"Spam Mail Classification Using SVM and Genetic Algorithm"

Neha Karadkar nehakaradkar2000@gmail.com
Akanksha Yeole akankshayeole14@gmail.com
Manasi Tilekar tilekarmanasi23@gmail.com

Zeal College of Engineering and Research Narhe, Pune

Abstract:

Feature selection is a problem of global combinatorial optimization in machine learning in which subsets of relevant features are selected to realize robust learning models. The inclusion of irrelevant and redundant features in the dataset can result in poor predictions and high computational overhead. Thus, selecting relevant feature subsets can help reduce the computational cost of feature measurement, speed up learning process and improve model interpretability. SVM classifier has proven inefficient in its inability to produce accurate classification results in the face of large e-mail dataset while it also consumes a lot of computational resources. In this study, a Genetic Algorithm-Support Vector Machine (GA-SVM) feature selection technique is developed to optimize the SVM classification parameters, the prediction accuracy and computation time. Spam assassin dataset was used to validate the performance of the proposed system. The hybrid GA-SVM showed remarkable improvements over SVM in terms of classification accuracy and computation time.

Keywords- Email classification, Feature Selection, Spam mail detection, Support Vector Machine.

I. Introduction

Most e-mail readers spend a non-trivial amount of time regularly deleting junk e-mail (spam) messages, even as an expanding volume of such e-mail occupies server storage space and consumes network bandwidth. An ongoing challenge, therefore, rests within the development and refinement of automatic classifiers that can distinguish legitimate e-mail from spam. Some published studies have examined spam detectors using Naïve Bayesian approaches and large feature sets of binary attributes that determine the existence of common keywords in spam, and many commercial applications also use Naïve Bayesian techniques. Email is one of the most popular forms of communication today. The surprisingly fast acceptance of this communication medium is best exemplified by the sheer number of current users, estimated to be as close to three quarters of a billion individuals and growing. This form of communication has the simple advantage of being almost instantaneous, intuitive to use, and costing virtually nothing per message. The current email system is based on the SMTP protocol RFC 821and 822 developed in 1982 and extended in RFC 2821 in 2001(. This system defines a common standard to unite the different messaging protocols in existence prior to 1982. It allowed users the ability to exchange messages with one another using a system based on the SMTP protocol and email addresses. These protocols allowed messages to flow from one user to another, making it practical and easy for different users to communicate independent of the service-provider or the client application.

As spam increased in volume and became more of a problem, anti-spam techniques were developed to counteract it. Tools to block spam were developed by a group of professionals. These tools were not always automated, but when used by system administrators of large sites, they could successfully filter spam for a large number of users. In response, spammers evolved their techniques to increase the number of spam delivered by working around and through the filters. As spam filters improved, spammers designed other methods of bypassing the filters and the cycle repeated. This resulted in the development of both spam and

anti-spam techniques and tools over a number of years. This evolutionary process continues today. Anti-spam tools use a wide range of techniques to reduce the volume of spam received by a user. A number of these techniques will be described in following section. There are several anti-spam techniques based on Open Source tool that we will examine in the light of the various techniques it uses to filter spam.

1.1 Motivation

Now a day's large number of people use email for communication and hence it has become the medium of target for hackers. In order to prevent such attacks, it is necessary to categorize the mails into two groups' viz. spam or not.

The problem with knowledge engineering method is that it requires constant updating of rules for classification which is very difficult. Over the last two decades, the application of Machine learning approach is increased due to various reasons like availability of large amount of data and the necessity of handling them in an efficient way.

1.2 Need

To develop a system that detects spam mails with maximum precision and with minimum processing time to help in the law enforcement fields.

