E-mail Spam Classification With Artificial Neural Network and Negative Selection Algorithm

Ismaila Idris

Dept of Cyber Security Science, Federal University of Technology, Minna, Nigeria.

Idris.ismaila95@gmail.com

Abstract

This paper apply neural network and spam model based on Negative selection algorithm for solving complex problems in spam detection. This is achieved by distinguishing spam from non-spam (self from non-self). We propose an optimized technique for e-mail classification; The e-mail are classified as self and non-self whose redundancy was removed from the detector set in the previous research to generate a self and non-self detector memory. A vector with an array of two element self and non-self concentration vector are generated into a feature vector used as an input in neural network classifier to classify the self and non-self feature vector of self and non-self program. The hybridization of both neural network and our previous model will further enhance our spam detector by improving the false rate and also enable the two different detectors to have a uniform platform for effective performance rate.

1. INTRODUCTION

The traditional way of detecting spam based on signature is no more efficient for today systems. Recent years, Researchers are interested in the field of immune system in achieving computer security. The function of computer security systems are meant to recognize and discard spam. Data mining, machine learning and the signature base techniques are proposed for spam detection Christodorescu et al [2]. The signature base spam detection technique is not too reliable in detecting new spam since the number of spam grows concurrently in such a way that the signature based spam detection technique cannot meet up with its security challenges, putting in mind the increase signature database or the time it will take before matching takes place in signatures. The data mining techniques keep in memory specific bytes sequence obtain in the file content and also monitor the behavior of suspicious program. Our proposed technique has the capability to detect formally unknown spam. There is a relationship between any detector selected to one feature dimension leading to large dimensionality of feature. This will result in the reduction of high false positive rate; urge cost of processing which result in to low false positive rate. There are different classification techniques proposed with Artificial Immune System which includes Naïve Bayes, Artificial Neural Network (ANN), Support Vector Machine (SVM) and

other hybrid approaches. (Kotter and Maloof,) and (Wang et al.) [3, 4].

The novelty of this paper is to use the Negative Selection Algorithm (NSA) technique to generate an array of two element self and non-self vector as the feature vector of spam detector. We assume the self and non-self detector memory of [1] as containing bit string which are spam and non-spam respectively. The new self and non-self vector are constructed by using self and non-self detector memory to go through the fixed length segment of a program. The two element self and non-self vector of the program is then connected to form a feature vector in other to identify spam.

2. RELATED WORK

The signature based techniques for spam detection is the most widely used method. It makes use of binary data mining to detect data when given a urge number of data and then use this data to detect data that looks similar in future detection. Henchiri and Japkowicz, [5]. The traditional technique as limitation in detecting spam adequately because it can only detect a small number of generics or extremely broad signatures. It finds it difficult in detecting new spam threats. The suspicious behavior technique provides protection from spam that are yet to exist in spam dictionaries which is not like signature based technique that is meant to detect existing spam. James Clark [6] proposed a neural network system meant for automated e-mail classification. He also presented an email classification NN-based system used for automated e-mail categorization problem. This system is referred to as LINGER. It is an architecture meant for all kind of text categorization. Linger is adaptable, flexible and most of its operation are configured. It recorded a urge success in automated e-mail filing and filtering spam mail. An anti-spam filtering techniques was presented for Turkish natives by Levent Ozgur [7]; His techniques are centered on artificial neural network (ANN) and Bayesian Networks. Algorithm that was created by levent are meant for specific user and they use the characteristics of the incoming e-mail to also make adjustment on themselves. Ian Stuart [8] used from one user a neural network techniques on a corpus of e-mail messages in his research. Descriptive characteristics of words are the feature set used to determine spam massages, This messages are also similar to messages that a reader will use in identifying spam. The experimental work used a corpus of

