E-mail Spam Classification using PCA-SVM and PCA- Hybrid of SVM RBF with Adaboost

A Dissertation Report

Submitted By

Sneha Singh

(11307124)

Submitted to

Department of Computer Science & Engineering

In partial fulfillment of the requirements for the award of the Degree of

Master of Technology in Computer Science

Under the guidance of

Sandeep Kaur (Asst. Professor)

(May, 2015)

ABSTRACT

E-mail enables users to send and receive messages over internet in a very fast and economical way. Although Email is a good source of information exchange some people send unsolicited bulk messages termed as spam to numerous recipients. Spam concept is diverse as spam may contain unwanted advertisements, chain letters etc. Spam causes wastage of resources and it is very annoying problem which is being faced by almost everyone having an email account. So filtering of spam email before sending it to the inbox of users is very important and challenging task. In this work, we have taken e-mail dataset from UCI spambase corpus. We have implemented and evaluated two models for e-mail spam classification, i. e. PCA-SVM and PCA- Hybrid of SVM RBF with Adaboost. PCA has been used for reducing data dimensionality. We have evaluated different type of existing classifiers without PCA for e-mail spam classification. Finally, comparative analysis of existing and proposed models has been done and the better approach for e-mail spam classification has been identified.

CERTIFICATE

This is to certify that **Ms Sneha Singh** has completed M.Tech Dissertation titled **"E-mail Spam Classification using PCA-SVM and PCA- Hybrid of SVM RBF with Adaboost"** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engg.

Date:

Signature of Advisor
Name: Ms Sandeep Kaur

UID:

ACKNOWLEDGEMENT

I would like to present my deepest gratitude to **Ms. Sandeep Kaur,** Assistant Professor (Department of Computer Science) for her guidance, advice, understanding and supervision throughout the development of this Dissertation study. I would like to thank to the **Project Approval Committee members** for their valuable comments and discussions. I would also like to thank to **Lovely Professional University** for the support on academic studies and letting me involve in this study.

DECLARATION

I hereby declare that the Dissertation proposal entitled **"E-mail Spam Classification using PCA-SVM and PCA- Hybrid of SVM RBF with Adaboost"** submitted for the M.Tech. Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____ **Investigator:** Sneha Singh

Registration No. 11307124

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Data mining is the process in which study and investigation of large quantities of data is performed in order to discover suitable, novel, potentially useful and ultimately understandable patterns in data. In the process of data mining, knowledge is fetched from massive amount of data and suitable patterns are generated. Generated pattern defines different types of relationships in data and represent it to user from many dimensions. In the field of Information technology, there are enormous amount of data available that can be used for various applications like market behavior analysis, fraud detection, customer retention, production control, science investigation etc. Data mining can be applied on different sources such as organization's data warehouse, on web pages etc.

Today World Wide Web is the most common place of information storage and retrieval. Web page contains various kinds of noises and harmful contents. These harmful contents must be addressed in order to be safe from severe problems.

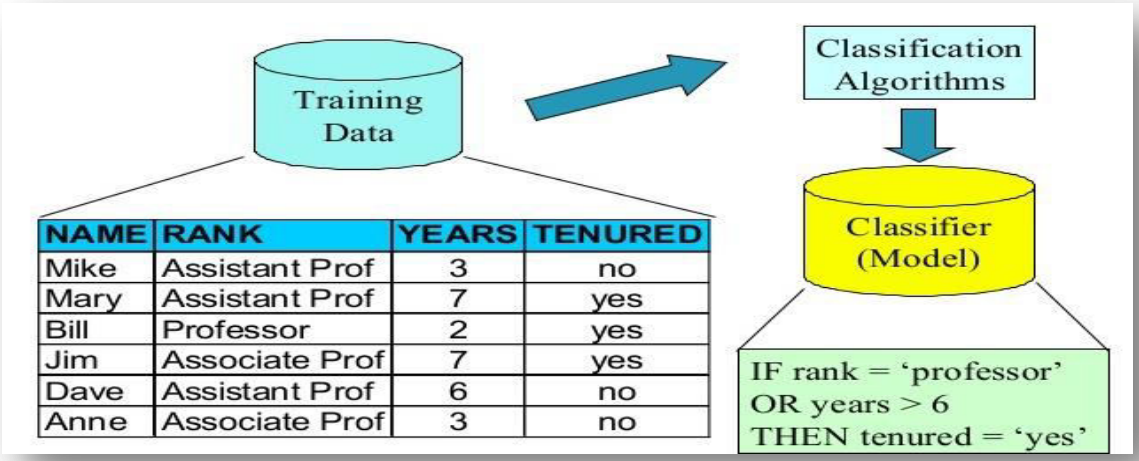
E-mail is a very important solution provided over internet. People can communicate with others residing at far geographical locations in a very less amount of time. Although e-mail is a very economical and fast method of communication, dirty e-mail or spam email may lead to serious and annoying problems. Illegitimate e-mail may cause malware downloads, unwanted advertisements etc. Various techniques are available to filter spam e-mail and to be safe from its negative consequences. Server side filters are used to detect spam e-mails before sending it to the inbox of recipients. This helps in keeping inbox of user clean and protected.

Machine learning is an important and broad field under data mining. In machine learning, various specialized algorithms are learned to make decisions and do predictions. These algorithms can be used to assign an e-mail either spam or ham class label.

1.1. DATA CLASSIFICATION- In machine learning, classification is a process of identifying the class label to which new [1] observations belong based on training data whose class labels are known. Input test data or unseen data is analyzed and assigned a class label by algorithm, which is termed as classifier. An example of classification is assigning a given email, either spam or legitimate class label.

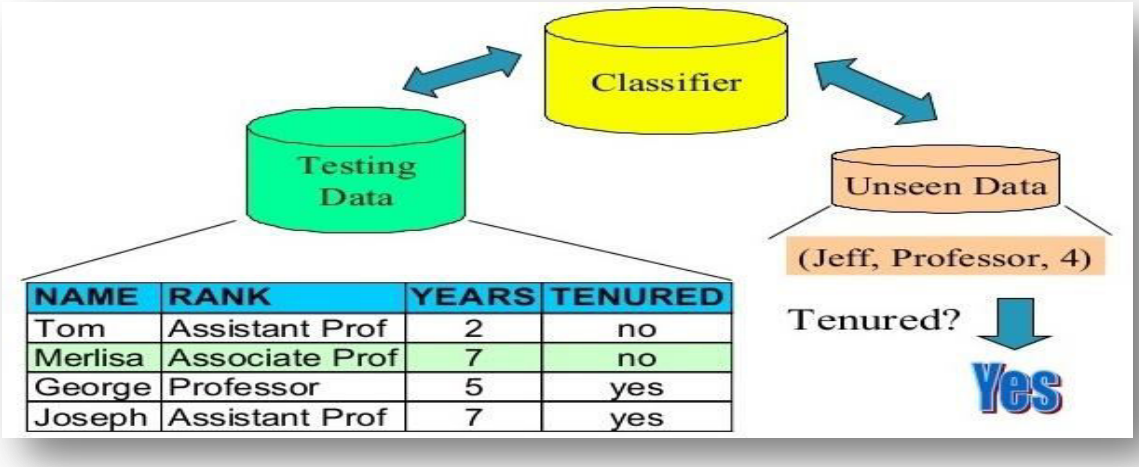
Data classification is a two [1] step process, i.e. learning step and classification step. In learning step, training data is provided to algorithm which builds a classification model. In classification step, generated model is used to predict class label for testing data.

Example: Two steps of classification have been represented in following diagrams. Classifier is provided some training data and model is built which in turn used to classify test data or data whose class labels are not known.



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Fig 1.1: Data Classification (Learning Phase)



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Fig 1.2: Data Classification (Testing Phase)

There are many application areas, where classifiers are used. Some of which have been mentioned below.

- Computer vision

- Speech recognition
- Pattern recognition
- Biometric identification
- Biological classification
- Document classification
- Handwriting recognition

1.1.1. CLASSIFICATION ALGORITHMS: Different types of algorithms are used for data classification purpose. Some of the algorithms which have been used extensively for e-mail spam classification have been listed and discussed below.

- Naïve Bayes
- Support vector machines
- Decision trees
- Meta Classifiers (Boosting)

1.1.1.1. NAÏVE BAYES CLASSIFIER: Naïve bayes classifier is a simple bayesian classifier. Bayesian classifiers are used to [22] predict the class label probabilities of input tuples. Bayesian classifiers are based on bayes' theorem.

Let D [22] is a set of training tuples with associated class labels. Suppose there are n classes, $A_1, A_2, \ldots, A_n$. Given a tuple X, the classifier will predict the class to which X belongs. Classifier predicts that tuple X belongs to class A_i if,

$$P(A_i|X) > P(A_j|X) \quad \text{for } 1 \leq j \leq n, i \neq j.$$

By Bayes theorem,

$$P(A_i|X) = \frac{P(X|Ai)\, P(Ai)}{P(X)}$$

Where P(X) and $P(A_i)$ are the prior probabilities of X and A_i respectively. $P(X|A_i)$ [22] is the posterior probability of X which is based on A_i. P (Ai|X) is posterior probability of Ai conditioned on X.

1.1.1.2. SUPPORT VECTOR MACHINES: Support Vector Machine (SVM) is used for classification and regression analysis. When training dataset is provided, SVM training algorithm builds a model which is used for data classification. SVM classifies

data in one of the two classes. SVM [23] constructs a set of hyperplanes in a high dimensional space, which are used for data classification or regression task. Very large number of separation lines termed as hyperplanes can be generated by SVM. A hyperplane [23] with maximum margin is selected and such a hyperplane is termed as Maximum Marginal Hyperplane (MMH). MMH provides best classification results. Following figure represents a set of hyperplanes.

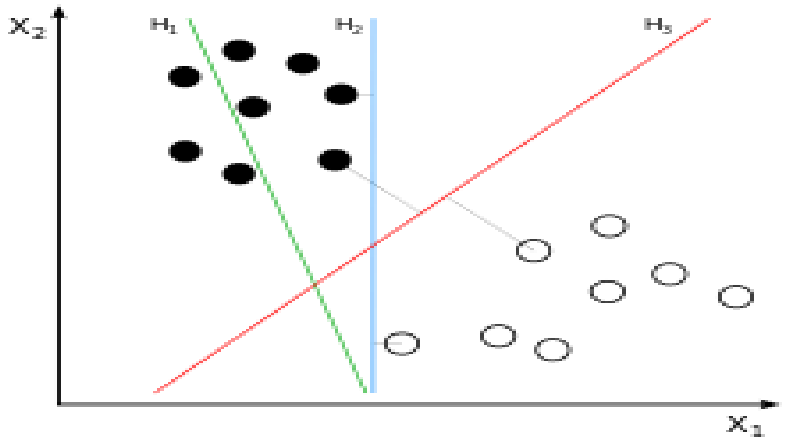


Fig 1.3: SVM Hyperplanes

Linear SVM: In linear SVM, from a set of hyperplanes a hyperplane with maximum margin is selected and used for data classification purpose. Maximum Marginal Hyperplane (MMH), classify data items in one of the two classes. Margin can be defined as, shortest distance from the hyperplane to the closest [23] training tuples on either side. Training tuples closest to the MMH are termed as support vectors. Support vectors are equally close to the MMH from both sides (classes).

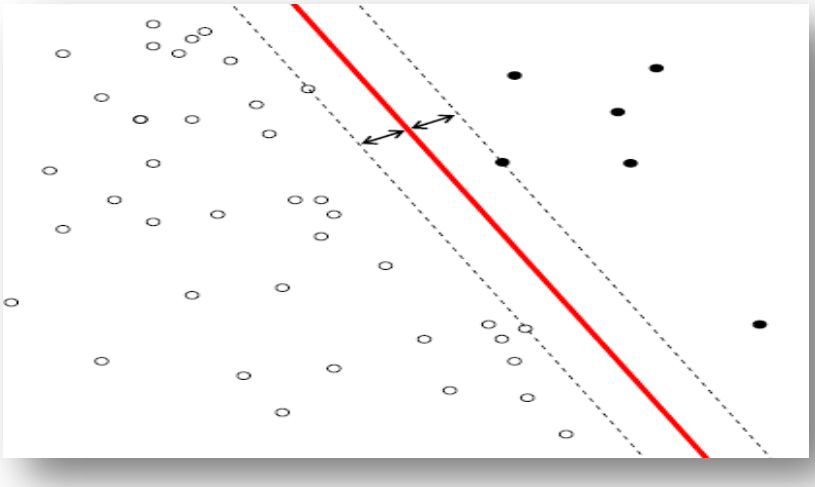


Fig 1.4: Linear Classification

In fig 1.4, circles at dotted lines are representing support vectors and dark line is MMH.

Non linear SVM: Data which is not separable linearly, non linear classification is used by extending linear SVM approach. Various kernel tricks are used with SVM to do non linear classification. Linear hyperplane is replaced by a non linear separation.

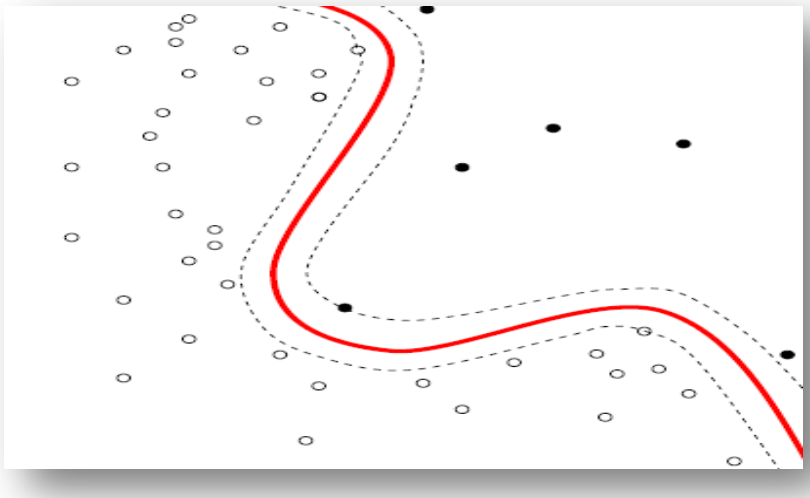


Fig 1.5: Nonlinear Classification

1.1.1.3. DECISION TREE CLASSIFIER: Decision tree is a tree like structure in which internal nodes represent test case and branch represents outcome of [25] test. Leaf nodes are used to define class labels. Decision trees can handle both numerical and categorical data. Decision tree may be binary or non binary. Suppose a tuple is given for which class label is not known, a path is followed from root to terminal node in order to determine class label of input tuple known,a path is followed from root to terminal node in order to determine class label of input tuple known,a path is followed from root to terminal node in order to determine class label of input tuple.