II. Literature Survey:

Simran Gibson, Biju Issac et al. [1] stated that electronic mail has eased communication methods for many organisations as well as individuals. This method is exploited for fraudulent gain by spammers through sending unsolicited emails. This article aims to present a method for detection of spam emails with machine learning algorithms that are optimized with bio-inspired methods. A literature review is carried to explore the efficient methods applied on different datasets to achieve good results. An extensive research was done to implement machine learning models using Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree and Multi-Layer Perceptron on seven different email datasets, along with feature extraction and pre-processing. The bio-inspired algorithms like Particle Swarm Optimization and Genetic Algorithm were implemented to optimize the performance of classifiers. Multinomial Naïve Bayes with Genetic Algorithm performed the best overall. The comparison of our results with other machine learning and bio-inspired models to show the best suitable model is also discussed.

A. I. Taloba and S. S. I. Ismail [2] stated that the upsurge in the volume of unwanted emails called spam has created an intense requirement for the development of more dependable and robust anti-spam channels. Machine learning methods of ongoing are being used to successfully distinguish and channel spam emails. The present a systematic audit of some of the popular machine learning based email spam filtering approaches. The audit covers review of the important concepts, attempts, effectiveness, and the research pattern in spam filtering. The preliminary discussion in the investigation background examines the applications of machine learning methods to the email spam filtering cycle of the leading internet specialist organizations (ISPs) like Gmail, Yahoo and Outlook emails spam channels. Discussion on general email spam filtering measure, and the various efforts by different researchers in combating spam through the use machine learning procedures was done. Our survey compares the qualities and drawbacks of existing machine learning approaches and the open research problems in spam filtering. We recommended profound leaning and profound adversarial learning as the future methods that can adequately handle the menace of spam emails.

J. K. Agarwal and T. Kumar et al. [3] stated that communication through email has become one of the cheapest and easy ways for the official and business users because of easy availability of internet access. Most individuals like to use email to share important information and to maintain their official records. Be that as it may, just like the two sides of coin, many individuals misuse this easy way of communication by sending unwanted and useless mass emails to others. These unwanted emails are spam emails that affect the normal user to face the problems like exorbitant usage of their mailbox memory and filtration of useful email from unwanted useless emails. Thus, there is the need of some autonomous approach that channels the extreme data of emails in the form of spam emails. In this

paper, an integrated approach of machine learning based Naive Bayes (NB) algorithm and computational intelligence based Particle Swarm Optimization (PSO) is used for the email spam detection. Here, Naive Bayes algorithm is used for the learning and classification of email content as spam and non-spam. PSO has the stochastic distribution and swarm behavior property and considered for the global optimization of the parameters of NB approach. For experimentation, dataset of Ling spam dataset is considered and evaluated the performance in terms of precision, recall, f-measure and accuracy. Based on the evaluated results, PSO outperforms in comparison with individual NB approach.

W. Feng, J. Sun, L. Zhang et al. [4] proposed that Electronic mail has eased communication methods for many organizations as well as individuals. This method is exploited for fraudulent gain by spammers through sending unsolicited emails. This article aims to introduce a method for detection of spam emails with machine learning algorithms that are optimized with bio-inspired methods. A literature survey is carried to investigate the proficient methods applied on different datasets to achieve great results. A broad research was done to implement machine learning models using Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree and Multi-Layer Perceptron on seven different email datasets, along with feature extraction and pre-processing. The bio-inspired algorithms like Particle Swarm Optimization and Genetic Algorithm were implemented to optimize the performance of classifiers. Multinomial Naïve Bayes with Genetic Algorithm performed the best overall. The comparison of our results with other machine learning and bio-inspired models to show the best suitable model is also discussed.

A. Wijaya and A. Bisri [5] proposed that the Email spam is an increasing problem because it disrupting and time consuming for user, since the easy and cheap of sending email. Email Spam filtering can be done with a binary classification with machine learning as classifier. To date, email spam detection actually challenging since the email spam actually happens a great deal and the detection actually need improvement. Decision Tree (DT) is one of famous classifier since DT able to handle nominal and numerical attributes and increasing the effectiveness of computing. However, DT has a weakness in over-sensitivity to the training set and the noise data or instance that can degrade the performance. In this investigation, they propose half breed combination Logistic Regression (LR) and DT for email spam detection. LR is used for decrease noisy data or instance before data feed to DT induction. Noisy data reducing is done by LR by filtering right prediction with certain false negative edge. In this investigation, Spam base dataset is used to evaluate the proposed method. From the experiment, the outcome shows that proposed method yield impressive and promising outcome with the accuracy is 91.67%. It can be concluded that LR able to improve DT performance by reducing noisy data.

