CHAPTER 2

REVIEW OF LITERATURE

2.1 INTRODUCTION

The purpose of the literature review is to increase the breadth of knowledge in the area of research and to identify the gaps in current knowledge. The literature review is part of any reliable research and it provides the research a prospect to review similar and related previous research to gain insights. The detailed review of the existing methods may specify the drawbacks of the methodologies used and it gives a challenge for predicting the novel approaches with the focus on the better improvement of the performance. The literature survey gives the detail description of the previous work with the summary and the critical evaluation of the work. The survey describes the existing methods with what are all the methodologies used and the dataset taken for experimental results. The literature survey was carried out in three phases on the hybrid feature subset selection, ensemble method of effective AdaBoost and merging both feature subsets into the ensemble method. The review of the literature provides the detail description of the existing framework with the summary and practical assessment.

This chapter briefly describes the relevant works in Hybrid attribute selection techniques implemented with various feature selection approaches and
AdaBoost ensemble methods. Section 2.2 provides the literature survey on Hybrid attribute selection methods, and section 2.3 entails work carried out in enhancing the Decision tree. Section 2.4 lists the survey on ensemble classifier and section 2.5 contains reviews about the AdaBoost algorithms.

2.2 HYBRID FEATURE SELECTION

Hui-Huang et al. (2011) have proposed a hybrid feature selection technique by merging feature selection approaches wrappers and filters. Filter method’s F-score and information gain (IG) are applied separately on a dataset to eliminate the redundant and irrelevant feature. Intersection (AND) and exclusive-OR (XOR) are used to combine the resulting two feature sets. Finally, the wrapper approach armed with inversed sequential floating search technique is applied to improve the classification precision. The authors have evaluated their algorithm by measuring the classification accuracy in diagnosing acute myeloid leukemia (AML) in protein disordered dataset and gene selection in the lung cancer dataset.

Feature selection improves the efficiency and accuracy of spam filtering. As the document frequency information based feature selections does not take into account the effect of term frequency information it always deduces unsatisfactory results. To address this problem Yuanning Liu et al. (2014) have proposed a hybrid method (called HBM), which combines the document frequency information and term frequency information. The stop words and execute words in the document are removed using a Porter Stemming algorithm. Optimal document frequency-based feature selection method is used in feature selection. Next term frequency-based feature selection (TFFS)
method is used to select the additional features by a TFFS. The authors have proposed a new threshold parameter optimization method, called feature subset evaluating parameter optimization (FSEPO). FSEPO determines whether considering the term frequency information in the feature selection process or not. It combines the high accuracy of wrapper metric in feature selection. Moreover, FSEPO takes less time on searching the optimal threshold and precision than grid search method.

Zhongyi Hu et al. (2015) have proposed a hybrid filter–wrapper methodology for short term load predicting feature selection. In this approach, partial mutual information based filter approach is used to separate the redundant and irrelevant attributes. In order to further decrease the redundant or irrelevant features wrapper approach is applied through a firefly procedure. The authors have chosen support vector regression as the modeler to implement their approach. Real-time energy load datasets from a North American electrical value and the global energy predicting competition 2012 datasets were used to evaluate the accurateness of the suggested method.

To improve the classification accuracy in the classification of the cancer tissues in breast mammograms Aswini Kumar et al. (2013) have proposed a hybrid attribute selection model. The proposed method uses branch and bound method to identify and select optimal attributes and genetic algorithm (GA) to select the most excellent features with crossover and mutation methods to combine. The classification accuracy of the proposed method outperformed the classification accuracy of well-known classifiers like K-nearest neighbor (KNN), SVM, rough set, artificial neural network (ANN), decision tree (DT),
and ant miner. The experiment was conducted with a dataset of 300 mammograms images.

The hybrid feature selection procedure proposed by Li-Yeh et al. (2011) deals with Deoxyribonucleic acid (DNA) microarray data. To reduce redundant features and to achieve excellent classification accuracies, the authors have combined CFS and the Taguchi GA technique. In addition they have used KNN with the leave-one-out cross-validation (LOOCV) as a classifier to classify eleven classification profiles and to estimate the classification accuracy. Eleven gene expression data sets were taken as benchmark data sets to test the proposed method against various classifiers like non-SVM with a neural network (NN), KNN and probabilistic NN, multi-class SVM with back propagation NN. The result of the experimental study is found to be better compared to other methods considered the experimental study.

A two-stage hybrid model for identifying erythematous squamous diseases was proposed by Juanying Xie et al. (2013). The suggested model implements classification tool SVM based on searching techniques such as sequential backward floating search (SBFS), sequential forward search (SFS) and sequential forward floating search (SFFS). To evaluate the importance of separate feature the generalized F-score is used. 10 selected feature subsets obtained after conducting the 10-cross validation experiments are combined into one new subset. The hybrid techniques formed by combining Generalized F-Score (GF) with new SBFS, SFS and SFFS are used as the feature selection techniques on the new subset. The authors have demonstrated that classification accuracy of the proposed method is better than the available methods for diagnosing erythematous-squamous diseases.
Chien-Pang Lee and Yungho Leu (2011) have proposed a novel hybrid technique for attribute selection in microarray data analysis. GA with the dynamic parameter is used to build few feature subsets and to rank the features as by their frequency of occurrences in the subsets of features. Next they use the $x^2$ test method to choose high ranked features for data analysis. Finally, SVM is used to confirm the effectiveness of the selected features. Six types of microarray datasets are used to evaluate the performance of the proposed technique against the existing methods.