Decision tree induction is the process of learning decision trees. Decision trees are learned from labeled input tuples. Various algorithms of decision tree are [25] ID3 (Iterative Dichotomiser), C4.5 (successor of ID3), CART(Classification and Regression tree) etc.

Following is a simple decision tree, which is used to take decision whether a customer will buy computer or not. Test cases are given on internal nodes and class labels are given on leaf nodes.

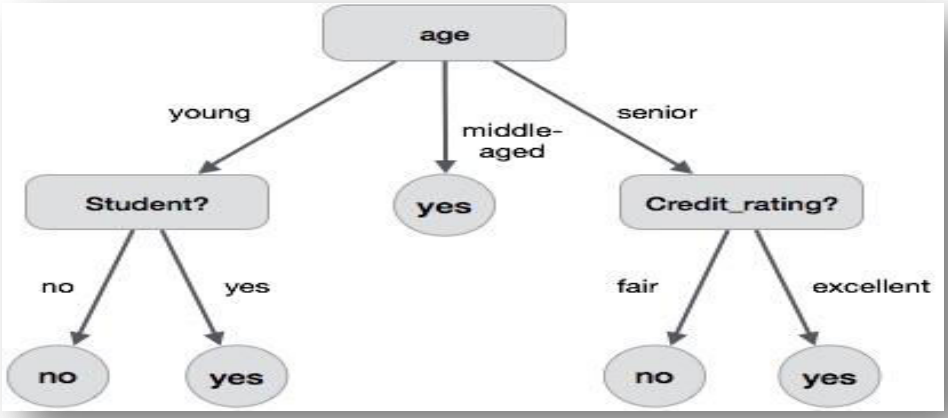


Fig 1.6: Decision Tree

1.1.2. TECHNIQUES TO IMPROVE CLASSIFICATION ACCURACY- There are some tricks which can be used to increase the accuracy of classifiers. Techniques have been listed below.

1.1.2.1. ENSEMBLE METHOD: Ensemble methods are used to improve the classifier performance. An ensemble is a combination of multiple classifiers. Ensemble generates more accurate results than its individual classifier. Various well known ensemble classifiers are, bagging, boosting and random forests.

The ensemble makes predictions by considering votes of its base classifiers. Following is diagrammatical representation of ensemble classifier. Here m number of classifiers is being trained with some sample of training data. A new composite classifier is created and is being used to do classification when test set is provided to it.

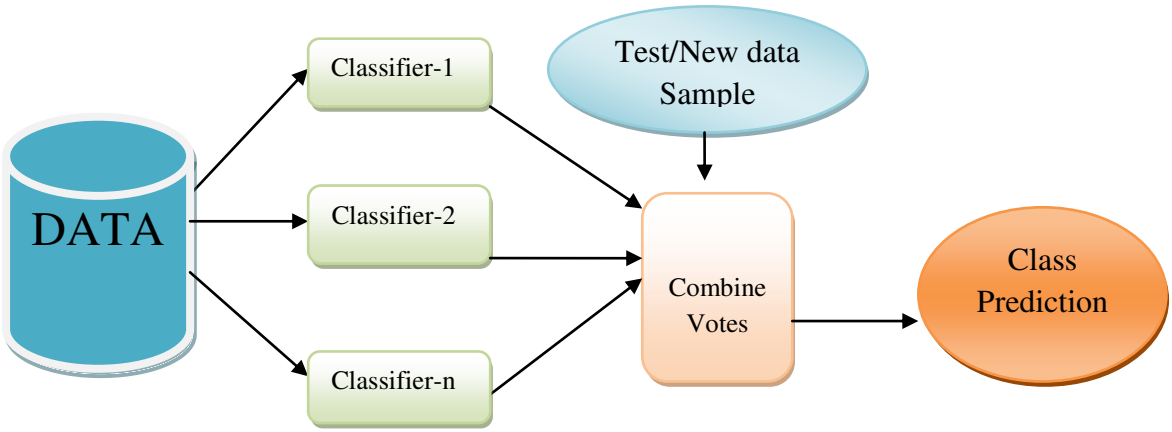


Fig 1.6: Ensemble Classifier

BAGGING- Bagging is also termed as bootstrap aggregation. It is a meta algorithm and used to enhance the accuracy of various base classifiers.

Suppose a set of training [1] data D having n tuples have been provided, bagging generate a new set of training data D_i having n tuples for each iteration i (i=1,2,…k). Di is created from original training data by using method of sampling with replacement [20]. Newly created training set may not contain some of the tuples present in D and some duplicate tuples may also be there. A series of classifiers are learned using training set Di. Each base classifier is learned using different training data set. Each base classifier is assigned one vote and final decision is made on the basis of total votes.

BOOSTING- Boosting meta classifiers are used to increase the performance of weak learners. In boosting, each training tuple is assigned some weight. N numbers of classifiers are learned iteratively [1] and weights are assigned to each classifier. Final classifier combines the weighted vote of each classifier and makes prediction.

Aaptive Boost (Adaboost) is a popular boosting algorithm. Suppose [1] dataset D has k class labeled tuples, (X_1,Y_1), (X_2,Y_2), ……(X_k,Y_k), where X_j represents tuple and Y_j represents its class label. Initially Algorithm goes through N rounds to generate N classifiers. Initially each tuple is assigned some weight, which is equal for all tuples. Weights of training tuples are updated based on how they are classified. A tuple's weight is increased if it was misclassified and decreased if it was correctly classified. Weights of classifier M_i is used to decide the training samples for classifier $M_{i+1.}$

Error rate of model M_i can [20] be given by, sum of weights of each tuples of N that has been misclassified by M_i

$$\text{Error of Mi} = \sum_{k=0}^{N} \text{wj} * \text{err(Xj)}$$

Where err(Xj) is misclassification error of tuple Xj. Err(Xj) value is 1 if tuple was misclassified and it is 0 if the tuple was correctly classified.

Each base classifier's vote is assigned some weight based on its performance and weighted votes are used to do final decision making.

1.1.3. METRICS FOR EVALUATING CLASSIFIER PERFORMANCE- Various measures [1] are used to check the accuracy of classifier in prediction of class labels for given tuples. These measures have been discussed below,

Accuracy = (TP + TN)/(P+N)

Precision = (TP)/(TP+FP)

Recall = (TP)/(TP+FN)

F-score = (2*Precision*Recall)/(Precision + recall)

Positive (P): This refers to the number of positive tuples.

Negative (N): This refers to the number of negative tuples.

True Positive (TP): This refers to positive tuples that were correctly labeled by classifier.

True Negative (TN): This refers to negative tuples that are correctly labeled by classifier.

False Positive (FP): This refers to negative tuples that are incorrectly labeled as positive by classifier.

False Negative (FN): This refers to positive tuples that are incorrectly labeled as negative by classifier.

Confusion Matrix: Confusion matrix is a tool to analyze the performance of a classifier. TP and TN show that classifier is performing well where FP and FN show that classifier's performance is getting worse.

		Predicted	
		Positives	Negatives
Actual	Positives	TP	FN
	Negatives	FP	TN

Fig 1.7: Confusion Matrix

1.2. SPAM: Spam is unsolicited and annoying message sent by spammers to numerous recipients. Spam may be unwanted advertisements or any phishing messages. Spam sometimes has malware attachments which are vulnerable to system resources. Spam is very annoying problem faced by any individual user of e-mail. Spam appears in different media such as e-mail, messaging system, social networks etc.

1.2.1. E-MAIL SPAM: Spam email is the junk mail or unsolicited commercial message containing advertisements and other irrelevant content. Spam emails are termed as unsolicited bulk email as it is sent in very large quantities by spammers. Spam email is an umbrella term under which various categories of e-mail come such as unwanted advertisements, asking users for confidential data, chain letters etc.

1.2.1.1. APPROACHES TO PROTECT AGAINST SPAM EMAIL

Various methods are used to protect from spam e-mail or phishing e-mail. These methods can be used to be away from various disadvantages of spam e-mail. Following list of methods are used to be safe from spam e-mails, phishing e-mail and e-mails having malware attachments.

1. **Network Level Protection:** In this approach, a set of IP addresses are blocked by network administrator from sending messages. Concept of domain name blacklisting is used to protect from spam emails or phishing emails.
2. **Authentication:** Authentication based [14] approaches are used to make sure whether email has been sent by a trusted party or not. Two type of authentication are domain level and user level. Domain level authentication is implemented by e-mail provider, Microsoft has used concept of sender id to identify phishing emails. User level authentication can be done by verifying digital signatures of sending party. Digital signature is signed by organization's private key.
3. **Server side classifiers and filtering tools:** These techniques work on a set of spam email features. A model is built from training [14] tuples and classification is done by generated model. Model is able to classify unseen and new emails, either as spam or ham. Various classifiers such as support vector machines, decision trees, naïve bayes, boosting and many more can be used to classify e-mails. Multitier classifier and hybrid classifier approach is also used to classify spam emails before sending it to the account of individual users. These server side tools do not always perform accurate classification.
4. **Client side tools:** Various browser based toolbars are used to classify emails in one of the two categories, i.e. legitimate or malicious. Blacklist or whitelist are maintained by browsers to make the user safe from illegitimate emails.

CHAPTER 2

REVIEW OF LITERATURE

In this chapter some of the existing and related research works done in the field of data classification and e-mail spam detection have been described. In this section, numerous studies have been reviewed and some of the papers have been considered that has been taken as motivation towards this study.

Serkan Günal et al. **(2006),** Authors have proposed two methods for selection of important features for efficient e-mail classification. Reduced dimensions result in less computational costs. Common Vector [3] Approach has been used for feature reduction. E-mails containing 140 features have been used from which 88 and 86 features have been selected to take part in classification process. Selected features have resulted is better selection performance of classifier. Dataset has been taken from SpamAssassin corpora. Training set contains 2000 and test set contains 750 e-mails. In future other classification algorithms can be used with feature selection methods.

Alireza Saberi et al. (2007), Authors have used three different learning methods and one ensemble method to detect phishing emails. Three data mining [4] algorithms have been used to detect phish email (scam) namely, K nearest neighbor, Poisson probabilistic theory and Bayesian probabilistic theory. These three text classification algorithms have been explained in this study. Spam and ham email dataset has been taken from Enron-spam whereas scam samples have been taken from a web phishing repository. Algorithms have been used to categorize data in two parts, i.e. frauds (phishing email) and non frauds (ham and spam email). Then by using majority voting ensemble classification algorithm have been used, in which their results were merged in order to increase the accuracy of classification.

Ma et al. (2009), In this paper authors have paid their attention on finding malicious URLs based on two types of features, i. e. host based features and lexical features. They have told the reason of not considering other type of features for URL detection. They have considered features derived by other researchers in this area. They have used different classifiers named as support vector machine, naïve bayes and logistic regression. Seventeen categories of features have been selected by them to do their work. They have evaluated the results obtained by different classifiers. SVM and LR classifiers are better than naïve Bayesian when more number of features is being used. They have also

evaluated the error rate of classifiers when training and testing data are taken from different sources. They explained false positive and false negative rates along with reasons for them in their experiment. Finally they explained the significance of machine learning algorithms over non machine learning approaches and they are willing to work on online learning algorithms.

Hsu Wei-Chih et al. (2009), Authors have done spam filtering using SVM based on textual content based features. Taguchi method, an approach of evaluation, implementation and design of process products [6] has been used. Goal of the method is to improve the quality of product by eliminating errors rather than causes of errors. Two important tools used in Taguchi Method are signal [6] to noise ratio and orthogonal vectors. Vector space model have been used for feature representation. Dataset has been taken from SpamAssassin corpus. Authors have changed two parameters of SVM for achieving better results. Orthogonal table has been used for selecting parameters and large table provides better selection of parameter. Proposed [6] approach has performed well when parameters are, $log2(C) = 8$ and $log2(\gamma) = -14$ respectively [pep]. Parameters must be selected carefully in order to get good performance.

Wenjia Wang et al. (2010), Authors have used ensemble of Bayesian classifier for e-mail spam classification. Both heterogeneous and homogeneous ensemble methods have been used for spam classification. Two kinds of [7] heterogeneous ensembles have been used in which, Naïve Bayes (NB) with Decision Tree (DT) and Bayesian Network (BN) with Decision Tree (DT) have been used as base classifiers. Three homogeneous ensembles of NB, BN and DT have been used. Ten datasets have used for validating classifier performance. Method for building heterogeneous ensemble has been explained in paper. Dataset has been taken from UCI spambase corpus. Heterogeneous ensemble models have given better classification results than base classifiers and homogeneous ensembles.

Justin ma et al. (2011), Authors have explained host based and lexical features of malicious URLs and given techniques to find them. Detailed definition and works of URL have been explained. In this need for dynamic feature classifier have been specified. Explained the problems associated with static feature selectors and some of the online learning algorithms have been considered for evaluation purpose. Comparison of online algorithms along with their advantages has been given. They collected various features of URL and performed classification using online and batch algorithms in MATLAB.

Methods used for feature collection and feature representation are also given. They have considered live and huge sources of labeled URLs for implementation purpose. Their work is important as they have used more comprehensive set of features and used online algorithms for implementation.

Monther Aldwairi et al. (2012), Authors have proposed a lightweight system to detect malicious websites online based on URL's lexical and host features and call it MALURLs [9]. They defined a malicious web page as a page that downloads a file, uploads a file, collects data, installs an application, opens a pop window(s), and displays an advertisement or any combination of the above without the knowledge or consent of the user. [9]. They have used naïve bayes classifier along with genetic algorithm to do the work. Initially they took small data set and increased it at a high speed using Genetic Algorithm. They trained naïve bayes classifier and used completely different dataset for testing purpose. Data sources chosen for the work are AKLEXA and Phishtank. They discussed about the significance of using naïve bayes classifier and Genetic Algorithm in their work. They used PHP as front end MySQL as backend for implementation purpose. They measured the precision of their work and discussed importance of using other parameters such as Genetic Algorithm.

