Classifying Spam Emails Using Artificial Intelligent Techniques

Introduction

Recently, the quantity of unwanted bulk email messages, i.e. spam messages has increased in web communication. These unwanted messages are degrading the reliability and authenticity of genuine email messages [1]. Spams have affected the organisations heavily.As a result, the data allocation capacity of the internet is getting difficult, which adds a high financial loss for companies. Many business models depending on the spam commercializing sector, have the advantage as the expense of sending an email needs less cost and can be sent with a large numbers**.** The above reasons forced few nations to amend some legislative changes [2]. In the literature, two types of spam filter techniques can be found: machine learning and non-machine learning (SMTP) methods. However**,** machine learning techniques have more popularity than non-machine learning approaches[3]. The subdivision of machine learning strategies can be done further in content based and non-content based. Plenty of research has been carried out on machine learning based spam filters[1][3]. False positives are the main concern as they can take to different implication while filtering spams. Therefore, there might be circumstances, where ham messages may be classified as spams. Though such situations can be rather lessened by the simultaneous application of an amount of dissimilar classification techniques, but it continues to be a matter of debate [4] [5]. In addition to that, the execution of a classifier also depends on the training set. The deportment of the spammer

Data set

In this work, the spam data set namely *spambase* has been collected from UCI machine learning repository for the experiment purpose[14].ELM technique uses the basic feed-forward neural network for classification and regression [11]. The total number of email instances are 4601.Out of which 1813 are actually spam emails which holds 39.4% of total email instances and remaining 60.6% i.e. 2788 are non-spams. The total number of features

is 58, of which 57 attributes are continuous and there exists one titular class label (d). The different type of dataset on spam emails are as follows,

Table 1. Different types of spam corpora[16][17]

Data Set	No of Emails	Spam Emails	Legitimate Emails
PUI	1099	481	618
Ling Spam	2893	481	2412
U5Spam	982	82	90
UCI	4601	1813	2788

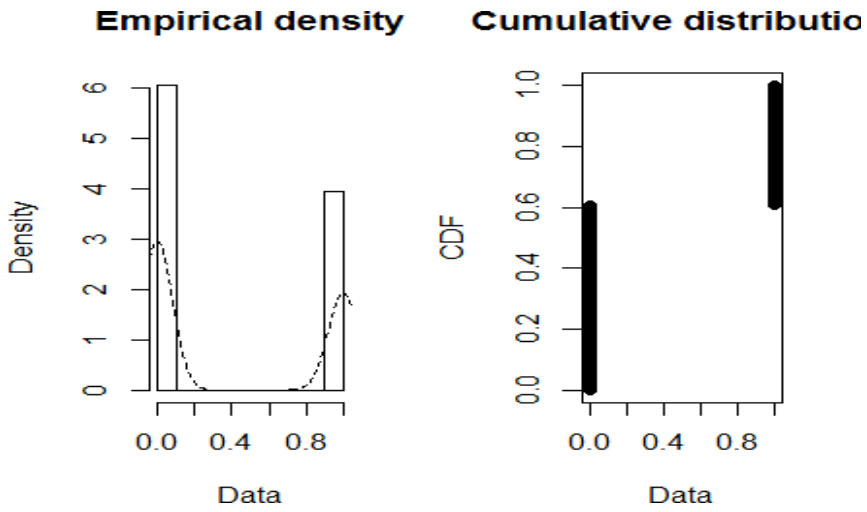
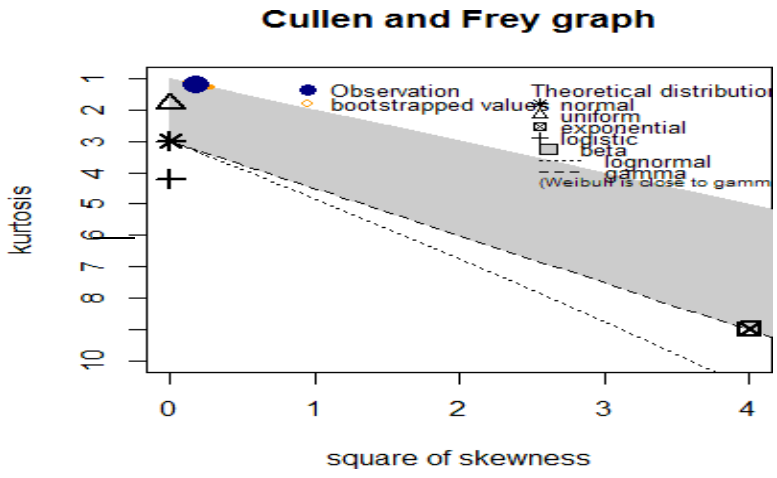