Haitao Liu et al. (2013) have proposed a hybrid attribute selection method to handle mixed features dataset. First the authors suggest a new method for computing the mutual information between discrete and continuous variables. This mutual information is used to coin a new correlation measure for handling mixed attribute data. Features are filtered using the new correlation measure. A Case-base reasoning approach is proposed to finalize the number of parameters in the filtered set. The proposed method is found to be more stable, interpretable and with better estimation accuracy.

Hannah et al. (2014) have proposed supervised particle swarm optimization (PSO) based Relative Reduct (SPSO-RR) and supervised PSO based Quick Reduct (SPSO-QR) for the disease diagnosis. The proposed SPSO-RR algorithm calculates a reduct set without generating all possible subset. For each particle $X_i$, the dependency of selected features is computed based on dependency of decision features. If the dependency is not equal to 1, Pbest the highest relative dependency value of each particle is retained and the best value of the entire population is retained as the global best value. The SPSO-QR algorithm computes a reduct set. It starts with an empty set and it adds one
particle at a time till the maximum iterations is completed. In every iteration a feature with highest fitness is selected and fitness of all possible combinations of the selected feature with the other features is calculated. If the current particle’s fitness evaluation is better than the Pbest, then this particle becomes the current best which is compared with population’s overall previous best fitness. If the current value is better than gbest, then this becomes the best feature subset encountered so far, and is stored in R. The dependency of each attribute subset is calculated based on the dependency on decision attribute and the best particle is chosen. The experimental analysis clearly proved that proposed approaches diagnosis the disease better over the other existing approaches.

Siow-Wee Chang et al. (2013) proposed an oral cancer prognosis model. The authors designed a hybrid model by combining attribute selection and machine learning methods. In stage one, five feature selection approaches, namely Pearson’s correlation coefficient (CC) and reliefF as the filter model, GA as the wrapper model and the hybrid methods ReliefF-GA and CC-GA approaches were applied to the oral cancer prediction dataset. Features selected by the Classifiers artificial NN, adaptive neuro-fuzzy inference system, logistic regression and SVM are applied in the second stage. The authors, after experimenting with all combinations of feature selection methods with classifiers found that the hybrid model of ReliefF-GA-ANFIS (Neuro Fuzzy Inference Systems) resulted in the best accuracy.

Yang Yang et al. (2011) improved the accuracy of fault diagnostics by using a hybrid attribute selection scheme. By measuring the significance and similarity of the features is respect to the multiple clustering solutions. In the
The proposed method the original feature space is divided into two features and hyper-spherical cluster and hierarchical cluster are used to classify data instances in the two subsets. The significance of the instances in the two subsets is computed using linear correlation coefficient, the symmetrical uncertainty (SU) and the Davies–Bouldin index. Based on the significance all features are ranked. These processes are repeated until populations of feature rankings, named sub-decisions, are achieved. Then the subdivisions are combined using combiners majority vote, simple average and weighted average. Finally, those high ranked but less independent features are eliminated. The experimental study proved that HFS scheme is superior to feature ranking from multiple view with respect to the same evaluation criterion.

Wenzhi Zhu et al. (2013) suggested feature evaluation criteria the neighborhood effective information ratio (NEIR). The proposed method after analyzing the neighborhood granulation and the structure information of neighborhood granules the neighborhood decision uncertainty (NDU) of each information granule is established based on the decision distribution of corresponding granule. The neighborhood effective information ratio (NEIR) is established as the sum of the NDU values of the information granules induced by the corresponding feature space. The experimental study demonstrates that the proposed algorithm is able to select subsets of better quality with similar subset sizes, sometimes even with smaller sizes.

Kusum and Pramod (2015) proposed a hybrid dimension reduction method by incorporating attribute selection with a feature extraction technique for text clustering. As the union selects features from various subsets there is a possibility of selecting more features or entire features. Similarly, intersection
technique may ignore few important features. To address these problems authors have proposed a novel integration method known as a modified union. The modified union selects top ranked features and the intersection is applied on remaining feature subsets. The attribute selection techniques term variance (TV) and document frequency (DF) is applied for calculating attributes significance score. Next, Principal component analysis (PCA) a feature extraction technique is used to reduce the dimensions of the feature subset. Three different datasets used such as a worldwide knowledge base (WebKB), classic4 and Reuters-21, 578 for evaluating the performance of the proposed method.

Senthamarai Kannan and Ramaraj (2010) suggested memetic framework (MA-C) which combines Genetic algorithm and Local search. The local search in the proposed framework ranks the feature using correlation values. The authors have designed the framework with objectives to improve classification performance and to accelerate the search to identify important feature subsets. To accelerate the search the authors have implemented symmetric uncertainty (SU) based correlation measure based on the information-theoretical concept of entropy. From the comparative study of experimental results it is inferred that redundant attributes are removed efficiently MA-C shows compared to SU-CFS, genetic algorithm (GA) with Naïve Bayes as subset selection criterion and WFFSA-R wrapper-filter (WFFSA-R) feature selection algorithm with reliefF as filter ranking method and GA as wrapper method. Moreover, it is also inferred that as the number of attributes increases, the reduction in attribute and efficiency of resulting attributes is better than feature selection algorithms. The gene expression datasets namely Breast, central nervous system (CNS), Leukemia, Leukemia_3c, Leukemia_4c, Ovarian, small round blue cell tumors
(SRBCT), and mixed-lineage leukemia (MLL) were used for the experiential study.