Juan Carlos Gomez et al. (2012), Work is focused on e mail classification using text content features only. Classifier uses principal component analysis document reconstruction (PCADR). It has been shown that PCDAR is able to extract and synthesize the important features of document for efficiently [10] representing any class. PCADR approach has been tested on different e mail corpora such as PU1, Ling Spam, SpamAssassin, Phishing and TREC7 [10] spam corpus. Every experiment is compared with linear SVM in order to evaluate performance of PCDAR. All the experiments were performed using a Core i7 1.7Ghz PC with 4GB in RAM using Windows and Java. PCDAR proved to be better than SVM in terms of classification accuracy and classification time (when classifying test examples). PCADR is well suited when training and testing data are from different sources. In future PCDAR can be applied to other text classification tasks and it can be also used with other classifiers to produce weighted decision about class.

R. Kishore Kumar et al. (2012), In this paper, authors have analyzed various machine learning spam classification algorithms. E-mail spam [11] dataset has been taken from

UCI machine learning repository and TANAGRA data mining tool has been used to analyze existing algorithms. Different feature selection algorithms namely Fisher filtering, ReliefF, Runs Filtering and Step disc has been used to select appropriate features from dataset. Various spam classification algorithms have been applied on the data set before and after feature selection and results are compared. The Runs tree classification is considered as a best classifier, as it produced 99% accuracy.

Venkatesh Ramanathan et al. (2012), Authors have proposed 12] a new server side methodology to detect phishing attacks namely phishGILLNET. PhishGILLNET consists of multiple layers in which the first layer (phishGILLNET1) makes use of Probabilistic Latent Semantic Analysis (PLSA) [12] to build a topic model. The second [12] layer of phishGILLNET (phishGILLNET2) uses AdaBoost to build a classifier in which probability distributions of the best PLSA topics have been used as features. The third layer (phishGILLNET3) makes a classifier from labeled and unlabeled examples by employing Co-Training. For experiment four email dataset and one phish URL dataset have been used to evaluate the performance of phishGILLNET. Ham email dataset has been taken from SpamAssassin corpus and Enron Email Dataset whereas Spam email dataset has been taken from PhishingCorpus and SPAM archive. Phish URL dataset has been taken from Phishtank. PhishGILLNET1 was compared with SVM, where phishGILLNET1 performed better. phishGILLNET2 [12] supports both 3-class (phish, spam, good) and binary (phish, not-phish) classification. phishGILLNET3 can handle unlabeled data. Performance of phishGILLNET has been compared with ten state of art methods and phishGILLNET found to be best classifier among all other classifiers. PhisGILLNET has achieved F- measure of 100% [12] and it can be used to detect phishing at blog posts, chats and social networking posts.

Renato M. Silva et al. (2012), Authors have presented and evaluated various existing machine learning algorithms. Work is focused towards classifying [13] websites as ham or spam based on its content based features, link based features and transformed link based features. For experiment they used WEBSPAM UK2006 collection dataset. Monte carlo cross validation is used to define the size of training and testing subsets. Recall, precision and F- measure has been calculated for each model. Among all classifiers aggregation techniques such as bagging of trees and adaptive boost gave best result whereas SVM gave worst results.

Ammar Almomani et al. **(2013),** Authors have explained phishing email and its lifecycle. They have discussed classification and evaluation methods of phishing email along with different features of phish [14] email such as, basic features, latent topic model features, dynamic Markov Chain Features. They have thrown [14] light on various protection measures against phishing e mail such as network level protection, authentication technique, client side tools and filters, user education and server side filters and classifiers have been discussed. Various existing machine learning approaches for phishing email detection have been presented and evaluated. Approaches presented and evaluated in this study are methods based on bags of word model, multi classifier algorithm, classifier model based features, clustering approaches of phishing email, multi layered systems and evolving connectionist system to detect and classify phishing e mail. Any existing methods are not found to be very effective. As future work they have suggested to develop new approach that can work in an [14] online mode and effectively solve the limitations associated with zero day phishing email detection.

Birhanu Eshete et al. (2013), Authors have pointed the various attacks done by malicious websites and have proposed a lightweight and holistic system termed as BINSPECT. This system is used to detect malicious web page containing threats like malware, drive by download attack etc. This device has achieved detection accuracy above 97% with low false signals and an average performance overhead of at most 5 seconds [15]. They have introduced novel features and also enhanced existing ones so as to improve their discriminative power in the characterization of malicious and benign web sites [15]. An example of malicious web page has been given in this paper. In BINSPECT, they have used 11 URL features out of which 3 features are new along with page source features and social reputation features. BINSPECT has three major components which are feature extraction and labeling, multi-model training, and classification [15]. Full description of working of BINSPECT system has been given. They have taken seven classifiers for checking their efficiency to find best set of classifiers for the BINSPECT. They have shown that, new features added by them have increased the efficiency of 4 classifiers. Performance of BINSPECT has been compared with other detection software, where BINSPECT is proved to be better.

Shubhamoy Dey et al. (2013), Authors have compared the performance of probabilistic classifiers with and without the help of various boosting algorithm. Data set has been taken from Enron email dataset. Genetic Search algorithm has been used to select important features, which selected 134 features out of 1359 features. Naïve bayes and

Bayesian classifiers have been evaluated first then boosting algorithms have been used to enhance the performance of these classifiers. Bayesian classifier has performed better than naïve bayes. Boosting with Resample using Bayesian Classifier has given best result among all, with an accuracy of 92.9%. Adaboost has also given better results. As future work, boosting algorithms can be used with other base classifiers to do the comparison of performance.

Sajid Yousuf Bhat et al. (2014), Authors have evaluated various ensemble classifiers for spammer detection in social network. Bagging and boosting ensemble learning approach can be used to enhance the capability of base classifiers. Dataset has been taken from Facebook in which spammer behavior has been injected by author. Instead of using content based features, new network structure based features have been proposed to detect the spammers. Some base classifiers (J48, IBK, and Naïve Bayes) available in WEKA have been evaluated. Ensemble learning approach of bagging and boosting with base classifiers (J48, IBK and Naïve Bayes) have been evaluated using given dataset. Bagging ensemble learning approach using J48 has performed better than other evaluated classifiers.

Jemal Abawajy et al. (2014), Authors have compared various meta classifiers and done case study to construct new multilevel classifiers. Different meta classifiers have been used as base classifier to generate new meta classifiers. These new set of classifiers are multi level meta classifiers and termed as AGMLMC. These classifiers are intended to be used in distributed networking and computing. Various base classifiers, meta classifiers and AGMLMC classifiers have been compared for spam email classification. SMO has been used to work as base classifier [18] for meta classifiers at lower level as it gave best result. All combinations of Adaboost, Bagging, Multiboost have been tested to generate multi tier classifier. Diverse meta classifiers have been used to work at different levels of multitier classifier. Bagging at middle level and Adaboost at top [18] level of Multilevel classifiers have been proved to be best combination for AGMLMC. AGMLMC have been found to be best among all base classifiers and meta classifiers for filtering phishing emails. Authors have suggested to take some other large data set for classification using AGMLMC.

CHAPTER 3
PRESENT WORK

3.1. PROBLEM FORMULATION

E-mail spam classification is the current area of research. Previous work on e-mail spam classification has different type of limitations. Limitations are listed below.

- Spambase dataset with large number of features have been used. This leads to high data dimension which increases the time complexity to complete the classification process.
- Various classification algorithms and SVMs have been used which does not provide optimal performance to classify spam e-mails.
- Although base classifiers have given good performance level but boundary decision errors persist while classifications.

Contributions made by us in this work has been mentioned below,

1. Dimensions of dataset have been decreased thus reducing the time taken to complete the process of classification.
2. Ensembles of SVM RBF Kernel with Adaboost have been used for e-mail spam classification. Ensemble approach has been used on original dataset and dataset with reduced features. Boundary condition decision errors have been reduced with ensemble approach.
3. SVMs with different kernel functions have been used on dataset with reduced features. Effects of dimensionality reduction on performance of classifiers have been noted.

3.2. OBJECTIVES- Objectives of this research work have been given below.

1. To analyze the performance of existing classifiers using WEKA tool.
2. To implement Principal Component Analysis repeatedly to reduce the dimensionality of data.
3. To implement and evaluate performance of SVM with different kernel functions with and without PCA.
4. To build a hybrid model of SVM RBF Kernel with Adaptive Boost (Ada_SVM).
5. To implement hybrid approach (Ada_SVM) with and without PCA.
6. To compare metrics precision, accuracy and recall of all implemented classifiers and identify best technique to classify spam e-mail.

3.3. METHODOLOGY- Approach proposed by us has been depicted in following diagram.

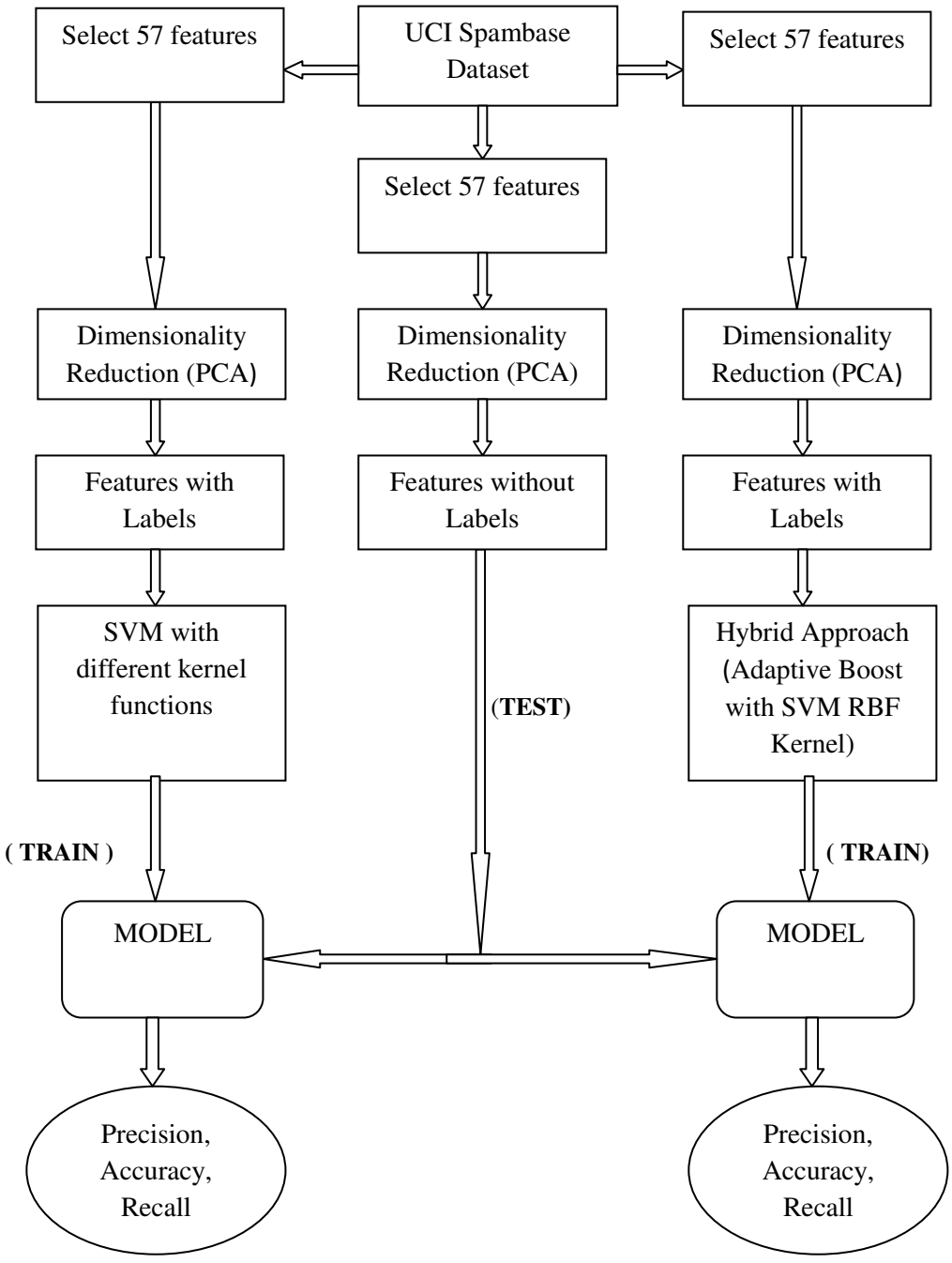


Fig 3.1: Flow Diagram of Proposed Approach

In this approach, dataset has been taken from UCI spambase corpora. Features will be reduced with the help of Principal Component Analysis (PCA) method. PCA will be used repeatedly to get different dimensions of data.

Support Vector Machines (with different kernel functions) will be used to classify spam e-mails with and without PCA.

Hybrid of SVM RBF with Adaptive Boost (Ada_SVM) will be used with and without PCA for e-mail spam classification. In this approach ensemble of SVM RBF will be made with Adaboost. In ensemble, different number of weak classifiers will be considered, i. e. 3, 5, 7, 10. Each and every ensemble will be tested on data with different dimensions.

We will evaluate different type of classifiers using Weka tool. Classifiers in Weka will be tested on original dataset; without PCA.

Finally, SVMs, SVMs-PCA, ADA_SVM, ADA_SVM-PCA and Classifiers in weka will be compared. Effects of data dimensionality reduction on the performance of classifiers for e-mail classification will be noted and presented by us

A brief introduction of PCA has been given below.

Principal Component Analysis (PCA) is mathematically defined as an orthogonal linear transformation that generates new set of axes for the data in which the greatest variance is represented by [10] first axis; second highest variance is represented by next axis and so on. Generated set of axes are termed as the principal components. PCA is a dimensionality reduction strategy which projects original data onto a smaller space.

Suppose that the data to be reduced consist of m attributes or dimensions. PCA finds m dimensional orthogonal vectors (principal components), where number of orthogonal vectors is less than m (attributes in original data). Generated principal components are stored in a sorted order of significance. Components with low variance can be eliminated to get the reduced data size.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1. DATASET- Dataset has been taken from UCI spambase corpus. Dataset consists of 4601 instances of e-mail messages. It consists of 57 attributes and one nominal attribute to show whether a given e-mail is spam (1) or not (0).

4.1.1. DATA DIMENSIONALITY REDUCTION - Principal Component Analysis (PCA) has been used to reduce the data dimensionality. Original dataset consist of 4601 instances. Firstly we have selected 2999 instances for training purpose and 1602 instances for testing purpose.