Figure 1. Empirical and Cumulative distribution ofdata setTable 2. Statistical output of the data set namely spambase

d	mean	0.3940448
	skewness	0.4338114
	sd	0.4886977
	kurtosis	1.187404

Figure 2. Cullen and Frey Graph applied to data set



Recall= $\frac{TP}{TP+FN}$.

Positive

Negative

McNemar test(P value): $\chi 2=^{(f-g)}$,

$TP+FP$

predictive value= $\frac{TP}{TP+FP}$

predictive value= $\frac{TN}{TN+FN}$

secondnegative test

(f+g)

and g is the outcome of first negative test and second positive test. The detailed comparisons between ELM and SVM has been shown in Table 4 and Table 5 ,where all the above parameters' values have been obtained.

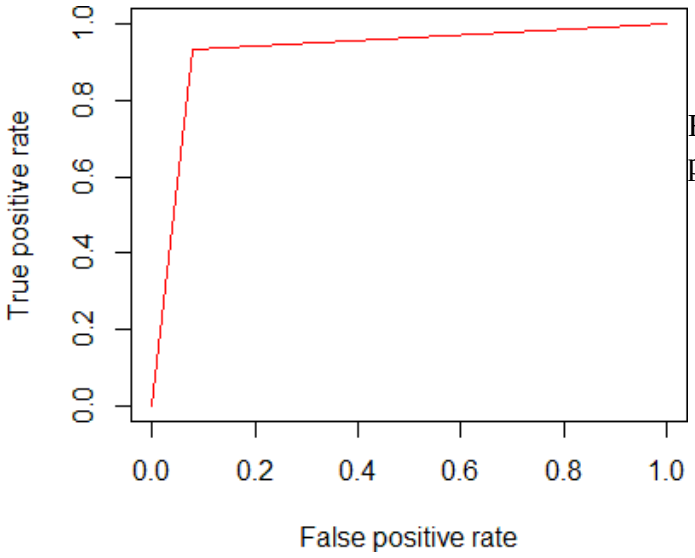
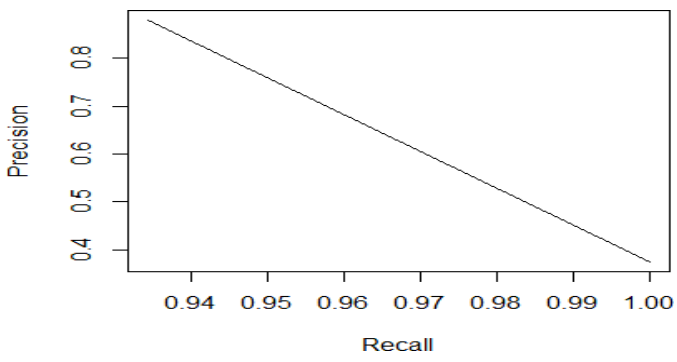