Mansour and Najmehi (2013) have suggested a method for time series forecast. This hybrid method uses the combination of genetic algorithm (GA) and ant colony optimization (ACO) algorithm for feature selection. As the number of neurons in the hidden layer of NN plays an important role in its performance PSO is used to find the optimal number of neurons in the hidden layer. This PSO optimized NN is used as a classifier. Experimental results show that the proposed hybrid model offers superior performance, in terms of mean absolute percentage error (MAPE), in time series prediction as compared to some recent researches in this field.

Pablo Bermejo et al. (2011) have proposed a Greedy Randomized Adaptive Search Procedure (GRASP) procedure for hybrid attribute subset selection in high dimensional datasets. GRASP is a multi-start method. The first phase of the multi-start method is known as construction phase and it starts with an empty subset and elements are added to it from all the possible candidates till a solution is achieved. To achieve a greedy randomized construction process, the algorithm selects at random from a list of the promising ones. The second phase is the improving phase, which works in direction to develop an improved solution. Local search is used to improved solution. Twelve high-dimensional datasets are taken for experimental study. Performance of FSS algorithms: Incremental wrapper-based subset selection (IWSS), IWSS with replacement (IWSSr), best agglomerative ranked subset (BARS), Linear Forward Selection (LFS), Fast Correlation-Based Filter (FCBF) and PCA are compared with four wrappers and two filter approaches.
2.2.1 Information Gain

Zhang Si and Song-Chun Zhu (2012) proposed a framework Hybrid Image Template (HIT) for learning a generative image from a small set of images. If the image patches feature statistics are found to be consistent, distinctive from negative patches and has less intra class variations, then the higher information gain is assigned to them. The learning stops once the information reaches the permitted statistical fluctuation. The classification performance of hybrid image templates produced was evaluated using histogram of oriented gradients (HoG) + SVM.

Harun Uguz (2011) proposed a two-phase attribute selection technique for text categorization by applying IG, PCA and GA to improve the text categorization accuracies. In stage one the texts found in the document were ranked using the Information Gain. In the next phase PCA and GA attribute selection and attribute extraction approaches are applied to the texts arranged in decreasing order of rank. When the result of the proposed method was compared with the result of C4.5 decision tree technique and KNN it is found that the proposed approach not only reduced the complexity of categorization but also the computational time.

Alok Sharma et al. (2014) proposed an attribute selection technique by applying enhanced normalized linear discriminant analysis technique to select dominant genes or features that are essential for human cancer classification. Classification accuracy of the proposed method is compared with various existing feature selection approaches such as IG, sum minority, gini index (GI), max minority, sum of variance and improved regularized linear discriminant
analysis (RLDA) and the classifiers J4.8, NB, KNN and SVM respectively. SRBCT, MLL, and Acute Leukemia datasets were used for experimental study.

Gaber et al. (2013) suggested an entropy-based approach to improve the overall performance of the Rain Forest (RF). Usually the number of trees and split in trees is used as parameters in RF. The authors have used information gain IG in estimating the predictive capability of feature. They have proved that that the predictive power of the features can be used in setting the spit in Random Forests.

Harun Uguz (2012) introduced a hybrid scheme based on IG and PCA for the classification of transcranial doppler signals. The hybrid scheme has significantly improved the classification efficiency and accuracy. The proposed method uses an IG technique for feature ranking and PCA technique for dimension reduction. Based on the feature’s importance for the classification IG technique ranked the features in the feature area. This ranking system eliminated least essential features and selected the highly ranked features. The PCA technique was used further to reduce the dimension of the highly ranked feature set. Transcranial doppler signals documented from the time-based section of the brain for 82 patients, also of 24 healthy persons were used to study the classification efficiency and accuracy of the proposed method and SVM classifier.
2.2.2 ReliefF

As large quantity of irrelevant information is gathered, selecting the most informative features may help users to understand the task, and enhance the performance of the models. Many feature selection algorithms incorporating margin based loss functions and various search strategies were developed. These algorithms were used to differentiate irrelevant and relevant features. However, there is no proper proof to establish the effectiveness of these algorithms. Pan Wei et al. (2014) compare 14 margin based feature selections from the viewpoints of reduction capability, classification performance of reduced data and robustness. After conducting experimental study, the authors concludes that greedy feature search method by minimizing loss function (GFS-MCL) have less the number of features and can acquire higher level of classification precision, feature selection method based on gradient descent by minimizing loss functions (FWL-GD) is the fastest and deals with high dimensional data or large-sample data, while GFS-MCL is the slowest one, Simba, ReliefF and FWL-GD takes the almost same time in selecting features and the features selected by GFS-MCL are much stronger than those by other algorithms.

Qinbao Song et al. (2013) have introduced a rapid clustering-based feature subset selection algorithm to handle high dimensional image, microarray and text data. Their algorithm works in two steps. In the first step minimum spanning clustering algorithm is used to distribute the features of the original dataset to different clusters. In stage two features that are found to be strongly bonded to that respective cluster is selected and a new feature subset is formed. The authors have compared their proposed method’s results with feature selection methods, such as reliefF, fast correlation-based feature and
CFS with respect to classifiers such as, instance-based (IB1), probability-based NB, rule-based RIPPER (Repeated Incremental Pruning to Produce Error Reduction), and tree-based C4.5. Their comparative study reveals that their work not only produces the most relevant feature subsets by also plays an important role in improving the performance of the classifiers.

