ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 2, March 2016

Combined Techniques For Detecting Email Spam Using Negative Selection and Particle Swarm Optimization

Dr. C. Palanisamy,
Dept of IT,
Bannari Amman Institute of Technology,
Sathyamangalam, India
cp_samy@yahoo.com

Mr. T. Kumaresan,
Dept of CSE,
Bannari Amman Institute of Technology,
Sathyamangalam, India
kumaresant@bitsathy.ac.in

S.E.Varalakshmi
Dept of CSE,
Bannari Amman Institute of Technology,
Sathyamangalam, India
rohiniebd@gmail.com

***Abstract*---Email is a communication throughout the entire world today. The email spam is both text and images requires for the media flow. Email spam is a subset of electronic spam. Involves nearly identical messages sent to numerous recipients by email. In this work introduces an email detection system based on an improved negative selection algorithm and particle swarm optimization (PSO).This implementation improves random detector generation in the negative selection algorithm (NSA). The combined NSA-PSO uses a local outlier factor (LOF) as the fitness function for the detector generation. The detector generation is terminated when the spam coverage is reached. A distance measure and a threshold value are employed to enhance the distinctiveness between non-spam and spam detectors. Our experimental dataset using Ling-Spam. The results showed that the accuracy of the proposed NSA-PSO model is better than the standard NSA model. The proposed model with the best accuracy is used to differentiate between spam and non-spam.**

***Keywords*---Negative selection algorithm, Particle swarm optimization, Spam detectors**

1. INTRODUCTION

Email spam includes malware as scripts or different possible file attachments. Definitions of spam sometimes embody the aspects that email is uninvited and sent in bulk. Thus Email spam detection techniques are used to detect the spam. NSA-PSO is used to detect the spam email with high accuracy in comparison with other optimization methods.

Classification can be used for predicting the class label of data objects. It is the process of finding a set of models (or functions) Which describe data classes or concepts for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model is also being represented in varied forms, like classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

The process of detector generation is select the random detector generation of the real valued negative selection algorithm for classifying non-spam and spam. Detectors that do not match of the non-spam set are accepted as viable detectors. Detectors that match of the non-spam set are discarded as unwanted detectors. The generation of detectors continues until the detector set reached the coverage in the spam space and then monitor the status of the system. The Euclidean distance is used as the matching measurement[6].

The Euclidean distance is defined as

$$L(x_i, dj) = \sqrt{(x_{i1} - c_{j1})\dot{2} + \cdots + (x_{in} - c_{jn})\dot{2}} \qquad (1)$$

The negative selection mainly focuses on anomaly detection, fault detection, and intrusion detection. Most work on negative selection and particle swarm optimization solves the problems of anomaly detection and intrusion detection[7]. The implementation of particle swarm optimization with negative selection to maximize the coverage of the non-self space[12].

The artificial immune system based on the mammalian immune system[2]. The main goal of the immune system is to distinguish between non-self and self-element. This research will replace self as non-spam in our system and non-self as spam in our system[5].

II. RELATED WORKS

Agrawal et al (2005) proposed a Common Vector Approach (CVA) for e-mail detection. Electronic mail is an important communication method for most computer users. Spam e-mails However consume bandwidth resource, fill-up server

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 2, March 2016

storage and are also a waste of time to tackle. The general way to label an e-mail as spam or non-spam is to set up a finite set of discriminative features and use a classifier for the detection. In most cases, the selection of such features is empirically verified. There are two different methods are proposed to select the most discriminative features among a set of reasonably arbitrary features for spam e-mail detection. The selection methods are developed using CVA which is actually a subspace-based pattern classifier[4].

Oda et al (2005) proposed the paper, Immunity from spam: An Analysis of an Artificial Immune System for Junk Email Detection proposed a random generation of detectors does not help in solving the problem of finding the best selected features. And the feature weights are updated during the matching process. The weighting of the features complicates the performance of the matching process. The experimental results performed with use of a spam-assassin corpus and a Bayesian combination of the detector weights. The immune system classifies correctly 90% of the messages. More specifically, it classifies 84% of the spam and 98% of the non-spam. The approach of scoring features or feature weighting during and after the matching process creates ambiguity in the selection of important features for spam detection due to its computational cost.

Sirisanyalak and Sornil (2007) proposed An artificial immunity-based spam detection system. The analysis of major work performed on negative selection algorithms with a combination of two different algorithms in a hybrid email spam model. An AIS based module that extracts features was designed and further used for a logistic regression model. the set of detectors was initially generated using terms that were extracted from the training message and using data from matched detectors that were used in the regression model. The experiment uses spam-assassin. The detection performance of the system is 0.91% and 1.95% of false positive and false negative rates.

