ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

Decentralized Secure Service Discovery of Resources Detection System

M.Sudha, R.P. Sathiya Priya,

Assistant Professor, Department of Computer Science, Muthayammal College of Arts & science, Rasipuram, Namakkal DT, India

Department of Computer Science, Muthayammal College of Arts & science, Rasipuram, Namakkal DT, India

ABSTRACT: Nowadays, email has the efficient become communication medium for exchange of information between people in organizations and the society throughout the world. With the increased email spam attacks the efficiency of communication and security is at stake. Spam emails are adaptive in nature as the spammers constantly update the spam contents that could bypass the spam filters. Negative Selection Algorithm is a one class classifier algorithm that is employed for email classification and here the classification is between Spam and Non-Spam which is one class classification. This research introduces an email spam detection system that shows an improvement in the performance of Negative Selection Algorithm. The performance can be improved by hybridizing this algorithm with additional two concepts which addresses the adaptive nature of unsolicited email spam. Hybridization is done by employing Firefly Algorithm in the Random detector generation step where Local Outlier Factor is used as the fitness function. The hybridization of Firefly Algorithm helps in selecting subset of features that will be used for the monitoring phase. The result shows higher accuracy than any other optimization algorithms for email spam detection systems.

KEYWORDS: Spam Detection , Negative Selection Algorithm , Firefly Algorithm , Feature Selection , Local Outlier Factor

I. INTRODUCTION

COMMUNICATION means exchange of information by speech, text, gestures and images. The invention of electronic communication devices and internet brought emails into existence. Realizing the speed and ease of electronic communication government bodies, organizations and the public started using emails for communication. Spamming of e-mails is an issue that causes several problems to the users and has become a serious threat for communication. The major aim of any spammer is to make the server busy by forcing the server with unwanted messages or unwanted searches or to destroy the information stored in storage medium of a particular user or a group of systems of an organization. Spams are of many types like unsolicited advertisements, phishing scams, Nigerian 419 scams, e-mail spoofing, Trojan horse e-mail, commercial advertisements, anti-virus spam, chain letters, porn spam, terrorist spam, etc. There are several spam techniques that have been used by spammers for sending spam mails to the targeted recipients like Appending, Image spamming, Backscattering. The basic two approaches for handling e-mail spam are Knowledge engineering based approach where a continuous maintenance of the database is needed and the Machine Learning (ML) based approach where many ML techniques were incorporated for classifying spam and non-spam entities. Spam mails are adaptive in nature as the spammers periodically update the contents that would bypass the spam filters.

Hence an efficient algorithm which will be helpful in mitigating the adaptive nature of the unsolicited mail is to be developed. In this work, ML based approach is used. Negative Selection Algorithm (NSA) is a one class classification algorithm where classification is done between two entities (i.e.) spam and non-spam. NSA is an algorithm where both training and monitoring phase are included. To increase the detection rate Firefly Algorithm (FA) and Local Outlier Factor (LOF) algorithms are used.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

The paper is organized as follows: chapter II presents literatures reviewed. Chapter III deals with the NSA, FA and LOF, Hybridized system design. Chapter IV discusses evaluation parameters and Chapter V describes the conclusion and future work.

II. LITERATURE REVIEW

Gebali et al [4] proposed a technique for the detection of spam by per-packet based classification of e-mails. In this approach, the e-mails are classified with respect to their contents as the classification is done using the packets. A middlebox used to control spam. The classification is done using naïve Bayes classifier. Both pre-classification support and fast e-mail class estimation support are used on per-packet estimation. The entire system is implemented in the packet level (i.e.) at layer 3.

Guzella T S and Caminhas W M [5] have proposed an extensive review of modern developments in the application of ML algorithms for filtering of spam e-mails by considering both text and image in the incoming e-mails. In this method instead of considering the spam detection as classification problem special characteristics like drift is considered in designing filters. This work particularly introduced two approaches which were not usually considered, they are the process of updating a classifier based on the representation of bag-of-words and naïve Bayes models. Ying et al [7] proposed a supervised learning algorithm namely ensemble approach based on decision tree, Support Vector Machine (SVM) and back propagation network is employed to distinguish e-mails between spam and non-spam. The main objective of this work is to reduce the problem in information security. This paper also has concerns about the wastage of network resources. The characteristics of e-mails are used in this approach to classify them. The proposed spam e-mails are classified into 14 different features on which ensemble approach is performed.