Abduelbaset M. Goweder, Tarik Rashed et al. [6] Stated that email is broadly becoming one of the fastest and most economical forms of communication .Thus, the email is prone to be misused. One such misuse is the posting of unsolicited, unwanted messages known as spam or garbage messages. This paper presents and discusses an implementation of an Anti-spam filtering system, which uses a Multi-Layer Perceptron (MLP) as a classifier and a Genetic Algorithm (GA) as a training algorithm. Standard hereditary operators and advanced strategies of GA algorithm are used to train the MLP. The implemented filtering system has achieved an accuracy of about 94% to recognize spam messages, and 89% to identify legitimate messages.

III Proposed Method and Algorithm:

1. Proposed Methodology:

In a proposed system, we are proposing experiment on detection of spam mails with limited set of supervised data.

We propose a new genetic as well as support vector machine based spam classification model for limited mails with higher accuracy.

The spam detection engine should be able to take email datasets as input and with the help of text mining and optimized supervised algorithms; it should be able to classify the email as ham or spam. Figure-1 represents the process that is followed to implement the model.

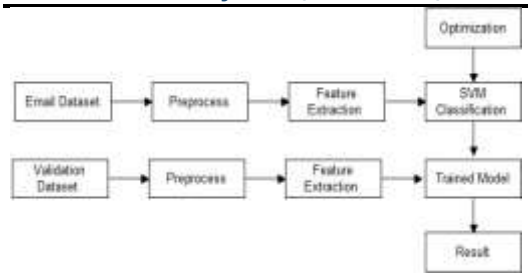

Figure1. Proposed Architecture

2. Algorithms

A. Pre-processing

I. Adding Corpus

This section will load all the email datasets within the program and distribute into training and testing data. This cycle will be accepting the datasets in '*.txt' format for individual email (Ham and Spam). This is to help understand the real-world issues and how might they be tackled.

II. TOKENIZATION

Tokenization is the method where the sentences within an email are broken into individual words (tokens). These tokens are saved into an array and used towards the testing data to identify the event of each word in an email. This will help the algorithms in predicting whether the email ought to be considered as spam or ham.

III. Stop Words Removal

This was used to remove the unnecessary words and characters within each email, and creates a bag of words for the algorithms to compare against. The module 'Check Vectorizer' from Scikit-learn assigns numbers to each word/token while counting and gives its event within an email. The instance is invoked to prohibit the English stopwords, and these are the words such as: A, In, The, Are, As, Is and so on, as they are not exceptionally useful to classify whether the email is spam or not. This instance is then fitted for the program to learn the vocabulary.

B. Classification

I. Genetic Algorithm

The GA algorithm is an evolutionary algorithm based on Darwinian natural selection that chooses the fittest individual from the given population. This involves the principle of variation, inheritance and selection. Detecting Spam Email With ML Optimized With Bio-Inspired Metaheuristic Algorithms (Chromosomes) that are binary addressed. The algorithm iterates through a fitness function where best individuals are chosen for reproduction of the offspring. The higher the fitness, the higher the probability.