Dataset has 57 features; PCA has been used repeatedly to reduce the dimensionality of the data. In this work, training data has been represented by X and testing data has been represented by Y. We have extracted different number of features repeatedly from dataset in order to evaluate the classifier's performance on different number of features. X1, Y1 contains 50 features, X2,Y2 contains 45 features, X3,Y3 contains 40 features , X4,Y4 contains 35 features, X5,Y5 contains 30 features, X6,Y6 contains 25 features, X7,Y7 contains 20 features, X8,Y8 contains 15 features, X9,Y9 contains 10 features, X10,Y10 contains 5 features.

4.2. EXPERIMENTAL SETUP- We have used two different tools, WEKA and MATLAB in this work. A brief introduction of both the tools has been presented below.

MATLAB- MATLAB [21] is short term for matrix laboratory developed by mathworks. Using MATLAB different operations like matrix manipulations, implementations of algorithms, plotting of function and data can be performed [21] efficiently. Other packages like simulink can be used with MATLAB for getting additional features. Researchers widely use MATLAB for completion of their work.

WEKA- Weka (Waikato Environment for Knowledge Analysis) [24] has been developed by University of Waikato, New Zealand. Weka is suit of machine learning softwares written in Java. An interactive GUI is provided along with different algorithms for data analysis and predictive modeling. [24] Using Weka, different tasks of data mining such as, data preprocessing, classification, clustering, visualization, feature selection etc can be performed [wiki]. Explorer window in Weka, provides GUI where different data mining

tasks can be performed efficiently.

4.3. EXPERIMENTS- Experiments have been performed in three phases, which have been discussed below.

1. Evaluation of performance of existing classifiers using Weka tool.
2. Evaluation of SVM with different type of kernel functions, with and without PCA using MATLAB.
3. Ensemble of SVM RBF Kernel with Adaptive Boost (Adaboost), with and without PCA using MATLAB.

4.3.1. PERFORMANCE EVALUATION OF CLASSIFIERS USING WEKA- In WEKA tool, we have evaluated performance of various classifiers for e-mail spam classification. We used UCI Spambase dataset and ten fold cross validation has been used to divide training and testing set. Various algorithm performance evaluation metrics such as TP rate, FP rate, precision, recall and accuracy have been used for comparison.

BAYESIAN CLASSIFIERS- Among all Bayesian classifiers, DMNB has performed best, with an accuracy of 93.0015 percents and Naïve Bayes Multinomial has given worst result.

CLASSIFIER	TP RATE	FP RATE	PRECISION	RECALL	ACCURACY
BaysianLogistic Regression	0.814	0.132	0.86	0.814	81.3736
Bayes Net	0.898	0.124	0.898	0.898	89.8066
Complement Naïve Bayes	0.792	0.226	0.792	0.792	79.2002
DMNB Text	0.93	0.083	0.93	0.93	93.0015
Naïve Bayes	0.793	0.152	0.842	0.793	79.2871
Naïve Bayes Multinominal	0.791	0.233	0.79	0.791	79.0915
Naïve Bayes Updatable	0.793	0.152	0.842	0.842	79.2871

Table 4.1: Bayesian Classifiers

FUNCTIONS- Simple Logistic has given best results and voted perceptron has performed worst.

CLASSIFIER	TP RATE	FP RATE	PRECISION	RECALL	ACCURACY
Logistic	0.924	0.089	0.924	0.924	92.4147
RBF Network	0.807	0.173	0.823	0.807	80.6564
Simple Logistic	0.926	0.088	0.926	0.926	92.5886
SMO	0.904	0.122	0.905	0.904	92.4151
SPegasos	0.914	0.108	0.914	0.914	91.3932
Voted Perceptron	0.49	0.334	o.758	0.49	49.0328

Table 4.2: Function Classifiers.

DECISION TREES- Random Forest has performed best with an accuracy of 95.4575.

CLASSIFIER	TP RATE	FP RATE	PRECISION	RECALL	ACCURACY
AD tree	0.921	0.09	0.921	0.921	92.1321
BF Tree	0.927	0.084	0.927	0.927	92.7407
Decision Stump	0.78	0.274	0.78	0.78	78.0483
FT	0.933	0.071	0.934	0.933	93.3493
J48	0.93	0.078	0.93	0.93	92.9798
J48Graft	0.933	0.077	0.933	0.933	93.2841
LAD Tree	0.921	0.089	0.921	0.921	92.0887
LMT	0.804	0	1	0.804	80.3708
NB Tree	0.932	0.074	0.932	0.932	93.1971
Random Forest	0.955	0.955	0.955	0.955	95.4575
Random Tree	0.909	0.096	0.91	0.909	90.9368
Rep Tree	0.929	0.081	0.929	0.929	92.8928
Simple Cart	0.924	0.087	0.924	0.924	92.4368

Table 4.3: Decision Tree Classifiers.

RULE BASED CLASSIFIERS- Conjunctive Rule has given accuracy of 88.7468, which is best among all rule based classifiers. DTNB, Decision Table, PART are proved to be good.

CLASSIFIER	TP RATE	FP RATE	PRECISION	RECALL	ACCURACY
Conjunctive Rule	0.887	0	1	0.887	88.7468
Decision Table	0.84	0	1	0.84	84.0153
JRip	0.798	0	1	0.798	79.7954
OneR	0.76	0	1	0.76	75.9591
PART	0.82	0	1	0.82	82.0332
Ridor	0.746	0	1	0.746	74.5524
ZeroR	0	0	0	0	0
DTNB	0.85	0	1	0.85	84.9744

Table 4.4: Rule Based Classifiers.

LAZY CLASSIFIERS- Lazy classifiers have performed well on spambase dataset. IB1 and IBK have given well and same performance level.

CLASSIFIER	TP RATE	FP RATE	PRECISION	RECALL	ACCURACY
IB1	0.908	0.103	0.908	0.908	90.7846
IBK	0.908	0.103	0.908	0.908	90.7846
KStar	0.804	0	1	0.804	80.4348
LWL	0.748	0	1	0.748	74.8082

Table 4.5: Lazy Classifiers.

META CLASSIFIERS- CV Parameter Selection, Grading, Multi Scheme have performed worst, i.e. could not do email spam classification. Adaboost has given good results while classification. Many algorithms have given precision of 1.

CLASSIFIER	TP RATE	FP RATE	PRECISION	RECALL	ACCURACY
AdaBoost	0.901	0.112	0.9	0.901	90.0674
Attribute Selected classifier	0.782	0	1	0.782	78.1969
Bagging	0.815	0	1	0.815	81.4578
Classification via Clustering	0.848	0	1	0.848	84.7826
Classification via Regression	0.804	0	1	0.804	80.3708
CV Parameter Selection	0	0	0	0	0
Dagging	0.827	0	1	0.827	82.6726
Decorate	0.84	0	1	0.84	83.9514
END	0.802	0	1	0.802	80.243
Filtered Classifier	0.778	0	1	0.778	77.8133
Grading	0	0	0	0	0
Logic Boost	0.772	0	1	0.772	77.2379
Multi Boost AB	0.808	0	1	0.808	80.7545
Multi Class	0.752	0	1	0.752	75.1918
Multi Scheme	0	0	0	0	0

Table 4.6: Meta Classifiers.

Among all classifiers evaluated in Weka, Random Forest has performed best with an accuracy of 95.5, precision of .95 and recall of .95. Decision trees like FT, J48, J48 Graft, NB tree etc. have given good performance. Other classifiers such as IB1, IBK, Simple Logistic, SMO, DBNB Text, Adaboost have performed well with spambase dataset.

4.3.2. SVM WITH DIFFERENT TYPE OF KERNEL FUNCTIONS- SUPPORT VECTOR MACHINES (SVMs) with different kernel functions have been used with and without PCA.

1. CLASSIFIERS WITHOUT PCA- Classifiers have been used to classify the dataset without reducing the dimension of dataset; all 57 features have been used to take part in decision making process.

2. CLASSIFIERS WITH PCA- PCA has been used to extract the features repeatedly from dataset, i.e. training and testing data. After feature extraction, classification algorithm has been used on each reduced dataset.

Various classifiers used in the study are listed below.

- SVM with Linear kernel
- SVM with Quadratic kernel
- SVM with Polynomial kernel
- SVM with Gaussian Radial Basis Function kernel (RBF)
- SVM with Multilayer Perceptron kernel

All the classifiers have been evaluated on original dataset, i.e. without application of PCA, and on dataset with reduced feature. We intend to analyze the effects of feature reduction. Reduced dimensionality leads to less learning time for the classifier. We will verify effects of dimensionality reduction on performance of classifier, i. e. accuracy, precision and recall.

4.3.1.1 SVM WITH LINEAR KERNEL-

A. WITHOUT PCA- Results have been given below.

Dataset	Accuracy	Precision	Recall
Train,Test	.8290	.3695	.6117

Table 4.7: Performance Metrics (Linear SVM with PCA)

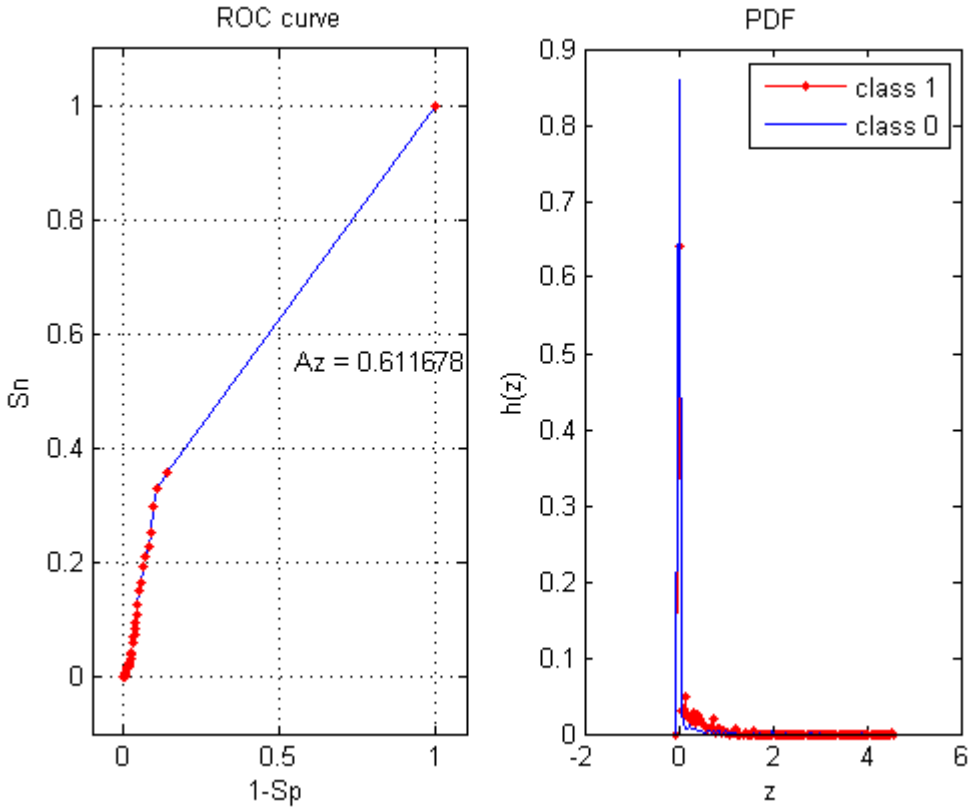


Fig 4.1: ROC Curve (Linear SVM without PCA)

B. WITH PCA- Results have been given below.

Dataset	Accuracy	Precision	Recall
X1,Y1	.4051	.5998	.6885
X2,Y2	.4064	.5998	.6885
X3,Y3	.4157	.5998	.6885
X4,Y4	.4076	.5998	.6885
X5,Y5	.4020	.5998	.6885
X6,Y6	.4032	.5998	.6885
X7,Y7	.4026	.5998	.6885
X8,Y8	.4007	.5998	.6885
X9,Y9	.4020	.5998	.6885
X10,Y10	.3602	.5998	.6885

Table 4.8: Performance Metrics (Linear SVM with PCA)

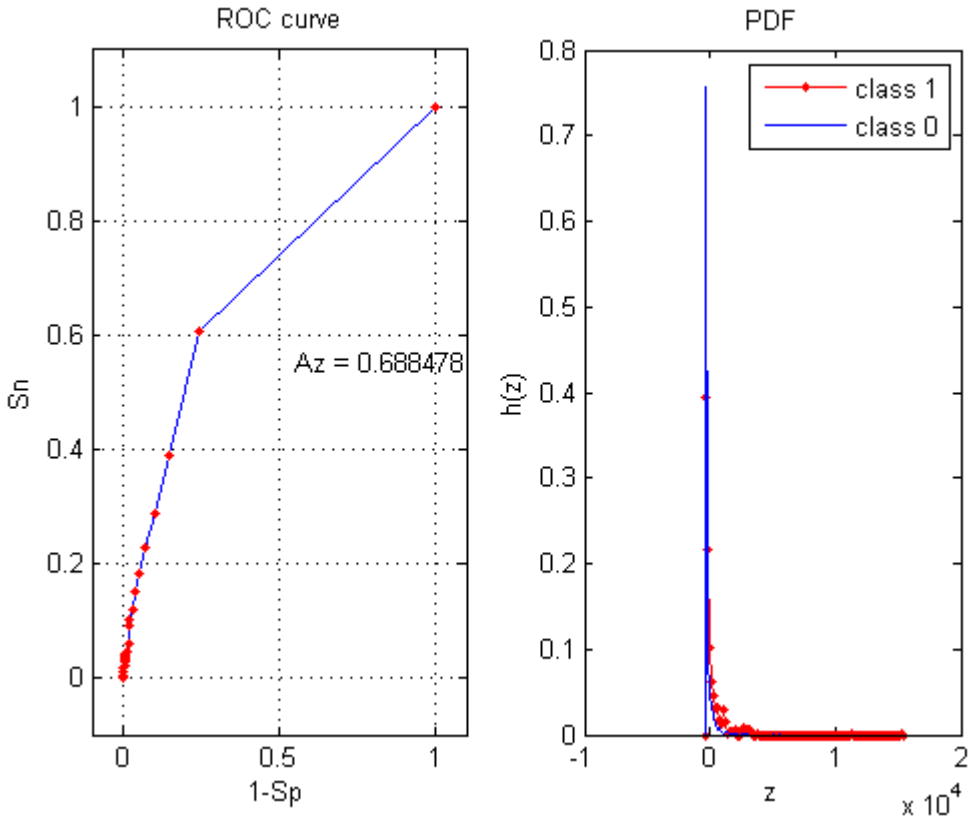


Fig 4.2: ROC Curve (Linear SVM with PCA)