Figure 3. False positive vs True positive



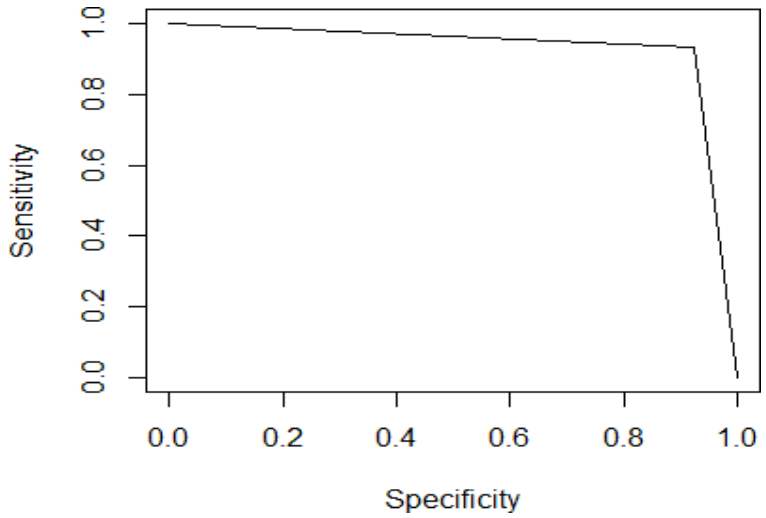


Figure 5. Specificity vs Sensivity

The Support Vector Machine Technique of Spam Classification

Support vector machine is a strong statistical learning theory used frequently for classification of data.SVM has been introduced by vapnik [11] and attaining popularity for its attractive features. In this work, the spam email classification has been limited by two classes : spam and non-spam.The data set is having the training vector S consisting two different classes that are spam or ham,

$$S = \{(x^1, y^1) \dots \dots (x^n, y^n), \text{ where } x \in R^n, y \in \{-1, +1\} \quad (4)$$

Where, $x \in R^n$, is a vector which has dimension n, with every instance either belongs to spam or non-spam(ham). In this work, our intention is to discover a generalized type of classifier that can separate two classes, i.e spam and non-spam {-1,+1} given the train data set.The distinguisher is nothing but a linear plane of the form,

$$f(x) = \text{w}.x + k = 0, \quad (5)$$

Where, w is the weight to segregate the hyperplane and k$\in R$ is a bias. In this work the hyper planes correspond to spam emails and non-spam emails can be written as,

$$\text{w}.x_\text{i} + k \geq 1, \ (for\ y_\text{i} = 1) \quad (6)$$

$$\text{w}.x_\text{i} + k \leq -1, \ (for\ y_\text{i} = -1) \quad (7)$$

Equation (6) & (7) can be combined together as,

$$y_\text{i}(\text{w}.x_\text{i} + k) \geq 1 \quad (8)$$

The slack variable is many times added to adjust the miss classification rate, therefore equation (11) can be updated as,

$$y_\text{i}(\text{w}.x_\text{i} + k) \geq 1 - £_\text{i} \quad (9)$$

For, spam class, the perpendicular distance from source to the hyperplane

$\text{w}.x_\text{i} + k = 1$, is

$\frac{|k-1|}{||w||}$, and for non-spam class the hyper plane can be written as,

i

w. $x + k = -1$ and the perpendicular distance can be written as, $\frac{|k+1|}{|w|}$.

Moreover the margin, $\rho(w,k)$ between the plane can be given as,

$$\rho(w,k) = \frac{2}{||w||}$$

To maximize this margin the following optimization problem forms,

Minimize $\frac{1}{2}||w||^2 + C\sum_{i=l} \pounds_i$

(10)

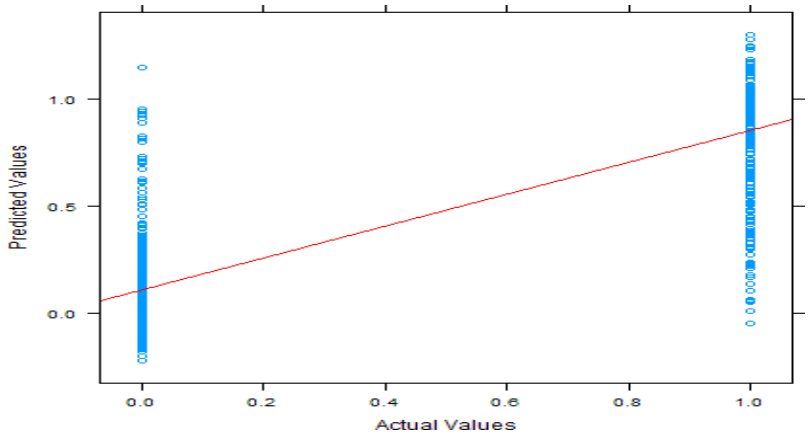
(11)