1654 e-mails which was over a period of some months received by an author. He states that the neural network like Naïve Bayes only need few features to get result. Neural network technique for classification of spam was also presented by D. Puniškis [9]. Attributes of the techniques comprises of the characteristics of the patterns that most network invaders deploy instead of making use of the context of keywords in the message. The dataset that was used in this experiment is corpus of 2788 non-spam and 1812 spam emails that was put together for several months. The result that was acquired from this experiment actually shows that ANN is good but is not the best as it is not suitable to be used alone as tool for filtering spam. Dynamically, suspicious behavior method is a way of knowing detection success which will as well depends on the observable element from an agent externally. Due to the efficiency in executing malicious intend, the proposed method went through criticism. (Jacob *et al*.) and (Schultz et al,) [10,11], proposed the most inspired spam detection technique whose framework comprises of three learning algorithms; The first frame work was the rule based learner that generate Booleans rule based on feature attributes. The second frame work was the probabilistic technique creating a probability of a class been giving some features and finally, is a multiple classifier system that put together results from other classifier to create a prediction. This method includes strings and byte sequence that are extracted from malicious executable on the dataset as different type of features. We actually relate the bytes sequence method with our work and excellent result was achieved with high accuracy. Kolter and Maloof [3] Malicious executable were detected by the use of data mining and n-gram analysis, sequences of bytes was extracted from the executable, and then is been transformed in to n-grams which are then treated as features.

3. Detector Library Generated and Proposed Architecture.

1.	**DEFINITIONS:**
2.	x is a self data set (spam)
3.	y is a non-self data set (non-spam)
4.	N is the number of matching data
5.	SM(0)=0, NSM(0)=0;
6.	**INPUT:**
7.	$\propto$ /* $\propto$ is a threshold
8.	b /* b is the detector of x;
9.	a /* a is the detector of y;
10.	**OUTPUT:**
11.	Finding matching detector of both
12.	self and non-self
13.	**BEGIN**
14.	Input N;
15.	Input SM(1), NSM(1) /*SM is self
16.	matching and NSM is non-self matching;
17.	For i=1 to N
18.	SM(i) = SM(i) + SM (1- i);
19.	Next;

20.	For i =1 to N
21.	NSM (i) = NSM (i) + NSM (i - 1);
22.	**If faffinity>= $\propto$**
23.	f affinity (x) = max;
24.	f affinity (y) = max;
25.	end if
26.	**if** fmatching = =.T.
27.	**(b,x) >= $\propto$**;
28.	else
29.	**(a,y) >= $\propto$**;
30.	End if
31.	End

Fig.1. Self and non-self detector library for Negative selection Algorithm.

The hybrid model proposed in this paper can be divided in to three major process: To generate a self and non-self detector libraries as shown in the algorithm above and proposed in Ismaila and Ali [1]; this was generated from the training dataset made available from machine learning repository at the center for machine learning and intelligent system for classifying e-mail as self and non-self. The 'spam base' last column indicated that the e-mail was considered spam (1) or non-spam (0). The dataset used in this technique has 4601 instances in which 39.4% are spams and each of the instances has 57 attributes. This data set was divided into two classes, we have the training dataset and the testing dataset which was divided in the ratio of 60% and 40% respectively. Secondly is to extract self vector and non-self vector each in training by the use of feature extraction and connect each of the vector to form a feature vector. Lastly is the use of the RBF neural network trained classifiers by using the self vector and non-self vector to detect the testing sample.
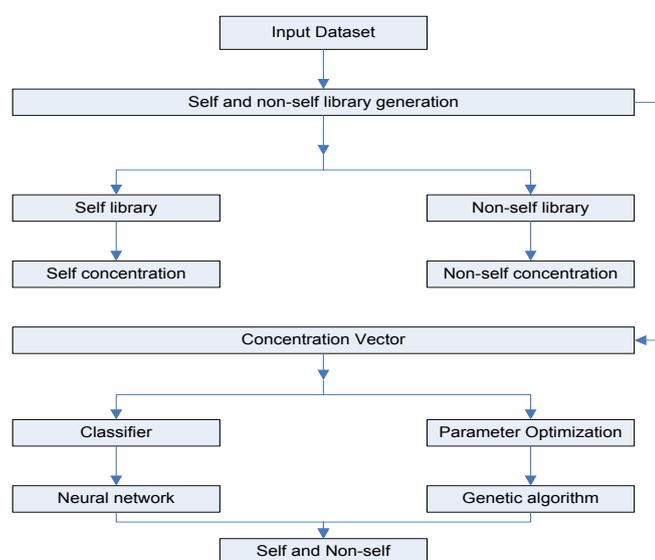
Fig.2. Architecture of the proposed technique.