Oscar Reyes et al. (2015) suggested a three ways of extending of the relief algorithm. The extensions were proposed to handle applications that contain multi-label data. The reliefF algorithm is a standard algorithm for feature estimation as well as a suitable algorithm for all domains. The three extensions proposed by the authors are reliefF multi label, ReliefF-ML pruned problem transformation-reliefF (PPT-ReliefF) and regression reliefF-multi-label (RReliefF-ML). Pruned problem transformation-reliefF transforms the multi-label issue into a single-label issue. ReliefF-multi-label and regression reliefF-multi-label are designed to directly handle multi label data.

Sushmita Paul and Pradipta Maji (2014) have proposed a gene ontology based quantitative index, termed as degree of functional diversity (DoFD). DoFD is used to quantify the functional diversity of a set of genes selected by any gene selection algorithm. The authors have also combined merits of both DoFD and rough set based maximum relevance-maximum significance (RSMRMS) and developed a new gene selection algorithm. The proposed new gene selection algorithm differentiates the relevant and significant genes those are also functionally irrelevant. The performance of various gene selection algorithms such as CFS, reliefF, Minimum Redundancy Maximum Relevance (mRMR) etc, was compared with a proposed method to measure the classification accuracy. In the experimental study conducted with 14 datasets
DoFD and RSMRMS based hybrid gene selection algorithm attains highest classification accuracy in 7 cases. All other cases, the accuracy is comparable with the highest accuracy obtained using any other existing gene selection algorithm.

Zilin Zeng et al. (2015) proposed a novel attribute selection approach based on neighborhood rough sets that can be used to search for interacting features. In this paper, feature relevance, feature redundancy, and feature interaction are defined in the framework of neighborhood rough sets. The neighborhood interaction weight factor reflecting whether a feature is redundant or interactive is proposed, and a neighborhood interaction weight based feature selection algorithm based on the neighborhood interaction gain (NIWFS) is designed. The authors compute the neighborhood mutual information between a feature and the class and then adjust it through the manipulation of interaction weight factor to evaluate whether the feature is redundant or interactive. They used adjusted relevance measure to assign a rank to features. The results from ten real world datasets indicate that NIWFS not only deals with mixed datasets directly, but also reduces the dimensionality of the feature space with the highest average accuracies.

2.2.3 Correlation-Based Feature Selection

Marie Galligan et al. (2013) proposed a methodology for feature selection which uses a greedy search algorithm, based on the generalized Dirichlet distribution. The proposed method is applied to two glycan chromatography datasets. The proposed feature selection method performs well for both glycan chromatography datasets compared to correlation-based feature
selection (CFS) method and the learning tree algorithm, recursive partitioning (rpart). The proposed feature selection method performs well for both glycan chromatography datasets. It is computationally slower, but results in a lower misclassification rate and a higher sensitivity rate than both correlation-based feature selection and the classification tree method.

Bolon-Canedo et al. (2014) have proposed a new general framework by adding a new term $\lambda$ to the evaluation function of the filter methods: Correlation-based Feature Selection (CFS) and Minimal-Redundancy-Maximal-Relevance (mRMR). If $\lambda$ is 0, it is ignored. If $\lambda$ is in the range 0 and 1, the influence of the cost is smaller and if it’s 1 the cost terms have the same influence. If $\lambda$ greater 1, the influence of the cost is greater than the influence of the other term. Classifier Support Vector Machine (SVM) is used in studying the behavior of the proposed framework. 17 heterogeneous classification datasets are used in studying the behavior of the proposed method. The results of the experimental studies show that the approach is sound and it allows the user to reduce the cost without compromising the classification error.

Omar et al. (2013) proposed a novel machine-learning and feature-selection algorithms to study the flare-prediction-capability of magnetic feature (MF) properties generated by solar monitor active region tracker (SMART). The authors determined the flare-prediction-capability of all 21 SMART MF properties using a machine-learning algorithm. Then they evaluated the prediction performance using standard forecast-verification measures and compared with the prediction measures of ASAP. The comparative study revealed that the combination of SMART MFs with machine-learning produced a more accurate flare-prediction than ASAP. A Similar study using feature-
selection algorithms demonstrated that a reduced set of six MF properties can achieve a similar degree of prediction accuracy as the full set of 21 SMART MF properties.

Rough set theory has been an extensively used tool for feature selection. The only problem this is the exponential computation time. To address this problem many heuristic feature selection algorithms have been proposed in rough sets. Though these algorithms do not suffer from the problem of exponential computational time they do suffer from high computation time. To address this problem Feng Jiang et al. (2015) have proposed a novel heuristic feature selection algorithm is called feature selection model of relative decision entropy (FSMRDE) in rough sets. To measure the significance of their work they designed a new model of relative decision entropy, which is an extension of Shannon's information entropy in rough sets. FSMRDE is applied to intrusion detection and other application domains to evaluate its effectiveness. Experimental results show that FSMRDE is efficient for feature selection and is able to achieve good scalability for large datasets

Irena Koprinska et al. (2015) proposed an approach which uses the correlation and instance-based attribute selection method for electricity load predicting. This is a two-step method where the first step categorizes a set of candidate features by applying a 1-week sliding window. The second step uses feature selection techniques mutual information (MI), RreliefF, and CFS to select feature subsets. Machine learning algorithms: NN, linear regression (LR) and model tree rules (MTR). The selected feature subsets are evaluated using the classifiers linear regression, neural prediction networks and model tree
rules. Two years of Australian electricity load data were used in the evaluation process. The best result is got when NN is followed by LR and MTR.