C is the capacity factor.

Table 3 . Detail outputs obtained by SVM

SVM	
parameter	epsilon = 0.1 cost C = 5
Kernel	Gaussian Radial Basis kernel function.
sigma	0.05
Number of Support Vectors	2264
Objective Function Value	-1757.389
Training error	0.081035
Cross validation error	0.066512

Below are the graphs (figures 6,7,8 and 9) which show the actual vs predicted plot,false positive vs true positive,precision vs recall and sensitivity vs specificity and accuracy plots once applied to test spam emails.



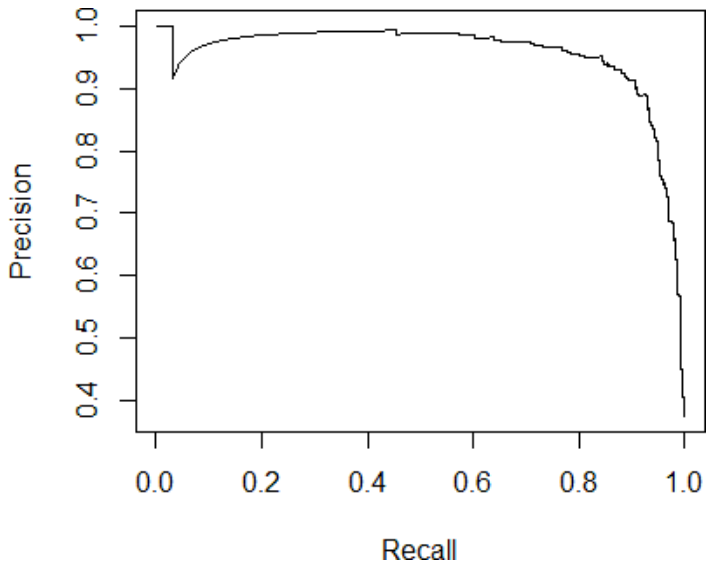


Figure 6 . Actual vs Predicted

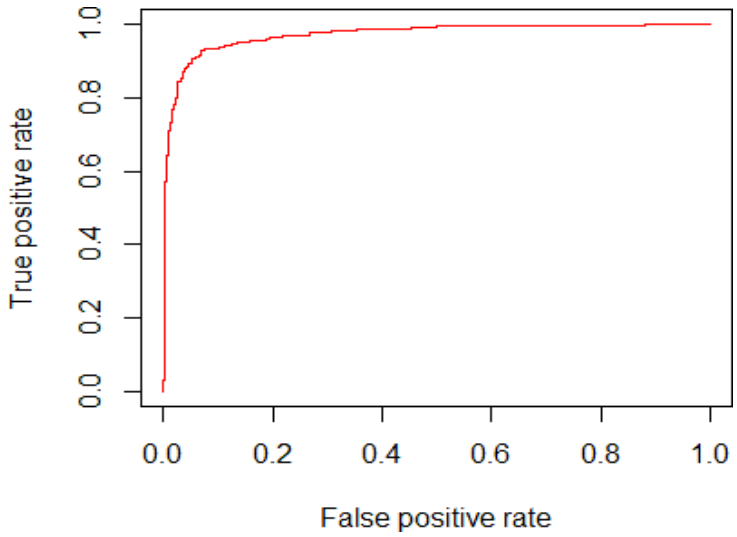


Figure 7. False positive vs True positive