4.3.2.1. SVM WITH QUADRATIC KERNEL-

A. WITHOUT PCA- Results have been given below.

Dataset	Accuracy	Precision	Recall
Train, Test	.7547	.3695	.6117

Table 4.9: Performance metrics (Quadratic SVM without PCA)

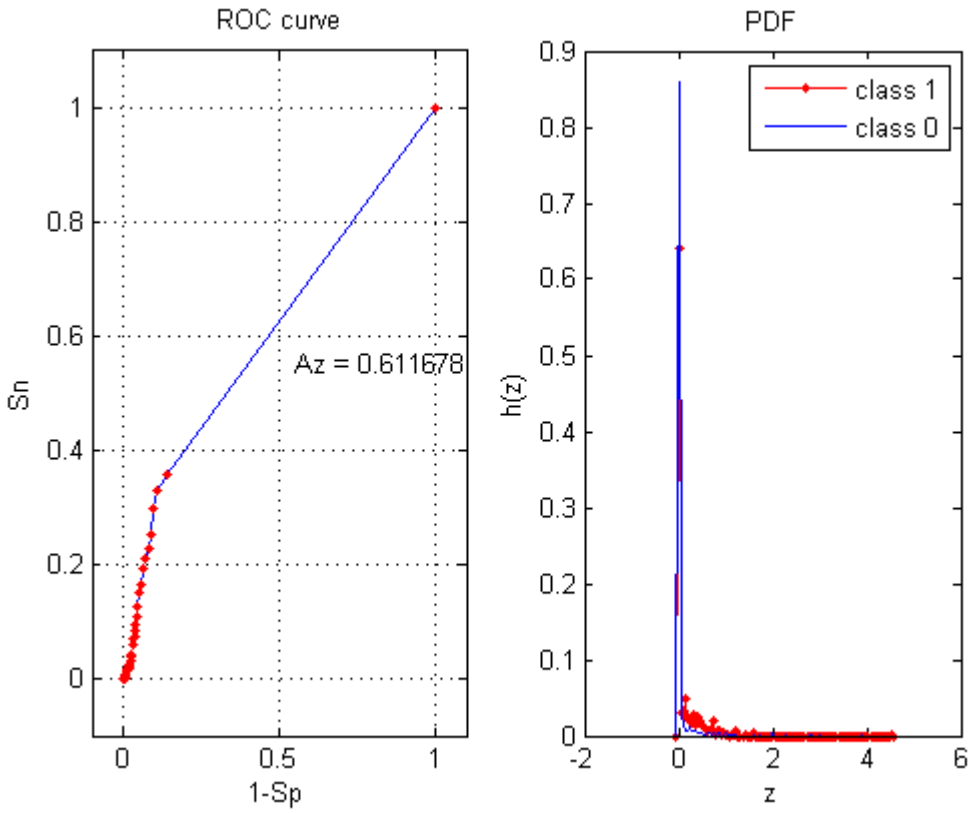


Fig 4.3: ROC Curve (Quadratic SVM without PCA)

B. WITH PCA- Results have been given below.

Dataset	Accuracy	Precision	Recall
X1,Y1	0.4775	0.5998	0.6885
X2,Y2	0.4482	0.5998	0.6885
X3,Y3	0.4457	0.5998	0.6885
X4,Y4	0.3801	0.5998	0.6885
X5,Y5	0.3770	0.5998	0.6885
X6,Y6	0.3870	0.5998	0.6885
X7,Y7	0.4120	0.5998	0.6885
X8,Y8	0.3390	0.5998	0.6885
X9,Y9	0.2971	0.5998	0.6885
X10,Y10	0.3727	0.5998	0.6885

Table 4.9: Performance metrics (Quadratic SVM with PCA)

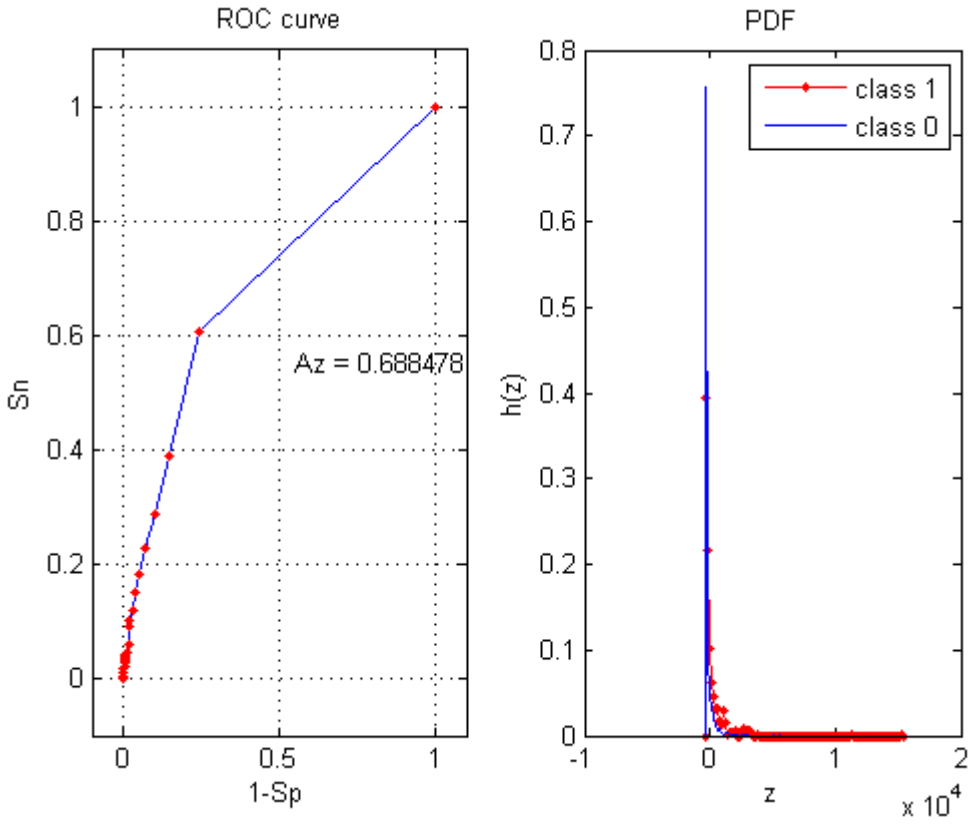


Fig 4.4: ROC Curve (Quadratic SVM with PCA)

4.3.2.3. SVM WITH POLYNOMIAL KERNEL-

A. Without PCA- Results have been given below.

Dataset	Accuracy	Precision	Recall
Train, Test	.5824	.3695	.6117

Table 4.10: Performance Metrics (Polynomial SVM without PCA)

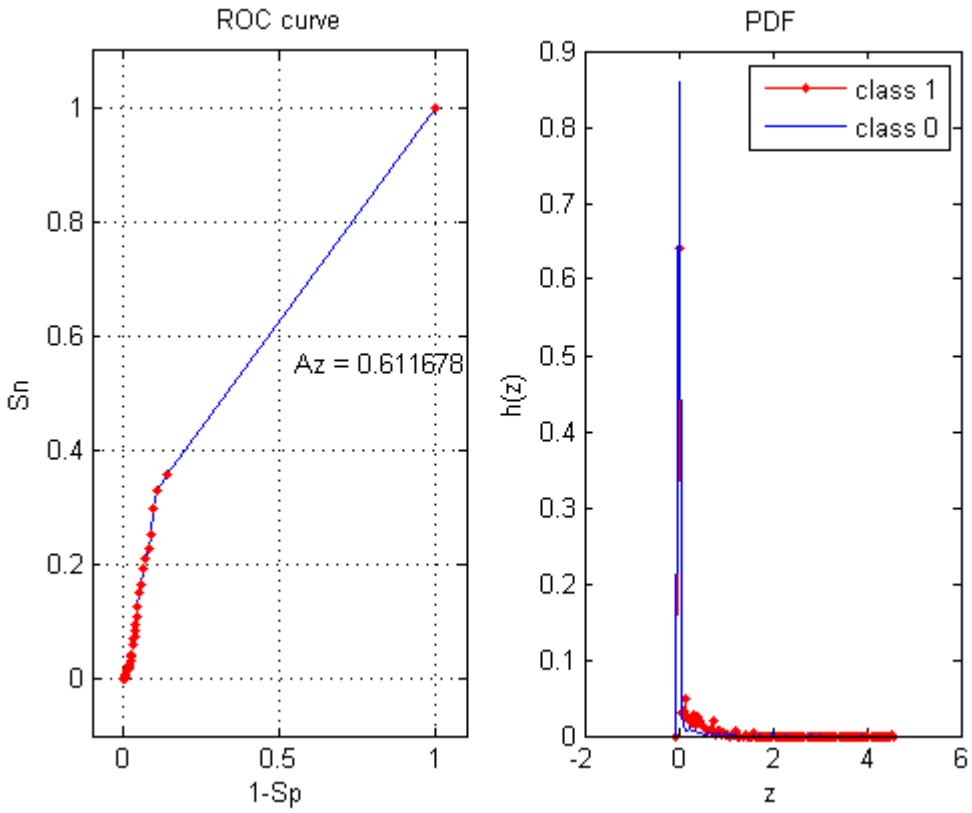


Fig 4.5: ROC Curve (Polynomial SVM without PCA)

B. With PCA- Results have been given below.

Dataset	Accuracy	Precision	Recall
X1,Y1	0.3901	0.5998	0.6885
X2,Y2	0.4026	0.5998	0.6885
X3,Y3	0.3664	0.5998	0.6885
X4,Y4	0.3933	0.5998	0.6885
X5,Y5	0.3995	0.5998	0.6885
X6,Y6	0.4757	0.5998	0.6885
X7,Y7	0.4875	0.5998	0.6885
X8,Y8	0.3989	0.5998	0.6885
X9,Y9	0.4164	0.5998	0.6885
X10,Y10	0.4713	0.5998	0.6885

Table 4.11: Performance Metrics (Polynomial SVM with PCA)

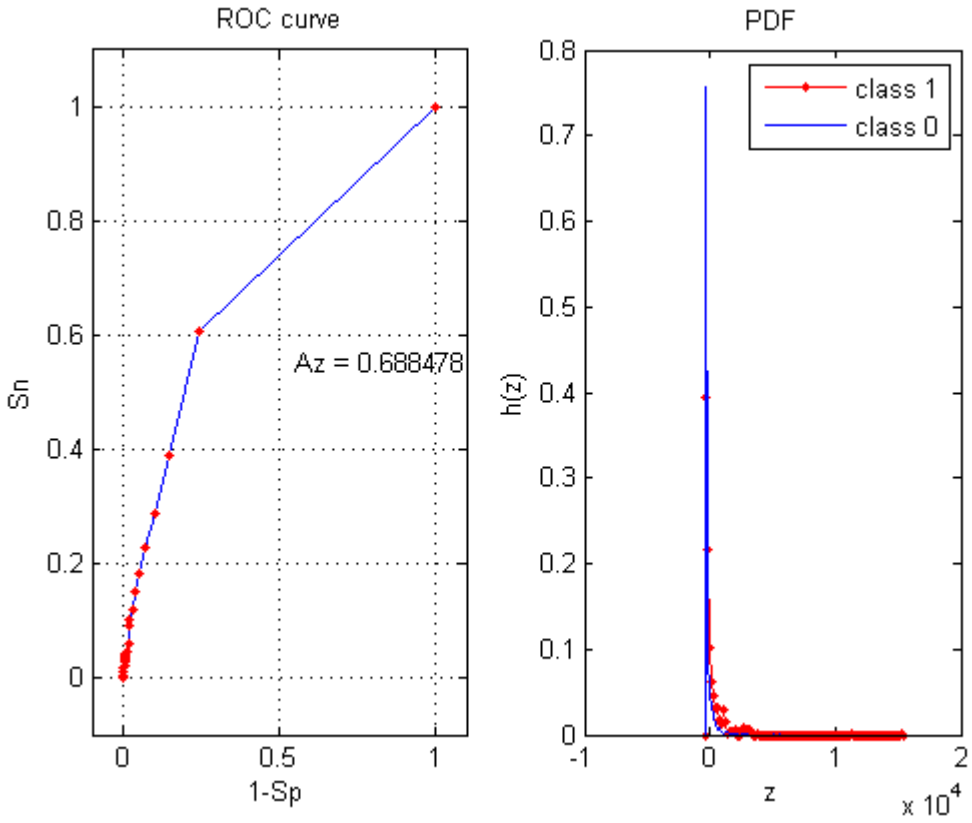


Fig 4.6: ROC Curve (Polynomial SVM)

4.3.2.4. SVM WITH RBF KERNEL

A. Without PCA- Results have been given below.

Dataset	Accuracy	Precision	Recall
Train, Test	.5175	.3695	.6117

Table 4.12: Performance Metrics (SVM RBF with PCA)

.