Dasgupta (2006) [1] proposed a survey on AIS and specifically as a branch Computational Intelligence (CI) was shown a progress. Several immunological principles has been adopted and based on those principles several CI models had been surveyed. Like AIS, analysis by several researchers to understand the mechanism of several Biological Immune System (BIS) based models were analyzed to find the natural processes and their dynamic behaviors were studied and reviewed with the existence of pathogens or antigens. Survey on several Biologically-Inspired techniques such as Genetic Algorithm (GA), Artificial Neural Network (ANN) and Cellular Automata were surveyed by the author. Several other specific AIS based approaches such as Immune Network Model, NSA and Clonal Selection Principle were also surveyed. The concepts were proposed to be used in many science and engineering problems to provide Computer Security and in many Fault Detection systems.

Selamat et al (2015) [14] proposed an idea in enhancing the NSA with PSO for detector generation with LOF as its fitness function for email classification. This approach showed an accuracy of 83.20%. But NSA without PSO showed 68.86% accuracy. Both the results show that the selection of feature selection algorithm influences the accuracy.

Therefore the problems identified in constructing spam detection system are 1) The adaptive nature of the spam mails are not handled 2) ML algorithms lack in efficient feature selection.

III. NSA-FA: A HYBRID MACHINE LEARNING ALGORITHM

A. NSA

NSA is an algorithm that imitates or simulates the process of negative selection in Immune system. Immune system uses antibodies to find and negate microbes such as bacteria and viruses. Antigens are substances that make an immune system to produce antibodies against them. An antigen may be produced within the body namely self-antigen or from the external environment called as non-self. Antibodies react with the external or internal substances called antigens and the entire combined substance is removed from body. Antibodies are the detectors which perform the action of finding the antigens and help in removing them from the body to protect human body.

This process of negative selection is employed for spam classification where the antigens correspond to spam mails and antibodies or the detector cells correspond to the non-spam mails. The selected detector cells by feature selection step of NSA helps in finding and removing spam contents from the system.

But, there is a serious problem in this algorithm. The problem is that the antibodies may also react with the self cells causing serious damage to the immune system. To avoid such serious damage negative selection is employed by

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

removing the immature antibodies for the system. Antibodies that don't bind with the self cells are alone allowed to circulate throughout the body. Thus, the negative selection ensures that the antibodies that don't bind with self cells are allowed to circulate. The pseudocode of the NSA is Algorithm 1.

There are three steps in this algorithm. They are

a. Defining Self Elements
b. Detector Generation
c. Monitoring

Algorithm 1 NSA

Input: a self-set, a monitor set, an integer n
Output: Either self or non-self
//training phase
1: d ⟵~~empty~~ set
2: **while** |D| < n **do**
3: d ⟵~~random~~ detector
4: **if** d does not match any element of S **then**
5: insert d into D
//classification phase
6: **for each** m ∈ m **do**
7: **if** m matches any detector d ∈ D then
8: **output** " m is non-self "(an anomaly)
9: **else**
10: **output** " m is self "

This algorithm fails due to the inefficient feature selection techniques used in the Detector Generation step of NSA. Hence, to increase the feature selection, one of the efficient feature selection techniques, Firefly Algorithm is used.

B. Defining Self Elements

In human immune system Antibodies are called the self elements and Antigens are called the Non-Self. Here in the email spam system, Non spam elements are the Self elements and the spam are the non-self elements.
The non-spam space is considered to be S and is defined in (1).

$$S_{ij}=\begin{pmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mn} \end{pmatrix} \qquad (1)$$

C. Feature selecion in NSA

FA is a metaheuristic optimization algorithm that imitates the flashing behavior of fireflies. The flash of one firefly attracts the other fireflies. The brightness is associated with the objective function. Some the assumptions made are as follows

a. All fireflies are unisexual
b. Attractiveness is proportional to brightness
c. If no fireflies brighter than a given firefly, it will move randomly