2.3 DECISION TREE

Leyli Mohammad Khanli et al. (2011) suggested the dynamic rule learning by applying a decision tree (DT) for resource management in grid computing. The proposed method uses Grid Java based Quality of service management by Active database (Grid-JQA) system architecture. The resource manager automatically in Grid-JQA selects the set of optimal resources among the set of candidate resources using event-condition-action rules (ECA) rules specified by the experience of experts and requests resource allocation. The authors use the decision tree to update the existing rules in a dynamic environment. Using the original rule set the proposed method built a set of decision trees in parallel on training data sets. Each learned decision tree is reduced to a set of rules and the conflicting rules are resolved. Results from cross validation experiments on a data set suggest this approach may be effectively applied to rule learning.

Alternating decision tree (ADTree) is an established ensemble algorithm which brings interpretability to boosting. However, all existing ADTree algorithms are applied to univariate decision nodes where potential interactions between features are ignored. Hong Kuan et al. (2015) have introduced a sparse version of multivariate ADTree. UCI datasets, as well as spectral datasets from the University of Eastern Finland (UEF) were used for experiential study. The experimental study shows that sparse ADTree achieves the best average rank in terms of prediction accuracy, second in terms of decision tree size and faster
induction time against both univariate decision trees: original ADTree, C4.5 and classification and regression trees (CART) and multivariate decision trees: Fisher's decision tree and a single multivariate decision tree from oblique Random Forest.

Kun-Huang Chen et al. (2014) developed a novel method by combining particle swarm optimization with a decision tree as the classifier. The authors designed this approach to identify genes that could help in identifying cancers. Their comparison tabulation demonstrates that the accuracy performance of the proposed method is higher compared to benchmark classification methods: support vector machine, self-organizing map, back propagation neural network, C4.5 decision tree, Naive Bayes, CART decision tree, and artificial immune recognition system.

2.3.1 C4.5

The C4.5 decision tree (DT) is a very common method used in discovering knowledge hidden in various domains and present it in human understandable form. However, the parameters used to make decisions vary from problem to problem. Generally, a rule of thumb or trial-and-error methods are used in setting these parameters. However, these methods may produce poor results due to poor parameter settings. To address this problem Shih-Wei Lin and Shih-Chieh Chen (2012) has proposed a novel scatter search-based approach (SS + DT). To evaluate the efficiency of the proposed approach, datasets in the UCI (University of California, Irvine) Machine Learning Repository are utilized. The experimental results demonstrate that the classification accuracy of C4.5 algorithm using the proposed method in
selecting its optimal is better than the results produced by the same C4.5 algorithm but using different methods in selecting the parameter.

Yudong Zhang et al. (2014) have proposed a spam detection method. The proposed spam detection process is a five step process followed to reduce the FP error of mislabeling nonspam as spam. First wrapper based feature selection approach is used to extract the only needed relevant features. Second, the DT based classifier C4.5 is used for training. Third, values from a cost matrix is assigned as weights to two error types such as false positive and false negative. In order to bring down the sample error rate K-fold cross-validation is sourced. Finally, PSO with modified binary particle swarm optimization (MBPSO) was used as the subset search approach.

Julian Luengo and Francisco Herrera (2015) have introduced an automatic extraction model to know how well a model will perform without applying it. This method totally eliminates the need to be applied or compared with other methods. Data complexity measures are used to identify the potential domains for a classifier. These domains are a collection of relevant and irrelevant characteristics of the problems which are obtained by linking the geometrical structure that exists in the data. After analyzing the data complexity prevailing in benchmark datasets the proposed method has identified 12 potential metrics to analyze the behavior patterns of the method and to obtain intervals of data complexity.

Jao-Hong Cheng et al. (2010) have suggested three models to construct a better predictive power trading system in terms of stock market timing analysis.
The first model of the proposed three models uses a probabilistic neural network (PNN), the second model uses rough set theory, and the third model combines PNN, rough sets and C4.5 decision tree and blends them into a new model. The experimental study demonstrates that the rough set method determines group of input variables, C4.5 filter noisy attributes from the above input variables, and also uses PNN, rough sets and C4.5 classifiers to generate trading rule sets, which is helpful to construct a better predictive power trading system for stock market timing analysis.

2.3.2 Decision Stump

Zaman Md et al. (2013) have processed hybrid decision forest where each base decision tree classifiers are integrated with an additional classifier model, the boosted decision stump. In this boosting, observation weights for subsequent iterations are updated according to the binomial log-likelihood (L2) loss function. A hybrid decision forest with larger feature space is constructed with the extra samples obtained by training boosted decision stump and subsample obtained by training the base tree classifiers on the original training. The comparative study of performance of hybrid decision forest with relevant prediction methods demonstrated that hybrid decision forest is capable of yielding creditable predictive performance.

Masqueraders who imitate other users are a serious threat to computer security and very difficult to be detected using firewalls or misuse-based intrusion detection systems. The anomaly detection techniques widely used in detecting masqueraders are not good performers due to their poor accuracy and relatively high false alarm rate. In the existing methods, characteristics of user behavior were entered, and an evaluation value was calculated by the support
vector machine (SVM) and the hidden Markov model (HMM). By comparing this value with predefined threshold, it is decided whether or not the user was a masquerader. Zhou Jian et al. (2007) proposed a rule-based approach for masqueraders detection, which compares n-grams of command sequence using a technique known as boosting decision stumps using the weighted voting of the decision stumps. The weights were learned by an application of the AdaBoost algorithm. Its “decision” is made by checking the presence or absence of a specified n-gram of the command sequence.