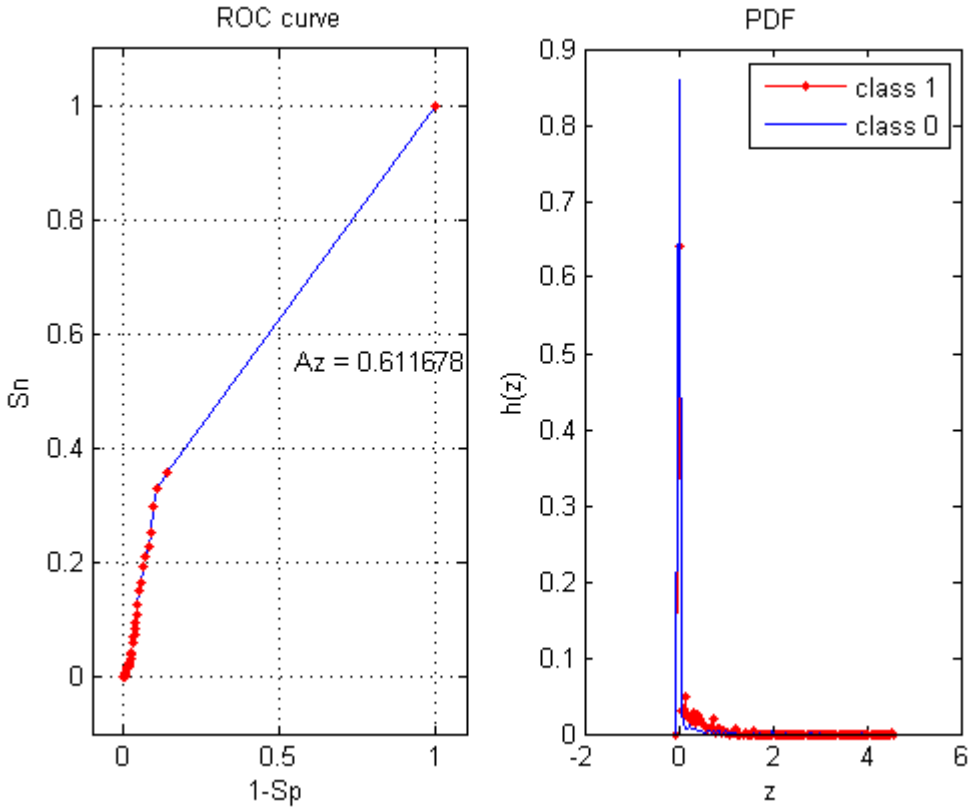


Fig 4.7: ROC Curve (SVM RBF Kernel without PCA)

B. With PCA-

Dataset	Accuracy	Precision	Recall
X1,Y1	0.3770	0.5998	0.6885
X2,Y2	0.3427	0.5998	0.6885
X3,Y3	0.3283	0.5998	0.6885
X4,Y4	0.3346	0.5998	0.6885
X5,Y5	0.3208	0.5998	0.6885
X6,Y6	0.3165	0.5998	0.6885
X7,Y7	0.3290	0.5998	0.6885
X8,Y8	0.3215	0.5998	0.6885
X9,Y9	0.3421	0.5998	0.6885
X10,Y10	0.4051	0.5998	0.6885

Table 4.13: Performance Metrics (SVM RBF with PCA)

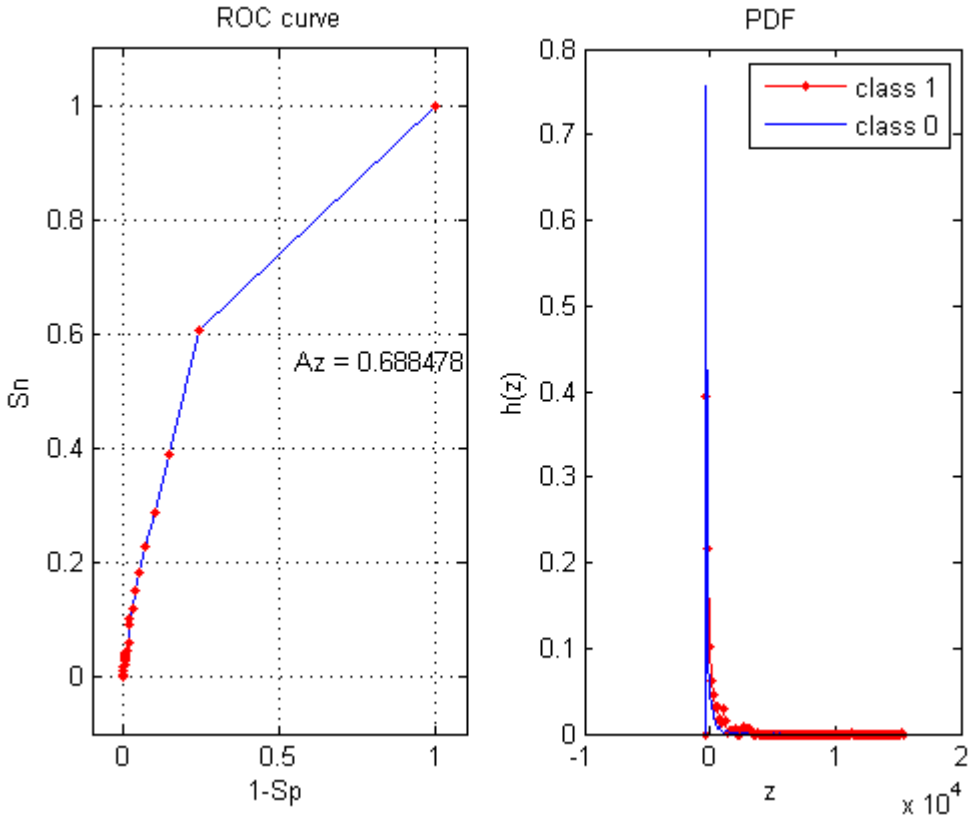


Fig 4.8: ROC Curve (SVM RBF with PCA)

4.3.2.5. SVM WITH MULTILAYER PERCEPTRON (MLP)-

A. Without PCA-

Dataset	Accuracy	Precision	Recall
Train, Test	.4363	.3695	.6117

Table 4.14: Performance Metrics (SVM MLP without PCA)

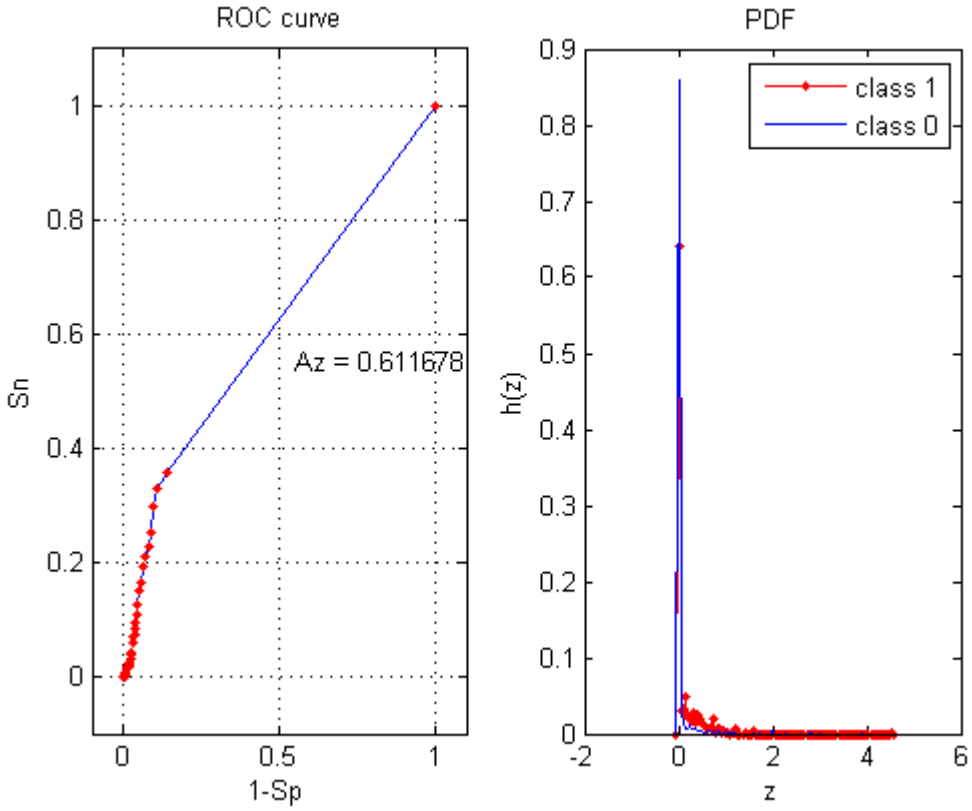


Fig 4.9: ROC Curve (SVM MLP without PCA)

B. With PCA-

Dataset	Accuracy	Precision	Recall
X1,Y1	0.5306	0.5998	0.6885
X2,Y2	0.5375	0.5998	0.6885
X3,Y3	0.4856	0.5998	0.6885
X4,Y4	0.5993	0.5998	0.6885
X5,Y5	0.4775	0.5998	0.6885
X6,Y6	0.5531	0.5998	0.6885
X7,Y7	0.5718	0.5998	0.6885
X8,Y8	0.4894	0.5998	0.6885
X9,Y9	0.5730	0.5998	0.6885
X10,Y10	0.5687	0.5998	0.6885

Table 4.15: Performance Metrics (SVM MLP with PCA)

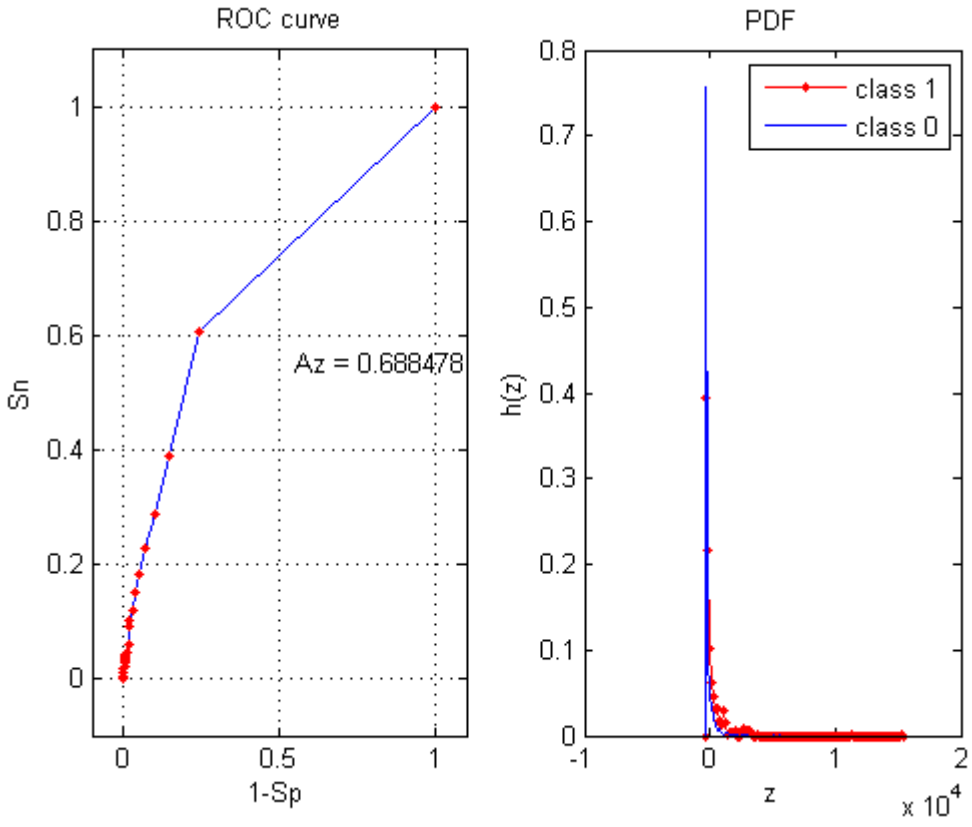


Fig 4.10: ROC Curve (SVM MLP with PCA)

SVMs WITHOUT PCA- Among all SVMs, Linear SVM has given best accuracy of 82.90 percent and SVM with multilayer perceptron kernel without PCA has given worst accuracy of 43.63.

SVMs WITH PCA- Accuracy of all SVMs has decreased whereas precision and recall has increased. PCA has been used repeatedly to get different dimensions of dataset. With each data sample, values of precision and recall are same for all classifiers.

We can see that all the classifiers are performing with same trend; after dimensionality reduction using PCA, accuracy gets decreased whereas precision and recall is increased. During e-mail spam classification, for achieving higher accuracy more features are required whereas a good level of precision and recall can be achieved with less number of features.

4.3.3. HYBRID OF SVM RBF KERNEL AND ADAPTIVE BOOST (Ada_SVM) – In hybrid approach (Ada_ SVM), ensembles of SVM RBF Kernel using Adaboost have been used for e-mail spam classification purpose. In this work we have used different number of SVMs in ensemble. Number of weak classifiers used in ensemble are 3, 5, 7 and10. We have evaluated the performance of hybrid classifiers each time with different number of weak classifiers.

4.3.3.1. Ada_SVM WITHOUT PCA (Dimensionality of dataset has not been reduced) - After all experiments, we came to know that with any number of weak classifier in ensemble learning we are getting same results. Performance is same for all ensembles, i.e. number of weak learners is not a constraint to decide performance of the classifier. Results have been given below.

Dataset	Accuracy	Precision	Recall
Train, Test	99.6242	99.4477	99.4090

Table 4.16: Performance Metrics (Hybrid Approach without PCA)

4.3.3.2. Ada_SVM WITH PCA (Dimensionality of dataset has been reduced)- Using PCA, dimensionality of spambase data has been reduced. Although Ada_SVM with different number of weak classifiers in ensemble has been used on different samples of dataset, performances of all classifiers on all dataset are same.

Dataset	Accuracy	Precision	Recall
X1,Y1	99.5955	99.4333	99.3936
X2,Y2	99.5955	99.4333	99.3936
X3,Y3	99.5955	99.4333	99.3936
X4,Y4	99.5955	99.4333	99.3936
X5,Y5	99.5955	99.4333	99.3936
X6,Y6	99.5955	99.4333	99.3936
X7,Y7	99.5955	99.4333	99.3936
X8,Y8	99.5955	99.4333	99.3936
X9,Y9	99.5955	99.4333	99.3936
X10,Y10	99.5955	99.4333	99.3936

Table 4.17: Performance Metrics (Hybrid Approach with PCA)

ENSEMBLE OF THREE WEAK CLASSIFIERS WITHOUT PCA- In hybrid approach (Ada_SVM), ensemble of three SVM RBF Kernel is used to classify spambase dataset. Training and testing errors have been represented by following graphs.