The pseudocode of FA is Algorithm 2.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

Algorithm 2 Firefly Algorithm

Begin
1. Objective function: $f(x)$,
$x = (x_1, x_2, ..., x_d)$;
2. Generate an initial population of fireflies
$x_i (i = 1, 2, .., n)$;
3. Formulate light intensity I so that it is associated with $f(x)$ (for example, for maximization problems, $I \propto \exp(-\gamma r)$; or simply $I = f(x)$;
4. Define absorption coefficient γ

While (t < MaxGeneration)
for i = 1 : n (all n fireflies)
for j = 1 : n (n fireflies)
if ($I_j > I_i$),
move firefly i towards j;
Vary attractiveness with distance r via $\exp(-\gamma r)$;
Evaluate new solutions and update light intensity;
end if
end for j
end for i
rank fireflies and find the current best;
end while
post-processing the results and visualization;
end

D. Computation of Fitness Function

LOF is used to calculate the fitness function for training the FA. An outlier can be defined as a data point that is not the same as others in a population with respect to certain measure. The pseudocode of LOF is Algorithm 3.

E. MONITORING

The monitoring consists of two processes. They are
a. Matching Process
b. Monitoring Phase

a. Matching Process

In order to keep a sufficiently small set of detectors and make sure a relatively constant size of it with the increase of protected string, exact non-matching cannot be adopted. The matching process is depicted in Figure 1.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

Algorithm 3 Local Outlier Factor

Input: G // Random particle population
S // Non-Spam space
i // i^{th} particle in G
Output: The degree of the local outlier factor for all records of the i^{th} particle

1. ***Begin***
2. *Population of random dataset G*
3. *Reach-distance : k = dist (G,i)*
4. ***For each i in G do begin***
5. *Reach-distance i = max (dist(s,i))*
6. *|G|' (Minpt(s,G))*
7. *Max-dist(i) ∈ (G)*
8. *Find population G with maximum reach-ability distance to s*
9. *Max(dist(i,s))*
10. *Find population G with maximum similarity with i*
11. ***End***
12. ***Return*** *$|G_{max}|$' similarity (I,G_{max})*
13. ***End***

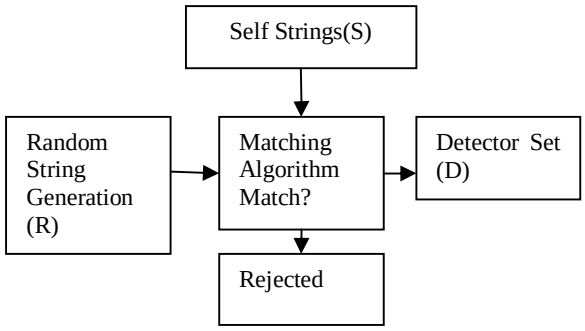

Figure1. Matching Process

b. Monitoring Algorithm

Two equal length strings match if they are equal in r contiguous position. The monitoring algorithm is depicted in Figure 2.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

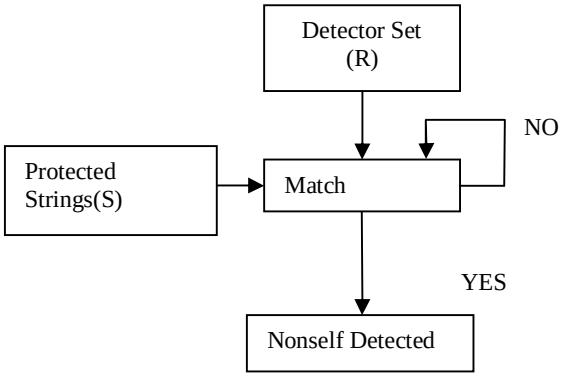

Figure2. Monitoring Phase

The hybridized system design is Figure 3.