4. Neural Network Classification.

The sample dataset are characterized by self and non-self feature vector after which are made on the feature vector using neural network. It is an adaptive system which changes the structure base on information that goes through the network in the process of learning as we are trying to simulate the biological function ability of neural network. The neural network topology uses the radial basis function as the activation function. It comprises of three layer: an input layer, a hidden layer with a non-linear activation function and a linear output layer. The sigmoid function represent the output layer of linear combination of hidden layer values, representing an inner probability that is made up of one node which serves as the label of the detected file.

4.1. Generation of Gene Library.

From our previous work, Ismaila and Ali [1], a self and non-self memory was generated. This two library is known as the self gene library and the non-self gene library. Both the self and non-self gene library is composed of fragment (words) with at most representation of non-spam e-mail and spam e-mail respectively. Fragments that are found in non-spam e-mail and rarely found in spam e-mail is a very good representation of non-spam e-mail. To reflect a fragment tendency, it can be generated by the difference of its frequency in non-spam e-mail minus that of the spam e-mail. Fragments are sorted out in order of their differences after each of the frequency for each fragment as been calculated.

The generation of gene libraries is describe in the algorithm below.

1.	**DEFINITIONS:**
2.	sm is a self detector library (spam detector)
3.	nsm is a non-self detector library (non-spam
4.	detector)
5.	x is the fragment (word) which is the sample of
6.	the training set (f_{sm} and f_{nsm})
7.	sm(0)=0, nsm(0)=0;
8.	**INPUT:**
9.	T /* Tendency
10.	b /* b is the gene library of x;
11.	a /* a is the gene library of y;
12.	**OUTPUT:**
13.	Finding the tendency of fragment x in
14.	both sm and nsm
15.	**BEGIN**
16.	Input x;
17.	Input sm(1),nsm(1) /*sm and nsm is its

18.	frequency appearing in both self detector and
19.	non-self detector;
20.	For i=1 to x
21.	sm(i) = sm(i) + sm (1- i);
22.	Next;
23.	For i =1 to x
24.	nsm (i) = nsm (i) +nsm (i - 1);
25.	**If faffinity>= T**
26.	f affinity (sm) = max;
27.	f affinity (smn) = max;
28.	end if
29.	**if** x = =.T.
30.	**(b,sm) >=T**;
31.	else
32.	**(a,nsm) >=T**;
33.	end for
34.	For each fragment x in the sample of
35.	training set do
36.	$f_{(T)} = f_{sm} - f_{nsm}$
37.	else
38.	if $f_{(T)} < 0$ then
30.	x + sm
40.	else
41.	x + smn
42.	end if
43.	end if
44.	end for
45.	Parameter (P_{sm} and P_{nsm}) are to be adjusted.
46.	We remove both P_{sm}% and P_{nsm}% in front and
47.	rear of the queue to form self and non-self gene
48.	library.

Fig.3 Gene library Generated

From the algorithm above, the tendency is acquired by the difference of its frequency in non-spam e-mail minus that in spam e-mail. The fragments are also sorted out accordingly in order of their difference after calculating the difference between each fragment frequency. For example, the two different fragment that are obtained from both front and rear of a queue with some population can be use to generate the self gene library and the non-self gene library.

4.2 Generating Feature Vector

The proportion of the number of fragment in an e-mail that appears in a gene library is referred to as concentration of the e-mail to the number of different type of fragment that exist in the same e-mail. This is represented as follows.

$$C = \frac{N}{F} \qquad (1)$$

C represent concentration, N is the number of fragment appearing in both e-mail and gene library while F is the

number of different fragment in the e-mail. From our previous work, we should note that the gene library could be either self gene library or non-self gene library. For the e-mail classification, we construct a self concentration which describe its similarity to non-spam and a non-spam concentration which describes its similarity to spam.