Mohak Shah et al. (2012) explored the premise of learning a combination of decision stumps in Occam's Razor, Sample Compression, and probably approximately correct (PAC), PAC-Bayes learning settings to obtain a small subset of attributes that can be used to perform reliable classification tasks. The proposed approaches were used for gene identification from DNA microarray data. The experimental study of the proposed method showed that it not only finds a much smaller of genes but also guarantees competitive classification accuracy.

2.3.3 Naive Bayes Tree

Liangxiao Jiang et al. (2009) proposed a novel Bayes model known as hidden naive Bayes (HNB). In HNB, a hidden parent for each attribute is created by combining the influences from all other attributes. The experimental study was conducted using the 36 UCI data sets. Classification accuracy of HNB was compared with classification accuracies of naive Bayes (NB), selective Bayesian classifiers (SBC), naive Bayes tree (NBTree), tree-augmented naive Bayes (TAN), and averaged one-dependence estimators (AODE). The experimental results show that HNB significantly outperforms
NB, SBC, NB Tree, TAN, and AODE. In data mining class the class probability estimation and ranking performance were measured by conditional log likelihood (CLL) and the area under the ROC curve (AUC). Comparative study of CLL and ROC augmented naive Bayes (NB), selective Bayesian classifiers (SBC), naive Bayes tree (NB Tree) with HNB showed that HNB significantly outperforms all of them.

Naive Bayes (NB) is an algorithm known for its simplicity, efficiency, and interpretability. NB tree is proposed to weaken the Naive Bayes’s feature independence prediction. The Shasha Wang et al. (2015) proposed a novel technique called Multinomial Naive Bayes (MNBB) tree which uses a multinomial naive bayes classifier on each leaf node to build DT and classification accuracies. In multinomial NB tree classification accuracy is replaced by information gain in order to reduce the tree building time. To further scale up the classification performance of MNBTree, the authors propose a multiclass learning version called multiclass multinomial naive Bayes tree (MMNBTree) by introducing the multiclass technique to MNBTree.

Sona Taheri et al. (2014) introduced the ML and data mining techniques, NB classifier approach is a widely accepted one. Performance of NB classifier and attributes subsets requires an increase in performance in which various methods and approaches are prescribed. This work proposes a novel attribute weighted NB classifier weights to the uncertain possibilities. This is based on the configuration of the NB classifier and the feature weights; the objective function is demonstrated and considered. The suggested method NB classifier reserved with a starting point. The optimal weights are persevering by a local optimization technique by applying the quasi-secant technique. Real-time
datasets in binary classification are measured; the experimental results showed very efficient and best precision for a presented method.

2.3.4 Random Forest

Adnan Idris et al. (2012) proposed a churn prediction model for the telecom industry. The proposed model is based on PSO, Minimum Redundancy and Maximum Relevance and RF. The authors conducted a detailed study to understand to measure how efficiently PSO based under the sampling method in collaboration with feature reduction techniques handles the imbalance data distribution. Principle Component Analysis (PCA), Fisher’s ratio, F-score and Minimum Redundancy and Maximum Relevance (mRMR) are the feature reduction techniques work in the hands with PSO. Two classifiers namely, RF and KNN are applied to the reduced feature dataset in order to analyze the classification accuracies. By using sensitivity, specificity and area under the curve based measures, the prediction accuracy is estimated. The proposed Chr-PmRF approach effectively handles the problems of customer churn prediction in telecommunication.

Miron Bartosz Kursa (2014) compared the relevant approaches Artificial Contrasts with Ensembles (RF-ACE) and Boruta methods and the minimal optimal approaches the Recursive Feature Elimination (RFE) and Regularized Random Forest (RRF). To evaluate the characteristics of various gene selection methods, tests were performed on four standard pre-processed microarray datasets: Colon, Leukemia, SRBCT and Prostate. Experimental study shows that the Boruta algorithm using the Random Forest important measure is the slowest whose computational training time ranged from hours to days,
especially for larger sets. Whereas the RF-ACE and RRF algorithms trained colon, leukemia and SRBCT datasets in 1 hour and prostate dataset in 2.5 hours. RFE and RRF methods selected a much smaller subset of genes than the all relevant methods.

Xiaoyong Pan et al. (2014) proposed a novel approach for binding site identification in proteins. The proposed method extracted local feature evolution information and global feature molecule weight from protein sequence. The extracted global features differentiate the non-interactive and interaction sites. A sub-sampling strategy based on submodularity framework was proposed by the authors to select representative data subset from redundant and unbalanced dataset. Finally, random forest was trained on optimally selected training subsets to predict interaction sites.

