

Lazy Bagging is one of Lazy ensemble approach used to build Lazy learner that build replicate bootstrap bags and uses same “type” of classifiers for prediction. So it is a good experience to build lazy learner but at some point it lacks diversity. Accuracy of classifier model in form of diversity is depending on type of base classifiers. If same type of base classifier is used then it contains similar prediction that reduces diversity, It. it is better to use different “type” of classifiers as base classifiers described by Xingquan (2008).

In this paper algorithm is proposed to address the class imbalance problem. The artificial data are created according to the distribution of the training dataset to make the ensemble diverse, and the random subspace re-sampling method is used to reduce the data dimension. In selecting member classifiers based on misclassification cost estimation, the minority class is assigned with higher weights for misclassification costs, while each testing sample has a variable penalty factor to induce the ensemble to correct current error. In our experiments with UCI disease datasets, instead of classification accuracy, F-value and G-means are used as the evaluation rule.

In this paper a recent technique for classification of datasets was investigated. One of the major factors to evaluate a classifier depends on how

accurately it can classify unknown patterns. There are a number of classification algorithms, both supervised and unsupervised. In most cases, a single classifier is trained on a part of the dataset and tested on the remaining part of the same dataset. It is observed that a single classifier performing excellently for the particular part of a dataset produces poor classification accuracy when presented with another part of the same dataset. In this paper, an ensemble approach of classification is presented. Consequently, various parts of the dataset are trained using individual k-nearest neighbors (k-nn) classifiers. Using bagging and majority of voting techniques, the classification accuracy of test dataset is evaluated. Three benchmark datasets are used for empirical study of the scheme. After extensive experiments with three ensembles having different number of classifiers, three different values of k for three different sizes of training datasets, selected randomly on five trials, it is observed that the ensemble of classifiers produces better classification accuracies than that of any individual classifier.

Harries et al (1998) described a methodology that identifies concepts by grouping classifiers of similar performance on specific time intervals is described. Weights are then assigned to classifiers according to performance on the latest batch of data. Predictions are made by using weighted averaging. Although this strategy fits very well with the recurring contexts problem, it has an offline step for the discovery of concepts that is not suitable for data streams. In particular, this framework will probably be inaccurate with concepts that did not appear in the training set.

An interesting idea is presented by Forman (2006), where a great number of “daily” classifiers are maintained in memory. Their predictions for incoming data are incorporated as additional features into the most recent classifier which will learn how to take advantage of them in order to make future classifications. Unfortunately, the framework is not designed for data

streams and in addition classifiers are built on small sample sets as they are not organized into concepts.

In this paper, we present BENCH (Biclustering driven Ensemble of Classifiers), an algorithm to construct an ensemble of classifiers through concurrent feature and data point selection guided by unsupervised knowledge obtained from biclustering. BENCH is designed for underdetermined problems. In our experiments, we use Bayesian Belief Network (BBN) classifiers as base classifiers in the ensemble; however, BENCH can be applied to other classification models as well. We show that BENCH is able to increase prediction accuracy of a single classifier and traditional ensemble of classifiers by up to 15% on three microarray datasets using various weighting schemes for combining individual predictions in the ensemble.

Hua-Jun et al (2003) applied ensemble classifiers to mining data streams. Data streams consist of data that is constantly being read in from a streaming source. In their method, they divide the data stream into chunks and then use these chunks to train the classifiers. Wang et al. also use the predicted error rate to determine the weight of each derived classifier in the ensemble. This method can be viewed as sampling data points from the full streaming dataset in order to create an ensemble. In their experiments, they used several types of base classifiers, naive Bayesian networks, RIPPER, and the C4.5 decision tree algorithm, on a dataset containing credit card transaction records for a one year period. They show that, with enough classifiers, the ensemble of classifiers will outperform single classifiers with respect to both speed and accuracy. However, this method has not been tested for underdetermined problems.

Jing et al (2008) combined parameter boosting with structure learning to improve the classification accuracy of BBN classifiers. They

construct an ensemble of BBN classifiers, starting with an empty set, and the algorithm goes though fixed number of iterations or stops if some criterion is met. At the beginning of each iteration, a training set and the set of corresponding weights for the data points are given to the TAN algorithm to build a BBN classifier. For base classifier i, the TAN algorithm adds the i edges with the highest mutual information to a na¨ıve BBN. The training error of the resulting TAN classifier is then used to determine the weight of the test data points in the next iterations. The algorithm will stop at this point if the training error increases. Jing et al. test their ensemble of classifiers using UCI datasets as well as two artificial datasets. According to their results, their boosted BBNs have comparable or reduced average testing error than naive BBN, TAN, and ELR on the 23 UCI datasets and the simulated datasets.

## 2.4 SUMMARY

This chapter discusses about the existing classification techniques available in the literature. The characteristic features of the existing classification approaches are thoroughly analyzed in this chapter. This chapter also discusses about the ensemble classification and its significance in various applications. Based on the available approaches, a novel ensemble based classification is developed in the present research work.