Wamli et al (2009) proposed A Novel Spam Email Detection System Based on Negative Selection this studies the possibility of using negative selection in email spam detection without prior information of the email spam. The negative selection algorithm is divided into four concurrent working modules with two repositories: the random detector generation module, the detector maturing module, the antigen matching module and the detector aging module, with a selves repository and a detectors repository. The TREC07 corpus was used in its implementation. After the initial 1/3 of the time during the learning period, the spam detection rate is over 80%, and it is over 70% most of the time.

Hamdan et al (2011) proposed Application of genetic optimized artificial immune system and neural networks in spam detection. A genetic optimized AIS culled old lymphocytes (replacing the old lymphocytes with new ones) and also checked for new interests for users, using an approach .To update intervals such as the number of received messages. An interval is updated with respect to time, user requests and other factors. Many choices were used in selecting the update intervals using the genetic algorithm. The experiment was implemented with a spam-assassin corpus that had 4147 non-spam messages and 1764 spam messages. The optimized spam detector with 600 generated detectors gives a false positive rate of 1.1% and a false negative rate of 3.7%, while spam detection with AIS and 600 generated detectors gives a false positive rate of 1.2% and a false negative rate of 4.9%.

IsmailaIdris and Ali Selamat (2014) projected an improved email spam detection model with negative choice algorithmic rule and particle swarm improvement. To address the trend of email spam, a unique model that improves the random generation of a detector in negative choice algorithmic rule (NSA) with the utilization of random distribution to model the info purpose victimization particle swarm improvement (PSO) was enforced. Distance live is utilized to boost the distinctiveness between the non spam and spam candidate detector. The detector generation method was terminated once the expected spam coverage is reached. The projected improved model is a higher replacement to United States intelligence agency model. Performance and accuracy investigation has shown that the projected improved model is ready to find email spam higher than the United States intelligence agency and PSO model.

III. PROPOSED SYSTEM

A combined negative selection algorithm and particle swarm optimization[1] was complimenting the parameters of each component of the system. This approach uses the advantages of the individual systems to achieve stability, consistency and accurate, intelligent systems that are usage in classification. The strength of the particle swarm optimization is combined with the negative selection algorithm to improve the weaknesses of both algorithms. A local outlier factor is also implemented as a fitness function for the particle swarm optimization[14].

The particles are composed of 57 features. The constant C value is 0.5. The position and velocity of the particle swarm optimization are represented in N-dimensional vector space by

$$P_i (P_i^1, P_i^2 \ldots P_i^n) \quad (2)$$

$$V_i (V_i^1, V_i^2 \ldots V_i^n) \quad (3)$$

Here, p_{id} denotes the binary bits i =1,2,...m (m is the total number of particles) and d=1,2,...n (n is the dimensionality of the data). Each particle in the generation updates its own position and velocity according to (2) and (3). The

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 2, March 2016

initialization of the real-valued PSO is established by the population of particles non-spam and spam[3]. All of the particles move in problem space to find the optimal solution over all of the iterations.

In PSO, for each iteration a fitness function is evaluated for all the particles in the swarm. The velocity of every particle was updated by keeping track of the two best positions. One is the best position a particle had traversed so far and called ''pBest''[11]. The second is the best position that any neighbour of a particle had traversed so far. It is a neighbourhood best called ''nBest''. When a particle took the whole population as its neighbourhood, the neighbourhood at best became the global best and was accordingly called ''gBest''. Hence, a particle's velocity and position were updated as follows

$$V_{i,d}(t+1) = v_{id}(t) + c(P_{id}(t) - x_{i,d}(t)) \quad (4)$$

$$x_{i,d}(t+1) = x_{i,d}(t) + v_{i,d}(t+1) \quad (5)$$

L ($x_i$dj) is compared with the non-spam space threshold Rs, which generates the match value $\propto$:

$$\propto = L(X_i d_j - R_s) \quad (6)$$

The detector d_j fails to match the non-spam sample X_i if $\propto > 0$. if d_j does not match any non-spam sample, it will be retained in the detector set. The detector threshold R^d, j of detector d_j can be defined as

$$R^d, j = \min(\propto), \text{if} \propto \leq 0 \quad (7)$$

If detector d_j matches the non-spam sample, it will be discarded. This process will not stop until a detector set that the desired spam space coverage is reached. The generated detector set can then be used to monitor the entire system.

