IMPROVED SPAMBASE DATASET PREDICTION USING SVM RBF KERNEL WITH ADAPTIVE BOOST

Sneha Singh[1], Sandeep Kaur[2]

[1]*Student, Dept. of CSE, Lovely Professional University, Phagwara, Punjab*
[2]*Assistant Professor, Department of CSE, Lovely Professional University, Phagwara, Punjab*

Abstract

Spam is no more garbage but risk as it includes virus attachments and spyware agents which make the recipients' system ruined, therefore, there is an emerging need for spam detection. Many spam detection techniques based on machine learning algorithms have been proposed. As the amount of spam has been increased tremendously using bulk mailing tools, spam detection techniques should deal with it. In this paper we have proposed Hybrid classifier Adaptive boost with support vector machine RBF kernel on Spambase dataset. We have also extracted the features first by Principal component analysis.

General Terms: *Email Spam classification.*

Keywords: *Adaboost, classifier, ensemble, machine learning, spam email, SVM.*

----------***----------

1. INTRODUCTION

This is the era of internet in which we can access different kind of information easily from anywhere. Email is one of the most important solutions provided over internet. Email enables users to send messages in a very fast and economical way. Although Email is a good source of information exchange some people try to misuse it and do illegitimate work. People who use email accounts for wrong purpose are termed as spammers and email sent by them is known as spam email. Spam is very annoying problem which is being faced by almost everyone having an email account. Spammers flood network with unwanted bulk emails which is also termed as junk email. Spam email may be phishing email, it may contain some malware or it may be just unwanted advertisement. So filtering of spam email before sending it to the inbox of users is very important and challenging task.

Various Machine learning methods are being used to classify spammer's emails from legitimate emails. Different type of classifiers to detect spam email has been used and evaluated in past research work. Although we have got good filtering techniques but still there is requirement of some better filtering techniques. So spam email filtering is major area to focus in the present field of research.

2. RELATED WORK

This section contains a brief presentation of previous work done by researchers for classification of spam emails.

In [1], various classification and evaluation methods of phishing email along with different features of phish email such as, basic features, latent topic model features, dynamic Markov Chain features have been discussed. Some light has been thrown 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 classifier. Various existing machine learning approaches for phishing email detection have been discussed. 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 online mode and effectively solve the limitations associated with zero day phishing email detection.

In [2], Authors have presented and evaluated various existing machine learning algorithms. Work [2] is focused towards classifying 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. Among all classifiers aggregation techniques such as bagging of trees and adaptive boost gave best result whereas SVM gave worst results.

In [3], Authors have 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 termed as AGMLMC. Various base classifiers, meta classifiers and AGMLMC classifiers have been compared for spam email classification. All combinations of Adaboost, Bagging, Multiboost have been tested to generate multi tier classifier. Bagging at middle level and Adaboost at top 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.

In [4], In this paper, authors have analyzed various machine learning spam classification algorithms. E-mail spam 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.

In [5], Authors have used three different learning methods and one ensemble method to detect phishing emails. Three data mining algorithms [5] have been used to detect phish email (scam) namely, K nearest neighbor, Poisson probabilistic theory and Bayesian probabilistic theory. 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 ensemble classification algorithm have been used, in which their results are merged in order to increase the accuracy of classification

In [6], Work is focused on e mail classification using text content features only. Classifier uses principal component analysis document reconstruction (PCADR), which is able to extract and synthesize the important features [6] of document for efficiently representing any class. PCADR approach has been tested on different e mail corpora such as PU1, Ling Spam, SpamAssassin, Phishing and TREC7 spam corpus. PCDAR proved to be better than SVM in terms of classification accuracy and classification time. PCADR is well suited when training and testing data are from different sources.

In [7], Authors have proposed a new server side methodology to detect phishing attacks namely phishGILLNET. PhishGILLNET consists of multiple layers in which the first layer makes [7] use of Probabilistic Latent Semantic Analysis (PLSA) to build a topic model. The second layer uses AdaBoost to build a classifier. The third layer makes a classifier from labeled and unlabeled examples by 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 [7] was compared with SVM, where phishGILLNET1 performed better. phishGILLNET2 supports both 3-class and binary 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.

In [8], Authors have evaluated various ensemble classifiers for spammer detection in social network. 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.

In [9], 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.

3. PARAMETERS TO EVALUATE THE PERFORMANCE OF CLASSIFIER

Parameters to evaluate performance of spam filtering tool have been described below,

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

Precision = (TP)/(TP+FP)

Recall = (TP)/(TP+FN)

Different abbreviations used above are as follows.

Positive (P): Total number of spam emails.

Negative (N): Total number of ham emails.

True Positive (TP): Total number of spam email correctly classified as spam.

True Negative (TN): Total number of ham emails correctly classified as ham.

False Positive (FP): Total number of ham emails misclassified as spam.

False Negative (FN): Total number of spam emails misclassified as ham.

Confusion Matrix: Confusion matrix is a tool to analyze the performance of a classifier.