The time analysis of sorting x fragment, where we represent x as the number of candidate fragments after preprocessing, whose process takes place once during training stage is represented according to algorithm 1 by:

$$0(xlogx) \hspace{4cm} (2)$$

Also from equation 1, the time analysis of generating self concentration and non-self concentration during the running phase is represented as

$$0(n_s * n_x + n_n * n_x) \hspace{3cm} (3)$$

ns and nn represent the number of fragment in both self gene library and non-self gene library respectively while nx is the number of fragment to be classified in the e-mail. Equation 3 is further represented as

$$0(n_s + n_n) \hspace{4cm} (4)$$

As $n_s + n_n < x$, the time analysis for generating a two element feature vector for an element is at most $0(x)$.

5. Experiment and Results

The dataset used is from the center of machine learning and intelligent system. It was used to test the proposed techniques. The corpus is made up of 4601 instances with spam rate of 39.4%. The corpus is divided into partitions with approximate number of instances and spam rate. The spam dataset after division as 1813 instances while the non-spam dataset as 2788 instances. This is as represented below. The red indicate spam (1) while the green represent non-spam (0). A performance index was used for Type equation here.Neural network and SVM to verify the effectiveness of the proposed technique. Clementine software package was used for SVM while Neural network is implemented with MATLAB of version R2009a
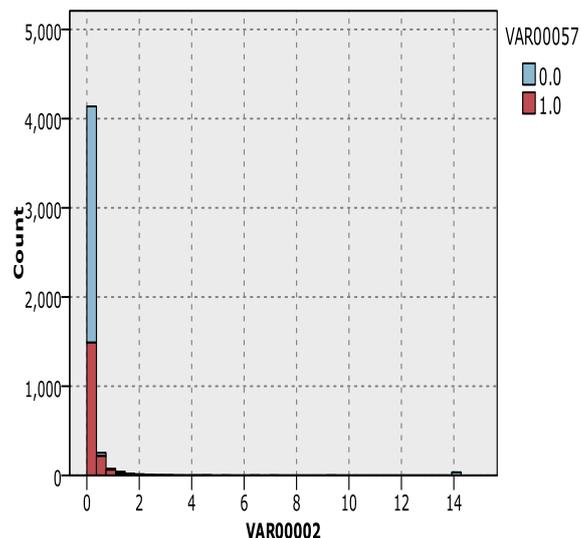

Fig. 4   Dataset Analysis

Self concentration and non-self concentration which corresponds to self gene library and non-self gene library with different classification are trained and tested using Neural network and SVM aiming to find the concentration with the best performance. The result of both Ps and Pn is as illustrated in the figure below.
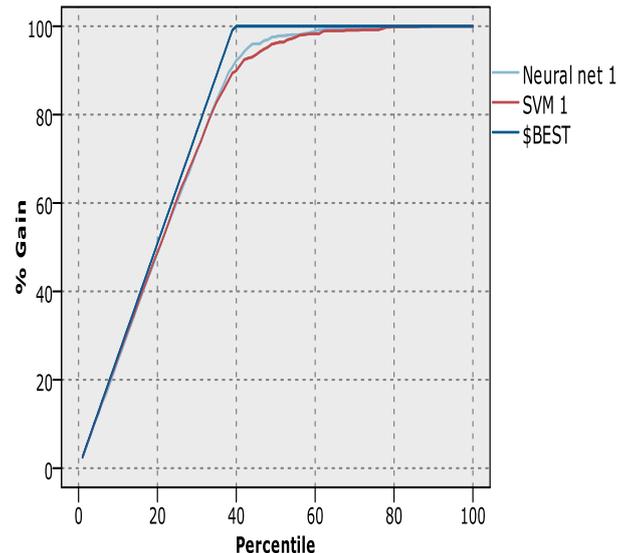