The local outlier factor (LOF) was employed to calculate the fitness function in a quest for purely normal data that will efficiently train our model. The outlier detection algorithm proposed as a fitness function of spam detection generation is very unique in computing the full dimensional distance from one point to another.[8]

$$\text{Ird}(i) = 1/\left(\frac{\sum s \in N_k(i) reachablity - distance_k(i,s)}{|N_k(i)|}\right) \quad (8)$$

Equation (8) is the average reach-ability distance of the candidate detector i from the non-spam element. This value is not the average reach-ability of the neighbour from i.

$$LOF_k(i) = \frac{\sum s \in N_k(i)(Ird_{(s)})/(Ird(i))}{|N_k(i)|} = \left(\frac{\sum s \in N_k(i) Ird(s)}{|N_k(i)|}\right) / \text{Ird}(i) \quad (9)$$

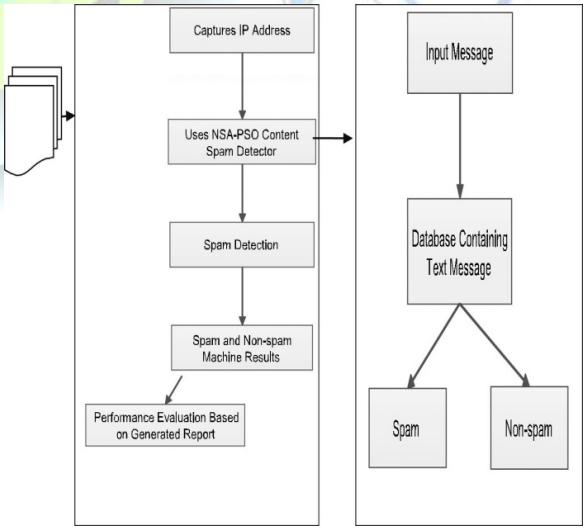
Equation (9) shows the average local reach-ability density of the neighbour divided by the local reach-ability density of the particle[1].

The complete flowchart of this proposed algorithm was detailed with the corresponding pseudo codes listed as follows:

a) Step 1: First initialize the non-spam space (normal pattern).
b) Step 2: From the population of spam and non-spam data, generate training and testing set with random candidate detectors.
c) Step 3: Using Equation (2) and (3) to initialize both the position and the velocity of the PSO.
d) Step 4: Calculate the reach-ability distance and the LOF for each candidate detector, as shown in Equation (8) and (9).
e) Step 5: Update each candidate detector position and velocity with Equation (4) and (5).
f) Step 6: Calculate the distance measure with Eq. (6) And the threshold value with Equation (7) to determine the P_{best} in the non-spam space S. If p_{best} does not match S, then it is a valid detector.
g) Step 7: Continuously generate and after the maximum coverage in spam space.

IV. SPAM DETECTION PROCESS

The spam detector process is presented in Figure 1. The NSA–PSO spam content detector takes input messages and verifies them with the files that are present in the database. The verification is performed message after message to determine the content of spam by the calculation of the different probabilities of spam occurrences. The flow is divided into three different modules. The modules are the master server, the client module, and the spam report module [1].



ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 2, March 2016

Fig 1.Spam Detection Process

The master server is used to keep track of client and spam details by the network administrator. Christo Ananth et al. [9] discussed about Improved Particle Swarm Optimization. The fuzzy filter based on particle swarm optimization is used to remove the high density image impulse noise, which occur during the transmission, data acquisition and processing. The proposed system has a fuzzy filter which has the parallel fuzzy inference mechanism, fuzzy mean process, and a fuzzy composition process. In particular, by using no-reference Q metric, the particle swarm optimization learning is sufficient to optimize the parameter necessitated by the particle swarm optimization based fuzzy filter, therefore the proposed fuzzy filter can cope with particle situation where the assumption of existence of "ground-truth" reference does not hold. The merging of the particle swarm optimization with the fuzzy filter helps to build an auto tuning mechanism for the fuzzy filter without any prior knowledge regarding the noise and the true image. The system identifies the compromised and non-compromised machines in the network.

The system different records of spam machines and non-spam machines and other records that contain the client's name, spam details, date, IP address and time stamps based on the NSA–PSO model, the spam is detected[13].

V. PERFORMANCE ANALYSIS

The proposed NSA-PSO model was compared to the spam detection algorithm such as, SVM, NB, DFS-SVM. The kernel of SVM was chosen as a Gaussian Radial Basis function with r = 1. The parameters of SVM were trained by quadratic programming method. The results are shown in Table 1. It showed that the accuracy of SVM is 90%.

TABLE 1
COMPARISON FOR NSA, PSO, NSA-PSO AND OTHER MODELS

CLASSIFIER	ACCURACY
SVM	90%
NB	78.8%
DFS-SVM	71%
NSA	80.86%
NSA-PSO	93.20%

The difference in the performance between the proposed NSA-PSO model and the NSA model are very significant. The best accuracy of the proposed model is 93.20%, while for the NSA model is 80.86%. Moreover, the NSA-PSO performed better than the four algorithms with an accuracy of 93.20%.