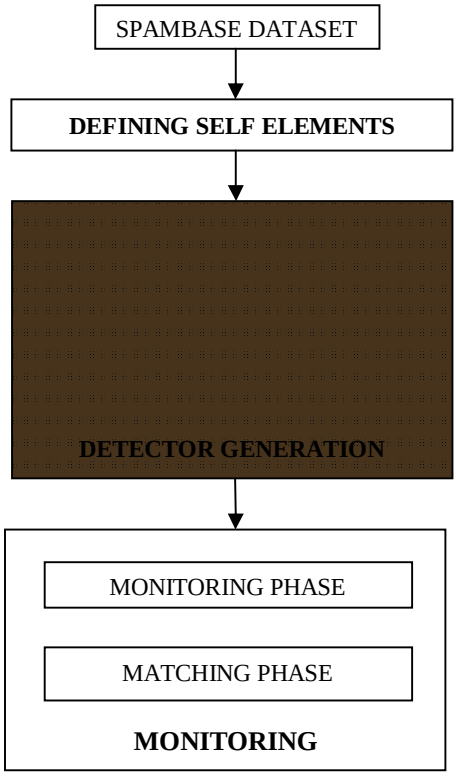

Figure3. Hybridized System Design

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

IV. PARAMETER EVALUATION

The common way for computing accuracy is based on the confusion matrix. Table 1 shows the confusion matrix.

Table1. Confusion Matrix

	Predicted Positives	Predicted Negatives
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

The formula for calculating the accuracy is shown in (2).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (2)$$

NSA-PSO shows 83.20% [14] accuracy. The performance of FA in feature selection is higher than that of PSO. Hence eventually the performance of NSA-FA is higher than that of the NSA-PSO.

V. CONCLUSION AND FUTURE WORK

The proposed work NSA-FA shows an increased accuracy when compared with the previous work NSA-PSO. The main drawback of NSA-PSO was the inefficiency of the feature selection algorithm. It has been overcome with FA which is an efficient feature selection algorithm. The major advantages of NSA-FA are 1) the adaptive nature of unsolicited emails has been addressed 2) Effective Features have been selected for monitoring.

In future, for even more effective feature selection, optimization algorithms like cuttlefish, bat can be used.

REFERENCES

[1] Dasgupta D., (2006), 'Advances in artificial immune systems', Computational Intelligence Magazine', IEEE, Vol. 1 No. 4 pp. 40–49.
[2] Bereta M., Burczynski T., (2007), 'Comparing binary and real-valued coding in hybrid immune algorithm for feature selection and classification of ECG signals', Engineering
[3] Applications of Artificial Intelligence, Elsevier, Vol. 20 No. 5 pp. 571–585.
[4] Marsono N M., El-Kharashi W., Gebali F., (2009), 'Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification', Computer Networks, Elsevier, Vol. 53 No. 6 pp. 835-848.
[5] Guzella T S., Caminhas W M., (2009), 'A review of machine learning approaches to spam filtering' Expert Systems with Applications, Elsevier, Vol. 36 No.7 pp. 10206–10222.
[6] Amayri O., Bouguila N., (2010), 'A study of spam filtering using support vector machines', Artificial Intelligence Review, Springer, Vol. 34 No. 1 pp. 73-108.
[7] Ying K C., Lin S W., Lee Z J., (2010), 'An ensemble approach applied to classify spam e-mails', Expert Systems with Applications, Elsevier, Vol. 37 No. 3 pp. 2197-2201.
[8] Dasgupta D., Yu S., Nino F., (2011), 'Recent advances in artificial immune systems: models and applications', Applied Soft Computing, Elsevier, Vol. 11 No. 2 pp. 1574–1587.
[9] Abaei G., Selamat A., (2014), 'A survey on software fault detection based on different prediction approaches', Computer Science, Springer, Vol. 1 No. 2 pp. 79-95.
[10] Idris I., Selamat A., Omatu S., (2014), 'Hybrid email spam detection model with negative selection algorithm and differential evolution', Engineering Applications of Artificial Intelligence, Elsevier, Vol. 28 pp. 97-110.
[11] Li D., Liu S., Zhang H., (2015), 'A negative selection algorithm with online adaptive learning under small samples for anomaly detection', Neurocomputing, Elsevier, Vol. 149 pp. 515-525.
[14] Selamat A., et al., (2015), 'A combined negative selection algorithm – particle swarm optimization for an email spam detection system', Engineering Applications of Artificial Intelligence, Elsevier, Vol. 39 pp. 33-44.