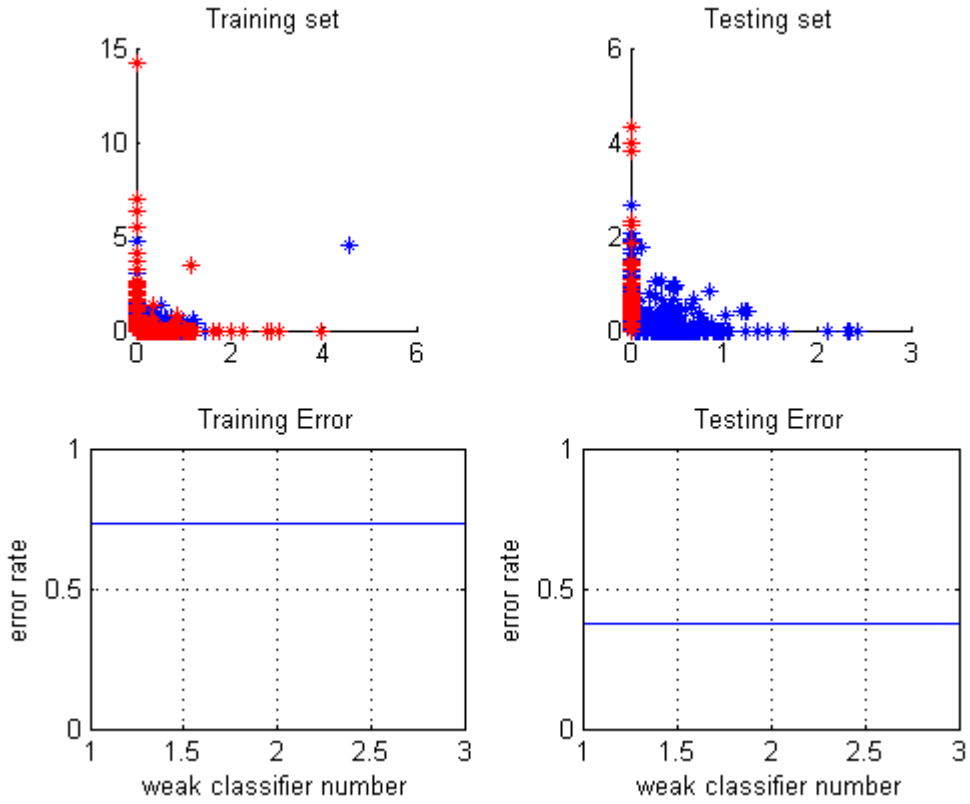


Fig 4.11: Ensemble of three classifiers (Without PCA)

ENSEMBLE OF THREE WEAK CLASSIFIERS WITH PCA- - In hybrid approach (Ada_SVM), ensemble of three SVM RBF Kernel with Adaboost is used repeatedly to perform classification process on dataset with reduced dimensions. Each time same performance level is obtained. So we have represented only one graph. Training and testing errors have been represented by following graphs.

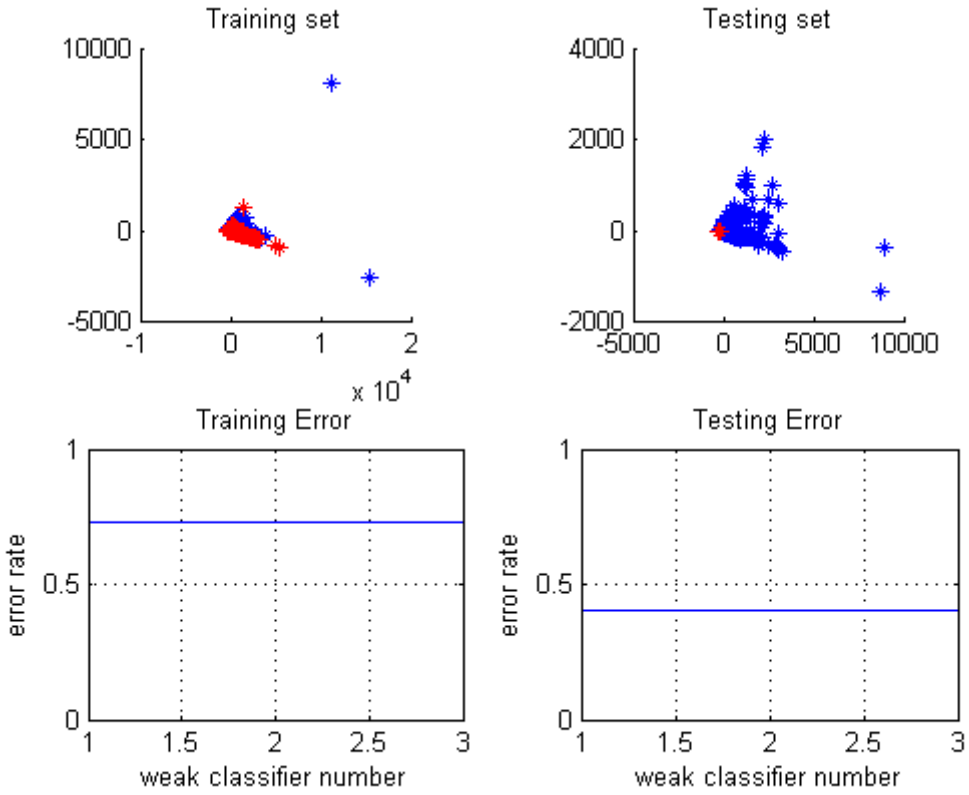


Fig 4.12: Ensemble of three classifiers (With PCA)

ENSEMBLE OF FIVE WEAK CLASSIFIERS WITHOUT PCA- - In hybrid approach (Ada_SVM), ensemble of five SVM RBF Kernel with Adaboost is used to classify spambase dataset. Training and testing errors have been represented by following graphs.

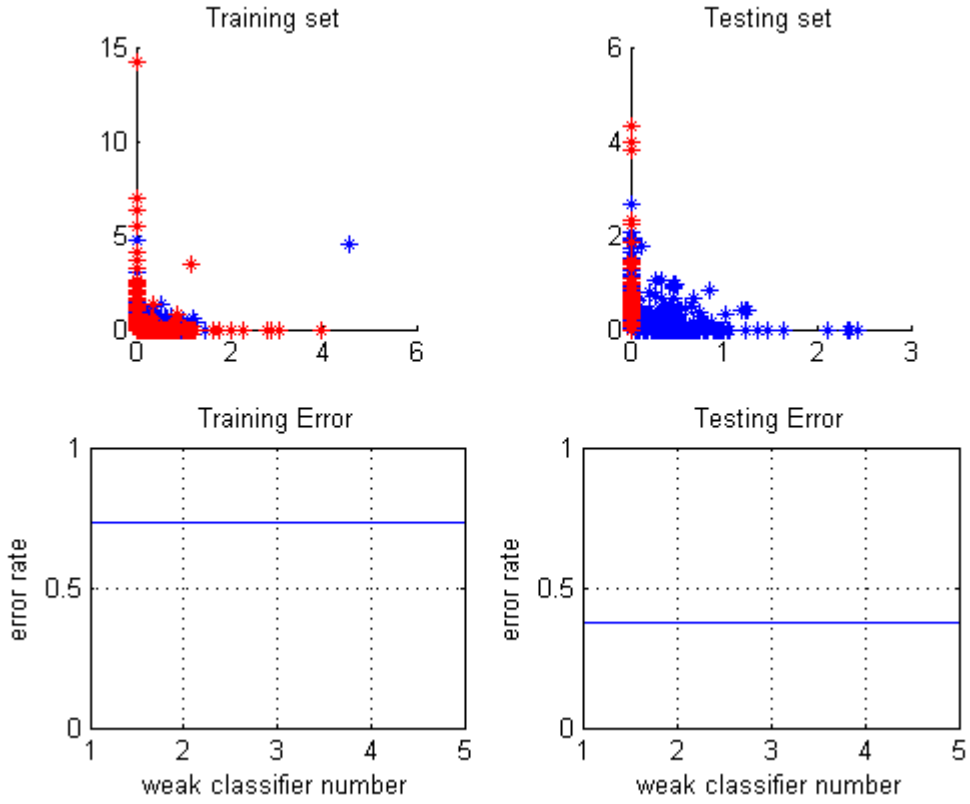


Fig 4.13: Ensemble of Five classifiers (Without PCA)

ENSEMBLE OF FIVE WEAK CLASSIFIERS WITH PCA- - In hybrid approach (Ada_SVM), ensemble of five SVM RBF Kernel with Adaboost is used repeatedly to perform classification process on dataset with reduced dimensions. Each time same performance level is obtained. So we have represented only one graph. Training and testing errors have been represented by following graphs.

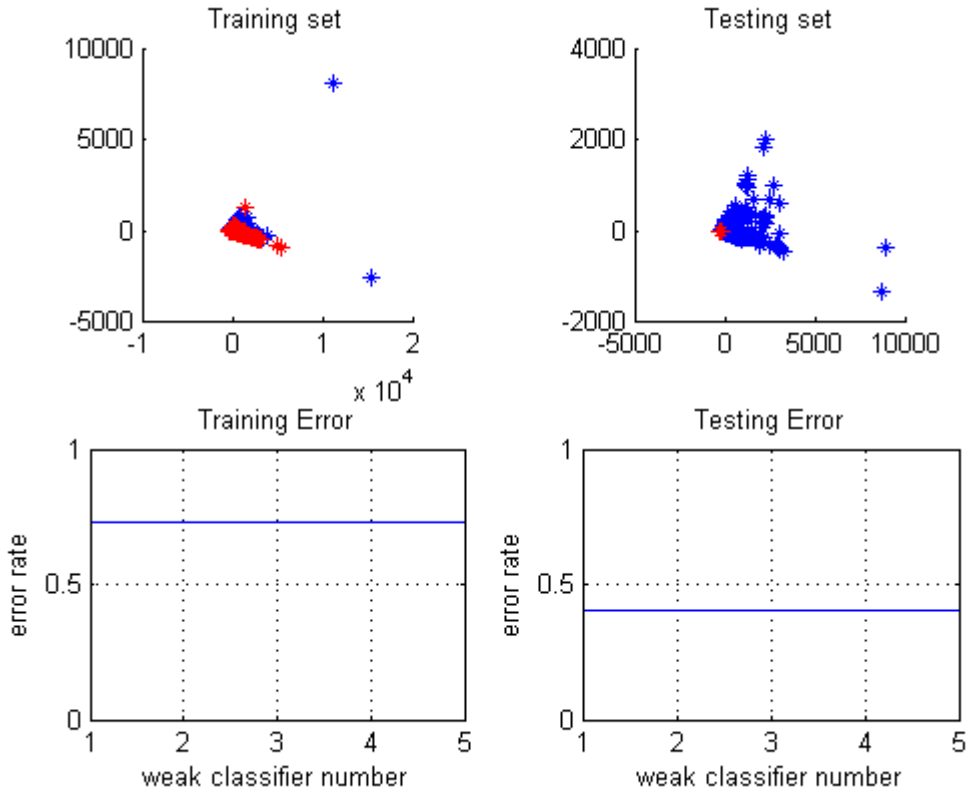


Fig 4.14: Ensemble of five classifiers (Without PCA)

ENSEMBLE OF SEVEN WEAK CLASSIFIERS WITHOUT PCA- - In hybrid approach (Ada_SVM), ensemble of seven SVM RBF Kernel with Adaboost is used to classify spambase dataset. Training and testing errors have been represented by following graphs.

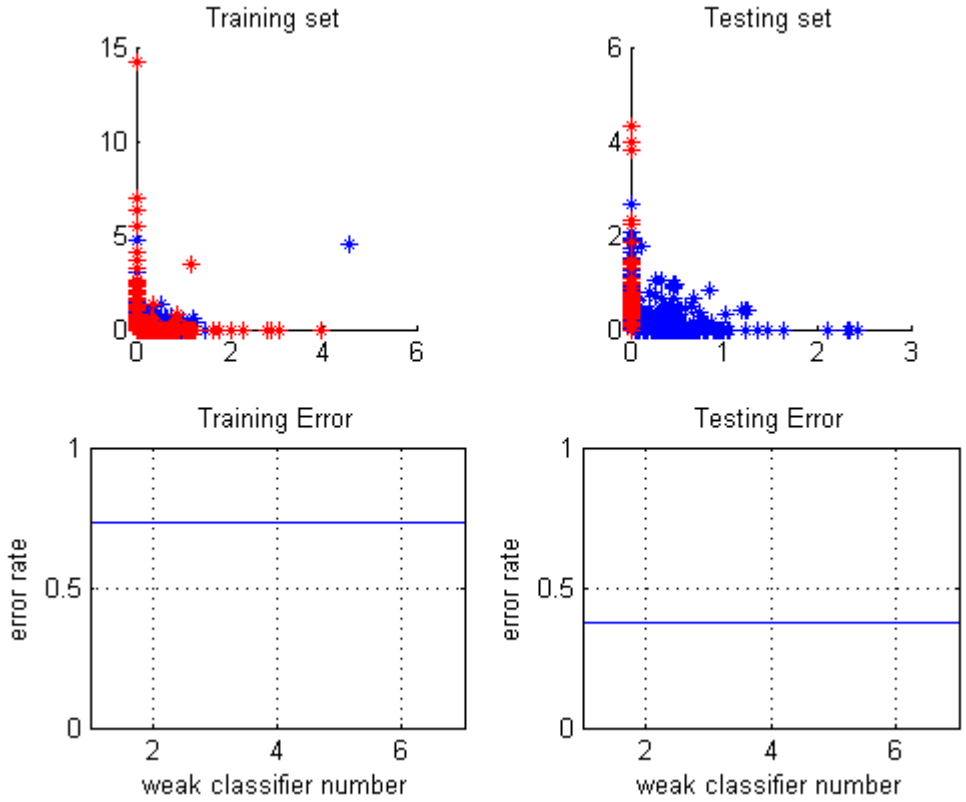


Fig 4.15: Ensemble of seven classifiers (Without PCA)

ENSEMBLE OF SEVEN WEAK CLASSIFIER WITH PCA- - In hybrid approach (Ada_SVM), ensemble of seven SVM RBF Kernel with Adaboost is used repeatedly to perform classification process on dataset with reduced dimensions. Each time same performance level is obtained. So we have represented only one graph. Training and testing errors have been represented by following graphs.