		Predicted	
		Positives	Negatives
Actual	Positives	TP	FN
	Negatives	FP	TN

Fig 1: Confusion Matrix

4. PROPOSED APPROACH

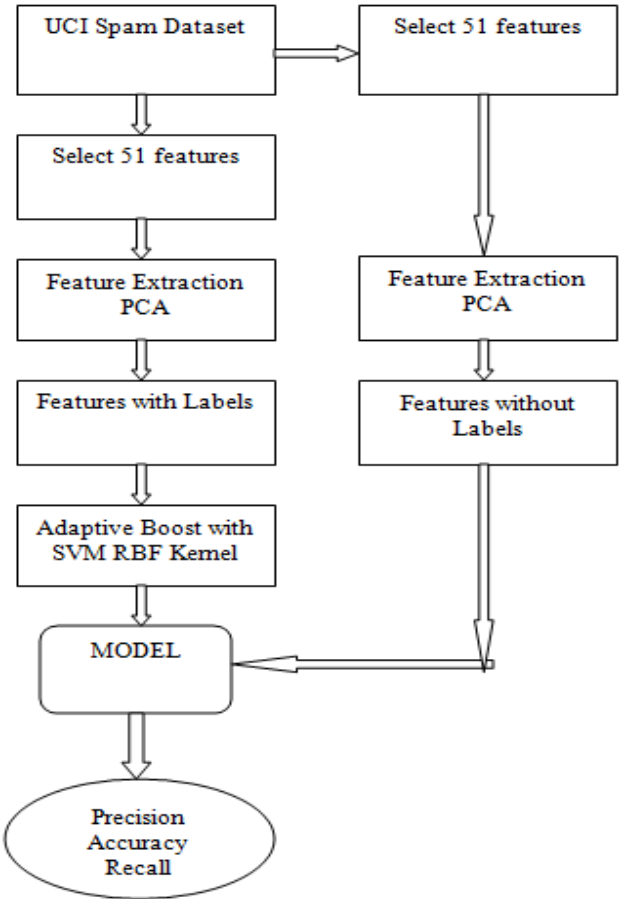

Fig 2: Flow Diagram of Proposed Approach

Introduction of Algorithms Used in Proposed Work

Combination of classifiers is being used to increase accuracy of classification results. Base classifiers can be used with meta classifiers to enhance the performance level of base classifiers.

A brief description of various concepts used in our proposed method have been described below,

4.1 Principal Component Analysis (PCA)

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.

4.2 Adaptive Boost (Adaboost)

Adaptive boost also termed as adaboost, is a very popular machine learning meta algorithm which can be used to enhance the performance of other learning algorithms. Using adaboost, weighted vote of multiple weak learners can be used to predict a class label in a more precise way.

In adaboost, boosted classifier is trained in a different way. A boost classifier is of the following form,

$$AN(x) = \sum_{n=1}^{N} \mathrm{aK(x)}$$

Where aK is a weak learner and x is input to weak learner. Training process will go through N iterations, where numbers of weak learners are N. At each iteration a weight is assigned to each sample of training set [11].

4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning model used for regression analysis and classification purpose. SVM generates a set of hyperplanes from which maximum marginal hyperplane is selected. It is also termed as binary linear classifier as it classifies test data in one of the two class labels.

Kernel tricks are applied on SVMs to make classification more accurate. Using kernel trick SVMs can perform nonlinear classification. Some important kernel tricks are Gaussian Redial Basis (RBF), polynomial and hyperbolic tangent.

The RBF kernel on two samples x and y, [13] is defined as

$$K(x, y) = \exp(-(\| x - y \|^2)/2\sigma^2)$$

Where $\| x - y \|^2$ is squared Euclidean distance between feature vectors [13] and σ is a free parameter.

5. CONCLUSION

In this paper, we have presented an optimal spam detection model based on Ada-SVM. We performed parameters optimization and feature selection simultaneously using PCA. In this Paper we have reduce the dimension of Features by features extraction.

REFERENCES

[1]. Almomani, Ammar, B. B. Gupta, Samer Atawneh, A. Meulenberg, and Eman Almomani. "A survey of phishing email filtering techniques." *Communications Surveys & Tutorials, IEEE* 15, no. 4 (2013): 2070-2090.

[2]. Silva, Renato Moraes, Akebo Yamakami, and Tiago A. Almeida. "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, 2012.

[3]. Abawajy, Jemal, Andrei Kelarev, and Morshed Chowdhury. "Automatic generation of meta classifiers with large levels for distributed computing and networking." *Journal of Networks* 9.9 (2014): 2259-2268.
[4]. Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar. "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. 2012.
[5]. Saberi, Alireza, Mojtaba Vahidi, and Behrouz Minaei Bidgoli. "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, 2007.
[6]. Gomez, Juan Carlos, and Marie-Francine Moens. "PCA document reconstruction for email classification." *Computational Statistics & Data Analysis* 56, no. 3 (2012): 741-751.
[7]. Ramanathan, Venkatesh, and Harry Wechsler. "phishGILLNET—phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training." *EURASIP Journal on Information Security* 2012, no. 1 (2012): 1-22.
[8]. Bhat, Sajid Yousuf, Muhammad Abulaish, and Abdulrahman A. Mirza. "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, 2014.
[9]. Trivedi, Shrawan Kumar, and Shubhamoy Dey. "Interplay between Probabilistic Classifiers and Boosting Algorithms for Detecting Complex Unsolicited Emails."*Journal of Advances in Computer Networks* 1, no. 2 (2013): 132-136.
[10]. en.wikipedia.org/wiki/Principal_component_analysis
[11]. en.wikipedia.org/wiki/AdaBoos
[12]. en.wikipedia.org/wiki/Support_vector_machine
[13].http://en.wikipedia.org/wiki/Radial_basis_function_kernel