The scale invariant feature transform (SIFT) method detects and describes local features of interest points in images. Moreover, it is invariant to image scale, rotation, illumination, noise, and occlusion. SIFT leads to a very large number of dimensions. Moreover, the classification of fingerprints leads a number of classes which corresponds to the number of individuals. To handle the very-high-dimensional points Thanh-Nghi Do et al. (2015) proposed multi-class random oblique decision trees. The proposed method induces a forest of binary oblique decision trees. This approach separate the C classes at each non-terminal node into two subsets of classes of size c1 and c2 (c1 + c2 = C). The experimental results showed that the multi-class random forest of oblique decision trees (RF-ODT) algorithm is very efficient than C4.5, random forest RF-CART, AdaBoost of C4.5, support vector machine and k nearest neighbors.
2.4 ENSEMBLE CLASSIFIER

Jiang Yu et al. (2015) proposed a spatial steganalytic scheme with redistributed residuals and a diverse ensemble classifier. In order to redistribute the residuals of local pixel predictors the proposed method combines shifting and suppressing techniques. The first-order statistic is applied to the resulting residuals reduces the computational complexity and at the same time preserves the long-range dependencies between pixels. Bagging and AdaBoost mechanisms help in obtaining diverse representation of final feature and enhance the individual weak classifier to improve the detection accuracy.

Subrajeet et al. (2014) proposed an ensemble classifier structure for initial analysis of critical lymphoblastic leukemia in blood microscopic images with the objective to increase the acute lymphoblastic leukemia (ALL) diagnostic precision. As the time taken by visualization examination is too long and the diagnosis is less accurate, the authors coined a quantitative microscopic approach. The objective of this work is to improve the ALL diagnostic accuracy by analyzing morphological and textural features from the blood image using image processing. The automated lymphoblast recognition method proposed by the authors uses image segmentation, feature extraction and classification to differentiate lymphoblasts (malignant) from lymphocytes (normal) in stained blood smear and bone marrow samples and to assist in the development of a computer-aided screening of ALL.

Bin Chen et al. (2014) suggested a fraternization high-dimensional attributes for JPEG steganalysis with ensemble classifier. The proposed method uses two subspaces for training the base learners of ensemble classifiers.
Firstly, the 1,034-dimensional PS feature space is generated by mixing SHI and extended DCT features. Next the second subspace is constructed by combining the high-dimensional feature spaces SHI, PEV and CF. Since the CF intra-block features are more sensitive intra-block features are added to the new feature space. The dimensional of the new subspace is reduced according to the correlation coefficient among different features parts to form the PSC feature space. Finally, the authors propose a proportion mechanism to manage the feature selection in two subspaces for training the base learners of ensemble classifier. Experimental results show that the PSC feature performance outperforms that of existing steganalysis approaches to obtain about 2% gain.

2.5 ADABOOST

The cascade AdaBoost approach based on Haar is the best technique for face detection. Zhong Li et al. (2015) have proposed a new cascade AdaBoost face detection approach based on multi-threshold weak classifiers, called multi-threshold AdaBoost (MTAdaBoost). The proposed approach chooses multiple thresholds, by computing the solutions of two optimal. The experimental results shows that the proposed method is an effective face detector with much less features, is very faster and takes less time for training.

Junyoung and Jin et al. (2014) proposed an AdaBoost based model in predicting bankruptcy in Korean construction organizations. Based on the investment organizations are divided into three types; small, middle, and large. The authors analyzed the predictive ability of the ANN, SVM, Decision tree Z-Score and AdaBoost for each group of companies. From the experimental results they have concluded that machine learning algorithms display better
predictive capability than Altman’s Z-score. Moreover the detailed study proves that AdaBoost is the best bankruptcy forecasting model.

Annalisa Riccardi et al. (2014) have proposed an additive modeling using a multiclass exponential (SAMME) boosting algorithm. This proposed approach uses a cost-sensitive approach to maintain a natural order in the targets. The proposed ensemble model uses Extreme Learning Machine (ELM) model with the Gaussian kernel and an additional regularization parameter as a base classifier. By using ELM as base classifier the necessity of generating a new training dataset at each iteration is avoided. The adoption of the weighted least squares formulation in estimating analytically the parameters connecting the hidden layer to the output layer at each iteration of the boosting algorithm has been introduced as an alternative approach to the already existing ELM boosting techniques. Three new algorithms, AdaBoost ELM for nominal classification and AdaBoost for ordinal regression based on ELM and cost model $i$ is denoted as (AdaBoost (ELM).ORC1), AdaBoost (ELM).ORC2 and AdaBoost (ELM).ORC3 were proposed to handle ordinal regression problems. Using ordinal regression datasets and a synthetic dataset (the toy dataset) as benchmark datasets experiments were conducted with A simple approach to ordinal regression (ASAOR), MultiClass ordinal support vector machines (MCOSvm), ORBoost with all margins (ORBoost-All), ORBoost with left-right margins (ORBoost-LR) and one of the proposed Extreme learning machine for ordinal regression (ELMOR). On comparing the results it is found that AdaBoost (ELM).ORC3 algorithm is the best among the proposed methods with the most effective cost model.
Ammar et al. (2011) have proposed a new approach in which particle swarm optimization (PSO) is used within an AdaBoost framework for object detection. AdaBoost performs an exhaustive search to find out good features to be used for constructing weak classifiers. To avoid the extensive search authors have proposed two methods based on PSO. First, they use PSO to evolve and select good features only and the simple decision stump to select weak classifiers. Next selection of good features and evolving weak classifiers are handled in parallel using PSO. These two methods are examined and compared on two challenging object detection tasks in images: detection of individual pasta pieces and detection of a face. The experimental results demonstrated that using PSO for selecting good individual features and evolving associated weak classifiers in AdaBoost is more effective for selecting features.