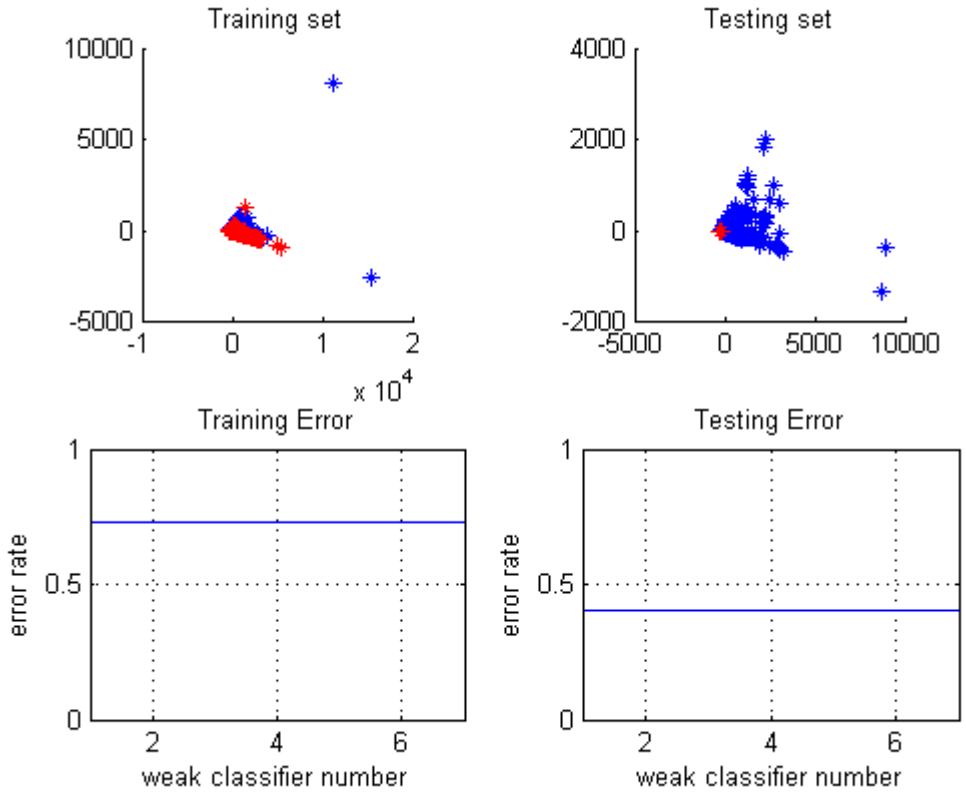


Fig 4.16: Ensemble of seven classifiers (With PCA)

ENSEMBLE OF TEN WEAK CLASSIFIERS WITHOUT PCA- - In hybrid approach (Ada_SVM), ensemble of ten SVM RBF Kernel with Adaboost is used to classify spambase dataset. Training and testing errors have been represented by following graphs.

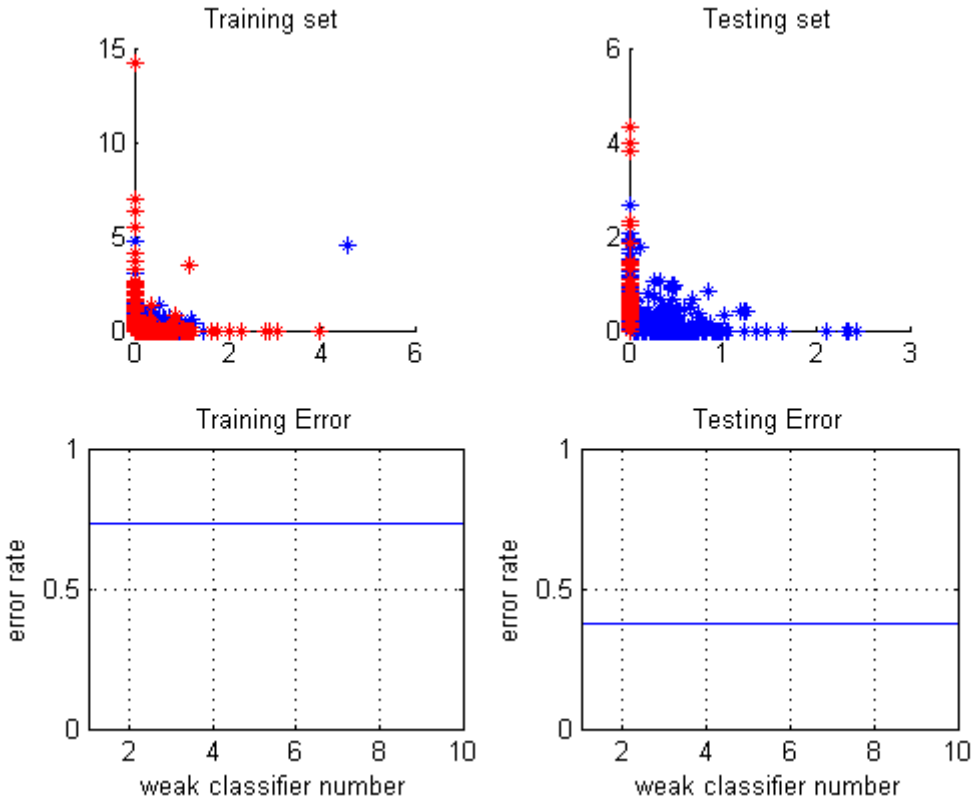


Fig 4.17: Ensemble of ten classifiers (Without PCA)

ENSEMBLE OF TEN WEAK CLASSIFIERS WITH PCA- In hybrid approach (Ada_SVM), ensemble of ten SVM RBF Kernel with Adaboost is used repeatedly to perform classification process on dataset with reduced dimensions. Each time same performance level is obtained. So we have represented only one graph. Training and testing errors have been represented by following graphs.

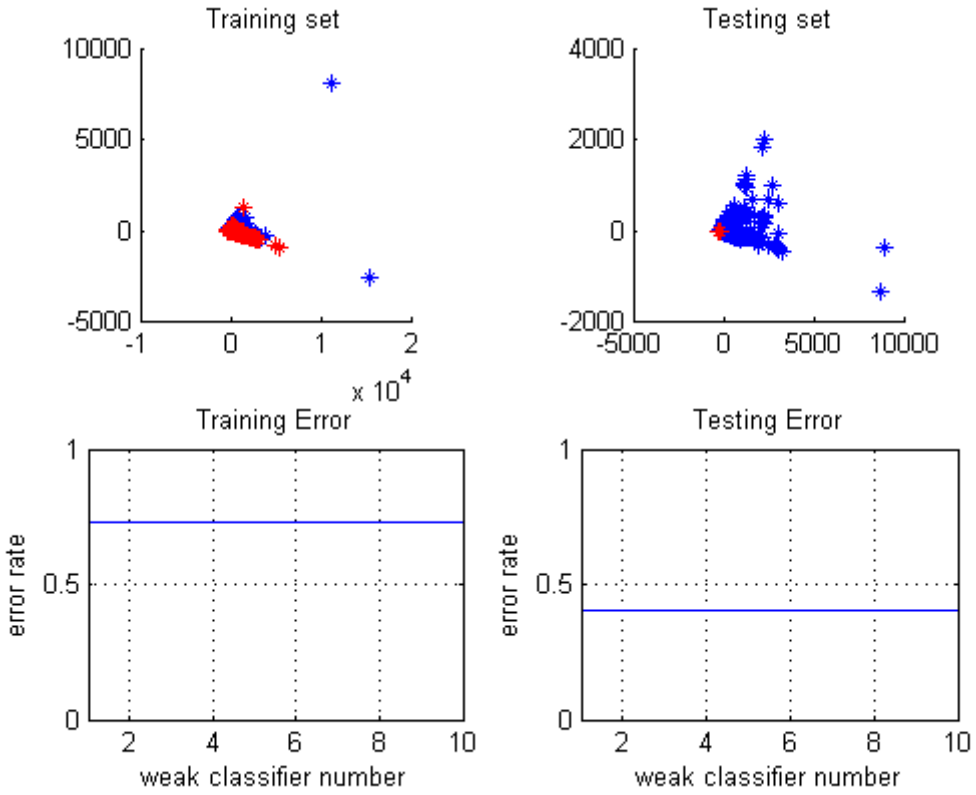


Fig 4.18: Ensemble of ten classifiers (With PCA)

Ensembles (in hybrid of Ada_SVM without PCA) of different numbers of weak classifiers i. e. 3, 5, 7, 10, have been used on data without reduction in dimensionality. Each ensemble has given same performance level.

Ensembles (in hybrid of Ada_SVM with PCA) of different numbers of weak classifiers i. e. 3, 5, 7, 10 have been used on data with reduced dimensions. Dimensions of data are 50, 45, 40, 35, 30, 25, 20, 15, 10, and 5. All ensembles have given same performance level on each and every data with reduced dimensions.

Experiments have been done repeatedly on dataset of various dimensions. Ensemble of large number of weak classifiers, require more time to complete classification process. Large data dimension increases time and space complexities incurred in process of classification. Ensemble of fewer weak learners can be used on data with fewer dimensions to achieve the good level of performance in less time but optimal performance can be achieved with all dimensions of data.

Results obtained by using ensemble learning approach, depicts that ensemble of SVM RBF Kernel with Adaboost have given best results when dimensionality of data has not been reduced. Hybrid model has given higher rate of accuracy, precision and recall.

Inference of this work can be summarized as, to achieve higher performance of hybrid model of SVM RBF Kernel with adaboost, data dimensions should not be decreased.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

We have done various experiments in this work; we have mentioned various relevant conclusions of this work below.

1. Among all techniques, hybrid approach (AdaBoost with SVM RBF) used to classify e-mails without reduction in data dimensions (without PCA) has given best results where accuracy is 99.6242%, precision is 99.4477% and recall is 99.4090%. Whereas accuracy, precision and recall given by hybrid approach with PCA are 99.5955%, 99.4333%, 99.3936% respectively while classifying spambase data. When dimensionality of data is reduced, performance of hybrid approach has also reduced slightly.
2. Number of weak classifiers (SVM RBF), i.e. 3, 5, 7, 10, in ensemble (in hybrid of ADA_SVM) is not a constraint to predict spambase data but dimensionality of data is a major constraint for the performance of hybrid classifier. So less number of weak classifiers in ensemble can be used on data with fewer dimensions to reduce the time taken for the classification process but optimal performance can be achieved only with all dimensions of data.
3. SVM with different type of kernel functions have been used for classification of spambase data. During e-mail spam classification, for achieving higher accuracy more features are required whereas a good level of precision and recall have been achieved with reduced dimensionality of data. SVM with linear kernel without PCA has given the accuracy of 82.90%, which is best among all SVMs.
4. Some of the classifiers evaluated in Weka have given better performance than SVMs. Random forest has given accuracy, precision and recall of 95.55%, 95.55% and 95.55% respectively, which is best among all classifiers evaluated in Weka.

In future we will enhance our approach by using ensemble of decision trees and use more set of features to classify spam e-mails.

CHAPTER 6

REFERENCES

I. BOOKS

[1] Han, Kamber and pei. (2012), Data Mining: Concept and Techniques, Morgan Kaufmann Publisher, Waltham USA.

[2] Berry, Linoff. (2004), Mastering Data Mining: The Art and Science of Customer Relationship Management, John wiley & sons, Singapore.

II. RESEARCH PAPERS

[3] Günal, S., Ergin, S., Gülmezoğlu, M. B., & Gerek, Ö. N. (2006). On feature extraction for spam e-mail detection. In *Multimedia content representation, classification and security* (pp. 635-642). Springer Berlin Heidelberg.

[4] Saberi, A., Vahidi, M., & Bidgoli, B. M. (2007, November). Learn to detect phishing scams using learning and ensemble? methods. In *Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences on* (pp. 311-314). IEEE.

[5] Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009, June). Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In*Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1245-1254). ACM.

[6] Wei-Chih, H., & Yu, T. Y. (2009, December). E-mail spam filtering using support vector machines with selection of kernel function parameters. In*Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on* (pp. 764-767). IEEE.

[7] Wang, W. (2010). Heterogeneous Bayesian ensembles for classifying spam emails. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*(pp. 1-8).

[8] Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2011). Learning to detect malicious urls. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 30.

[9] Aldwairi, M., & Alsalman, R. (2012). MALURLS: A Lightweight Malicious Website Classification Based on URL Features. *Journal of Emerging Technologies in Web Intelligence*, *4*(2), 128-133.

[10] Gomez, J. C., & Moens, M. F. (2012). PCA document reconstruction for email classification. *Computational Statistics & Data Analysis*, *56*(3), 741-751.

[11] Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March). Comparative study on email spam classifier using data mining techniques. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, pp. 14-16).

[12] Ramanathan, V., & Wechsler, H. (2012). phishGILLNET—phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training. *EURASIP Journal on Information Security*, *2012*(1), 1-22.

[13] Silva, R. M., Yamakami, A., & Almeida, T. A. (2012, December). An analysis of machine learning methods for spam host detection. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on* (Vol. 2, pp. 227-232). IEEE.

[14] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques. *Communications Surveys & Tutorials, IEEE*, *15*(4), 2070-2090.

[15] Eshete, B., Villafiorita, A., & Weldemariam, K. (2013). Binspect: Holistic analysis and detection of malicious web pages. In *Security and Privacy in Communication Networks* (pp. 149-166). Springer Berlin Heidelberg.

[16] Trivedi, S. K., & Dey, S. (2013). Interplay between Probabilistic Classifiers and Boosting Algorithms for Detecting Complex Unsolicited Emails. *Journal of Advances in Computer Networks*, *1*(2), 132-136.

[17] Abawajy, J., Kelarev, A., & Chowdhury, M. (2014). Automatic generation of meta classifiers with large levels for distributed computing and networking.*Journal of Networks*, *9*(9), 2259-2268.

[18] Bhat, S. Y., Abulaish, M., & Mirza, A. A. (2014, August). Spammer Classification Using Ensemble Methods over Structural Social Network Features. In *Proceedings of the*

2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02 (pp. 454-458). IEEE Computer Society.

III. WEBSITES

[19] https://archive.ics.uci.edu/ml/datasets/Spambassse

[20] http://en.wikipedia.org/wiki/Ensemble_learning

[21] http://en.wikipedia.org/wiki/MATLAB

[22] http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[23] http://en.wikipedia.org/wiki/Principal_component_analysis

[24] http://en.wikipedia.org/wiki/Support_vector_machine

[25] http://en.wikipedia.org/wiki/Weka_%28machine_learning%29

CHAPTER 7

APPENDIX

GLOSSARY OF TERMS-

A-

Algorithm- It is the step by step systematic procedure to solve some problem.

C-

Corpus- A large collection of writings or records of a specific kind or on a specific subject.

F-

Feature- Feature defines the property of object. An object may have one or more features.

K-

Kernel Method- kernel methods are a class of algorithms for pattern analysis**.**

M-

Machine Learning- Machine learning is a subfield of computer science which explores the construction and study of algorithms that can learn from and make predictions on data.

Malware- It is malicious software used for various intrusive and hostile activities. It can gain access to computer resources and can also fetch and steal personal information of user.

P-

Phishing- It is an attempt to get sensitive information of user or organization by impersonating as a trustworthy entity in electronic communications.

Principal Components – Principal components are set of axes (orthogonal vectors), which represent original data onto a smaller space.

ABBREVIATIONS

Ada_SVM- Hybrid of Adaptive Boost with SVM RBF Kernel

Adaboost- Adaptive Boost

GUI- Graphical User Interface

MATLAB- Matrix Laboratory

MLP- Multilayer Perceptron

MMH- Maximum Marginal Hyperplane

PCA- Principal Component Analysis

RBF- Redial Basis Function

SVM- Support Vector Machine

Weka- Waikato Environment for Knowledge Analysis