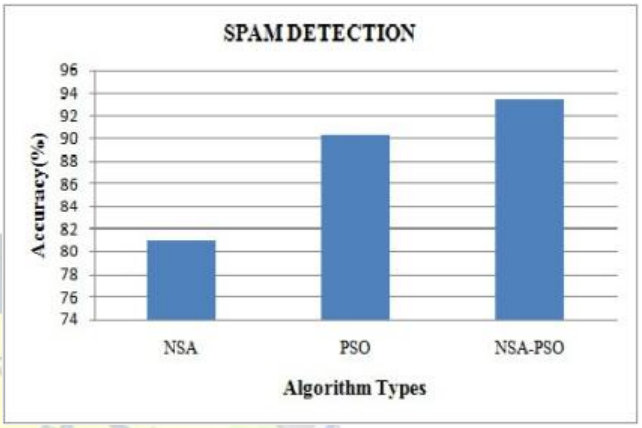


Fig 2.Accuracy report

VI CONCLUTION

The main contribution and technical innovation of this experiment falls within the following four points: feature selection method is used that can achieve high generality with a naive Bayesian classifier. The distance measure in which Euclidean distance is used as the matching measurement and use combined negative selection algorithm and particle swarm optimization (NSA-PSO) model can detect email spam. This proved that the feature selection results by NSA-PSO are better than those of SVM, NB, DFS-SVM, and NSA. This work can be enhanced to improve the computational efficiency using better optimization technique.

ACKNOWLEDGEMENT

I humbly submit all the glory and thanks to almighty for showering the blessing upon me and give me the necessary wisdom for accomplishing this project. I take immense pleasure to thank my guide Mr T. Kumaresan, Assistant Professor (senior grade), Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam for his guidance to complete this project successfully. I would like to enunciate special thanks to my friends, teaching and non teaching staff who have directly and indirectly contributed to the success of this project. Last, but not least I take it a great privilege to express my deep sense of gratitude to my beloved parents and relations for their blessings and support.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 2, March 2016

REFERENCES

[1] Ali Selamat, Ismaila Idris, Ngoc Thanh Nguyen, Sigeru Omatu , Ondrej Krejcar , Kamil Kuca and Marek Penhaker, 2015, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system", *Journal of Elsevier,* Vol.39,pp.33-44.

[2] Abi-Haidar A., and Rocha L, "Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics", Artificial Immune Systems, *Jounal of Springer*, Vol.5132, pp.36-47, 2008.

[3] Oda T, and White T, "Developing an Immunity to Spam", *Journal of Springer*, Vol.2723, pp. 231–242, 2003a.

[4] Dasgupta D, Yu S, and Nino F, "Recent advances in artificial immune systems: models and applications", *Journal of Application Software Computing*, Vol.11, no.2, pp. 1574–1587, 2011.

[5] Hamdan, Mohammad Adel, and Abu, Z.R., "Application of genetic optimized artificial immune system and neural networks in spam detection", *Application Software Computing*, Vol.11, no.4, pp. 3827–3845, 2011.

[6] Oda T, and White T, "Immunity from spam: an analysis of an artificial immune system for junk email detection". *Journal of Springer*, Vol.3627, pp. 276–289, 2005.

[7] Ramaswamy S, Rastogi R., and Shim K, "Efficient algorithms for mining outliers from large data sets", vol. 29, no.2 , pp. 427–438, 2000.

[8] Sajesh T.A, and Srinivasan M.R., "Outlier detection for high dimensional data using the Comedian approach", *Journal of Statistical Computation and Simulation*, Vol.82,no.5,pp. 745–757, 2011.

[9] Christo Ananth, Vivek.T, Selvakumar.S., Sakthi Kannan.S., Sankara Narayanan.D, "Impulse Noise Removal using Improved Particle Swarm Optimization", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 3, Issue 4, April 2014,pp 366-370

[10] Zhang L, Zhu J, and Yao T, "An evaluation of statistical spam filtering techniques", vol.3,no.4 ,pp. 243–269, 2004.

[11] Wang H, X. Z. Gao, Xianlin Huang, and Zhuoyue Song, "PSO-optimized negative selection algorithm for anomaly detection", *Journal of Springer*, Vol.52, pp. 13–21, 2009.

[12] Yevseyeva I, Vitor Basto-Fernandes and David Ruano-Ordas, "Optimising anti-spam filters with evolutionary algorithms", J*ournal of Elsevier*, Vol.40, no.10,pp. 4010–402, 2013.

[13] Ling-Spam dateset: http://csmining.org/index.php/ling-spam-datasets.html.

[14] Ma, W., D. Tran, and D. Sharma, "Negative Selection with Antigen Feedback in Intrusion Detection", *International Conference on Artificial Immune Systems*, 2008.