Ebenezer et al. (2014) introduced the neural-AdaBoost based facial recognition model. This model improves the accuracy and computational execution time of facial recognition systems. From thousands of facial features different facial patterns were extracted with Gabor feature extraction technique. Due to the large size of the Gabor wavelets, it is not practically possible to use all the wavelets as input to the classifier, for fear of misclassification and possible system crash. To speed up the classification and to select a few hundreds of the numerous extracted features an AdaBoost algorithm is applied. The selected features were fed into an elegant three-level NN classifier that is trained by a back propagation procedure. Japanese Female Facial Expression (JAFFE) dataset is used in evaluating the proposed approach. The Support vector machine, Image feature, 2D Discriminant locality, PCA and neural network, Regularized Discriminant Analysis (RDA) were compared with the results of the proposed approach. The comparative study shows that the performance of the proposed approach is outstanding.
Cao Ying et al. (2013) discussed the advance and prospects of AdaBoost algorithm. AdaBoost is one of the best outstanding boosting techniques. The authors have described unique features of AdaBoost. First, they claim that the AdaBoost by changing the boosting which was initially a mere conjecture into reality is one of the reasons for its success. They explain how the change in distribution of training error and generalization error has improved the precision in weak learning algorithm and paved way for designs of many other learning algorithms with training and simplification error. Their study brings to light the various hypothetical simulations of AdaBoost and the several alternatives derived from them. Fourth discussion is about how the binary class AdaBoost is also suitable for solving multi classification problem. The author’s discussion ends after identifying the application areas of AdaBoost and the possible research issues for the feature.

Jelte and Gerard (2015) have compared ML methods for real-time detection. AdaBoost and SVM are two best techniques applied to the different ML techniques. The framework identified that AdaBoost is the exceptional ML method for real-time or target detection when compared with SVM. Moreover, it is straightforward to use and training time is very low. Gisette and Dexter are the two datasets taken for experimental study where noise is introduced in the attribute set. The outcomes of the various ML methods are compared using the datasets with different classifiers namely AdaBoost, Radial SVM, C4.5, Linear SVM, 2nd Polynomial SVM, and 3rd Polynomial SVM.

Putthiporn and Chidchanok (2013) have proposed a method which eliminates imbalanced error domination effect and improves the testing accuracy. The authors observed that the sum squared error (SSE) function used
in the training processing of a supervised NN is the cause of classification error in majority class domination over the classification error in minority class. To equalize the number of the error terms in both classes the authors have used the boundary data of each class. To locate all relevant boundary data the authors adapted the concept of Hausdorff distance. Secondly, they observed that if the training data are not evenly distributed, then the process of the testing accuracy may be low. However, it is almost impossible to assume that the training data are evenly distributed. Therefore the authors decided to free the testing accuracy from the distribution of training and testing data by taking a natural standard deviation of the data, including both training and testing data. To estimate the natural standard deviation of the data, an efficient statistical technique called Bootstrapping is deployed prior to the training process. The experimental results demonstrated the superiority of the proposed method over the other methods. Moreover, the proposed method significantly achieved promising overall accuracy, F-measures of both classes, G-mean and average area under the curve (AUC). The proposed method achieved higher accuracy in both minority and majority classes than the other techniques.

Shuqiong and Hiroshi (2014) proposed a parameter to fast up the training of real AdaBoost using parameterized AdaBoost. The machine learning algorithm AdaBoost has achieved significant success in object detection and data classification. Modified Real AdaBoost algorithm decreases weights of correctly classified samples and increases the weights of misclassified samples but at the same time it may lead to misclassification of samples which were correctly classified in previous runs. A quicker training can be accomplished when this kind of misclassification is restricted during the boosting method. The authors have suggested a parameterized AdaBoost method to address this problem. The proposed method introduces a parameter to penalize the
misclassification of samples that have already been correctly classified. By penalizing the misclassification the proposed method is able to retain more positive margins compared to Real AdaBoost. Experimental results on eighteen datasets by using convergence cost and test error is compared with real AdaBoost and parameterized AdaBoost. Experimental results show that the training error in the proposed method converges faster.


Shuang Liu et al. (2014) proposed an efficient enhancement procedure based rainfall-runoff framework by applying an improved AdaBoost. RT (Adaptive Boosting for Regression problems and “T” for a threshold demarcating the correct from the incorrect) method for flood disaster warning. Particularly for flood periods and dry periods a single procedure based rainfall-runoff framework at sometimes fail to capture all the runoff characteristics. An operative multi-model ensemble method is essential to address this problem. The AdaBoost.RT technique of ensemble learning methods gave the idea for
designing of aggregation (AdaBoost-XXT) of a process-based rainfall–runoff model (AdaBoost-XXT) of a procedure based rainfall-runoff named XXT which is a fusion of TOPMODEL and Xinjiang model. Initially, the “hard” samples might be emphasized due to the weights of incorrect predicted samples remained improved somewhat than steady. The static threshold usually used to distinguish the correct from the incorrect was substituted with the active mean value of absolute errors. The PSO was used to define the model parameters. The AdaBoost-XXT was evaluated using datasets Linyi watershed. The experiential data from 2001 to 2003 and daily data from 2006 to 2007 were referred as calibration and validation sets. Moreover, youshuijie catchment, pan evaporation, the experiential daily rainfall and streamflow data from 1981 to 1983 and daily data from 1984 to 1985 were used as the validation set.

Hence this chapter is provided with an elaborated literature review covering all the possible dimensions used in this research. In the next chapter the detailed methodology of the proposed method is discussed.