II. Support Vector Machine

This algorithm plots each hub from a dataset within a dimensional plane and through classification strategy the cluster of data is separated by a hyperplane into their particular gatherings shows in equation 1.

$$H = Vx + c \qquad (1)$$

Where c is a constant and V is the vector. The SGD Classifier was loaded from scikit-learn library, which is the linear model with 'Stochastic Gradient Descent (SGD)', also known as the optimized version of SVM. This algorithm gives more accurate results than SVM (SVC algorithm) itself. Disadvantage of working with SVC algorithm is that it cannot handle a large dataset, whereas SGD gives productivity and other tuning opportunities. The algorithm uses the learning rate to iterate over the sample data to optimize the Linear algorithm and it is signified by the following equation-6 for the default learning rate as 'Optimal' showing in equation2.

$$\frac{1}{\alpha(t_0 + t)} \qquad (2)$$

Where t is the time step which is acquired by multiplying number of iterations with number of samples (Emails). The Learning Rate allows implementation of the parameter space during the training time. The α is addresses the regularization term and t0 is a hueristic approach.

IV. Mathematical Model

Let us consider S as a system for automatically recommends vehicle to customer. S={F,I,O,e, Φ}}

INPUT:
Identify the inputs F= f1, f2, f3 ....., fn— F as set of functions to execute commands.
I= i1, i2, i3 Sets of inputs to the function set
O= o1, o2, o3 Set of outputs from the function sets,
e = End of the program.
S1= I, F, O
I = Query submitted by the User, i.e. query mail
O = Output of desired query, i.e. Spam mail stage prediction
F = Functions implemented to get the output, i.e. genetic and svm for classification of data.

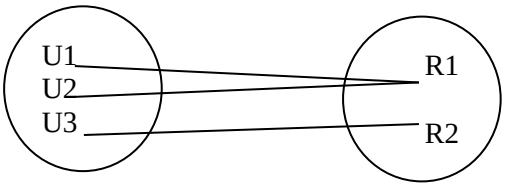

Figure2 Mapping Diagram

Where,
U=users
R=Cancer stage.
U1=Search image1 for get right spam type
U2= Search image2 for get right spam type
U3= Search image1 for get right spam type
R1= Result of right spam mail stage.
R2= Result of wrong spam mail stage.

Set Theory

S={s, e, X, Y,*Φ*}

Where,

s = Start of the program.

1. User Login with Credential.

2. User inserts the mail for checking spam mail or not.

3. System predict the stage of spam mail or not.

e = End of the program.

V. Conclusion:

We propose a, two classifiers, SVM and GA-SVM were tested to filter spams from the spam assassin dataset of emails. All the emails were classified as spam (1) or legitimate (-1). GA is applied to optimize the feature subset selection and classification parameters for SVM classifier. It eliminates the redundant and irrelevant features in the dataset, and thus reduces the feature vector dimensionality drastically. This helps SVM to select optimal feature subset from the resulting feature subset. The resultant system is called GA-SVM. GA-SVM achieves higher recognition rate using only few feature subset.

References:

1. Simran Gibson, Biju Issac "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms" Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. date of publication October 13, 2020.

2. A. I. Taloba and S. S. I. Ismail, ''An intelligent hybrid technique of decision tree and genetic algorithm for E-Mail spam detection,'' in Proc. 9th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS), Cairo, Egypt, Dec. 2019, pp. 99–104, doi: 10.1109/ICICIS46948.2019.9014756.

3. J. K. Agarwal and T. Kumar, ''Email spam detection using integrated approach of Naïve Bayes and particle swarm optimization,'' in Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS), Jun. 2018, pp. 685–690, doi: 10.1109/ICCONS.2018.8662957.

4. W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, ''A support vector machine based Naive Bayes algorithm for spam filtering,'' in Proc. IEEE 35th Int. Perform. Comput. Commun. Conf. (IPCCC), Dec. 2016, pp. 1–8, doi: 10.1109/pccc.2016.7820655.

5. A. Wijaya and A. Bisri, ''Hybrid decision tree and logistic regression classifier for email spam detection,'' in Proc. 8th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE), Oct. 2016, pp. 1–4, doi: 10.1109/ICITEED.2016. 7863267.

6. M. Goweder, Tarik Rashed , Ali S. Elbekaie, and Husien A. Alhammi, "An anti-spam system using artificial neural networks and genetic algorithms", Proc. Int. Arab Conf. on Information Technology, 2008.