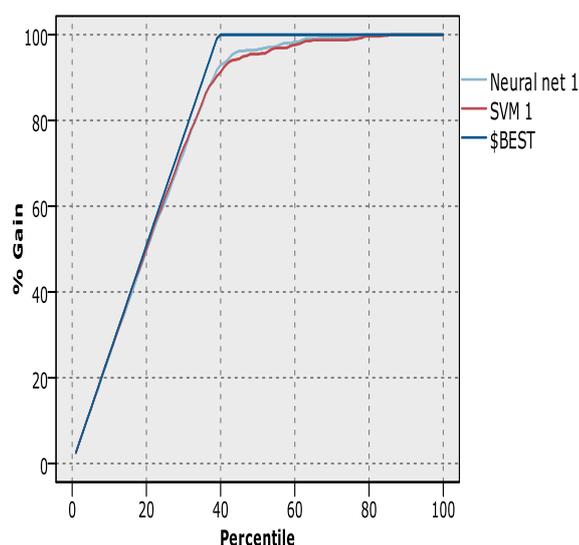Fig.5   Training result for both Neural Network and SVM

Fig.6   Testing result for both Neural network and SVM

Fig. 5 and Fig. 6 above illustrated both training and testing results of the proposed model. Classification using Neural network for training is at its best at 94.306% of accuracy while SVM is 92.654% accuracy. The testing accuracy for the Neural network is at 94.017% with false positive rate of 0.299% while SVM is at 91.378% with false positive rate of 1.274%.

5.   CONCLUSION

The proposed model is able to have a good accuracy and false positive rate with the combination of Negative selection algorithm and Neural network technique compare with the SVM techniques. Computer system complexity is fast becoming a worrying issue and as tremendous influence in spam propagation. Antivirus finds it difficult to detect malware these days as it has become invisible in our computer system. In this paper, we present the self and non-self in a way to create efficiency of detector generation through model and algorithm. The novelty of this paper is to generate a new antibody method that vaccinates randomly created antibody by introducing a new self detector method, with respect to self and non-self producing advance antibody. In consequent research, we shall be looking at constant upgrade of the existing model of antibody (self) in other to prepare it against new spam.

REFERENCES

[1].   Ismaila I. and Ali .S. A Spam Detection Model Based on Negative Selection Algorithm.. is accepted by IJMIA: International Journal on Data Mining and Intelligent Information Technology Applications (ISSN: 2234-3660) - (Accepted October 27, 2011)

[2].   Christodorescu, Mihai, Jha, Somesh, Kruegel, Christopher. Mining specifications of malicious behavior. Proceedings of the the 6th joint meeting of the European software engineering on The foundations of software engineering, New York, NY, USA, p.5–14, 2007

[3].   Kolter, J. Zico, Maloof, Marcus A., learning to detect and classify malicious executables in the wild. J. Mach. Learn. Res., 7:2721–2744. 2006.

[4].   Wang, Jau-Hwang, Deng, P.S., Fan, Yi-Shen, Jaw, Li- Jing, Liu, Yu-Ching, 2003. Virus detection using data mining techinques. Security Technology, 2003. Proceedings. IEEE 37th Annual International Carnahan Conference on, p.71 - 76. 2003.

[5].   Henchiri, Olivier, Japkowicz, Nathalie, a feature selection and evaluation scheme for computer virus detection. Data Mining, IEEE International Conferenceon, 0:891-895. 2006

[6].   James Clark, Irena Koprinska, Josiah Poon, A Neural Network Based Approach to Automated E-mail Classification

[7].   Levent Ozgur, Tunga Gungor, Fikert Gurgen, Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish, Elsevier, 2004.

[8].   Ian Stuart, Sung-Hyuk Cha, and Charles Tappert, A Neural Network Classifier for Junk EMail. Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 7th, 2004Stephanie Forrest and A.S. Perelson, Self nonself discrimination in computre. IEEE, 1994

[9].   D. Puniškis, R. Laurutis, R. Dirmeikis, An Artificial Neural Nets for Spam e-mail Recognition, electronics and electrical engineering ISSN 1392 – 1215 2006. Nr. 5(69). 2006

[10].   Jacob, Grlegoire, Debar, Hervle, Filiol, Eric, behavioural detection of malware: from a survey towards an established taxonomy. Journal in Computer Virology, 4:251-266 2008.

[11].   Schultz, Matthew G., Eskin, Eleazar, Zadok, Erez, Stolfo, Salvatore J., data mining methods for detection of new malicious executables. Securityand Privacy, IEEE Symposium on, 0:0038. [doi:10.1109/SECPRI.2001.924286] 2001