World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 3, No. 3, 56-59, 2013

A Negative Selection Algorithm Based on Email Classification Techniques

Victor. Onomza Waziri, Ismaila Idris
Department of Cyber Security Science,
School of Information and Communication Technology,
Federal University of Technology,
Minna, Niger State. Nigeria.

Mohammed Bashir Abdullahi,
Department of Computer Science,
School of Information and Communication Technology,
Federal University of Technology, Minna-Nigeria,

Hahimi Danladi, Audu Isah
Department of Mathematics/ Statistics,
School of Natural and Applied Sciences,
Federal University of Technology, Minna-Nigeria

Abstract— Aiming to develop an immune based system, the negative selection algorithm aid in solving complex problems in spam detection. This is been achieve by distinguishing spam from non-spam (self from non-self). In this paper, we propose an optimized technique for e-mail classification. This is done by distinguishing the characteristics of self and non-self that is been acquired from trained data set. These extracted features of self and non-self are then combined to make a single detector, therefore reducing the false rate. (Non-self that were wrongly classified as self). The result that will be acquired in this paper will demonstrate the effectiveness of this technique in decreasing false rate.

Keywords- Negative selection; E-mail Classification; Algorithm; Self, Non-Self, Artificial Immune System, Classification accuracy.

I. INTRODUCTION

Classification is one of the familiar techniques used in machine learning. Patterns belonging to different classes are discriminated due to the generation of decision boundaries, It is then divided in to training set and testing set randomly and then classification is made on the training set were as the testing set is used to assess performance of the generated classifier. Spam is often sent as a commercial content. Spam is defined as unsolicited e-mail [1]. These mails are sent randomly to individual numerous mailing lists and of recent times, flood of e-mails are very common. These could result in to mail server congestion; jeopardize our social security and so on. Spam is a very important subject that most be prevented by both technologist and the law of the land. In distinguishing spam messages, various techniques has been proposed by researchers in other to fight against e-mail spam whose results are not very effective due to constant change in patterns of the spam behavior. [2 , 3]. Some anti-spam filters relied on the manually constructed pattern matching rules, but these rules require time and expertise. In addition, spam characteristics changes with time, requiring the rules to be maintained. A system that learns automatically to separate spam from other "legitimate" messages presents important and profitable advantages.

Though, spam detection and e-mail classification problem are been solved using Artificial Immune System [4, 5]. A new e-mail classification technique based on Negative Selection Algorithm shall be designed and implemented. We shall first of all generate a spam and non-spam detector after which e-mail classification will take place by utilizing the non-spam and the spam accordingly in other to successfully reduce the false rate. Our improved classification techniques are also compared with the existing techniques. The experiment confirms the reliability and efficiency of our new techniques in minimizing false positives. The datasets used in this research is gotten from machine learning repository, Center for Machine Learning and Intelligent System.

II. RELATED WORK

Negative Selection Algorithm inspired by biological Immune System. [6] is an emerging learning techniques. Most of the classification algorithms are used for solving spam problems. Most of the techniques focus on machine learning techniques in spam filtering, this learning techniques are: Rule Learning, Naïve Bayes, Support Vector Machine, Artificial Neural Network, Artificial Immune System (AIS), DNA Computing, decision trees and combinations of different learning.

During these years, several classifiers such as naïve Bayes, text compression and artificial neural network have been proposed to detecting and handling the spam. These classifiers are based on probabilistic techniques and machine learning. The following paragraphs represent the related researches summarily.

[7] Used the SVM (support vector machines) algorithm for classifying the emails as spam or non-spam. In addition, besides this algorithm they used and compare it to the three other classification algorithms: Ripper, Rocchio and Boosting decision trees. They apply these algorithms on two different data sets: one of them constrained to 1000 best features and the other one constrained over 700 features. In terms of accuracy and speed, Boosting trees and SVM had the acceptable performance. But, compare to the other three algorithms, SVM's result shows the less training time against the other algorithms.

Perfect feature sequence and multiple features consider by Bayes classifiers, but other classifiers usually use pruning and text pre-processing [9]. Generally, Bayes-based techniques are well known to achieve high spam detection accuracy either as stand-alone classifiers or as parts of classifier ensembles. [8] used feature extraction for spam detection. Basically, they used AIS (Artificial Immune System) for feature extraction. This method extracts a comparatively small set of features which are used as inputs in classification to spam detection. These features are modeled by regular expressions of terms. Features are created from the content of spam messages by using the strings and character matching rules. One of the algorithm that are mentioned and compared with them in their research as the spam detection model is a Back propagation neural network.

III. E-MAIL CLASSIFICATION

During email classification, two mistakes occur by existing anti-spam method. It is either the email is recognized as self and is deleted or non-self and been accepted carelessly. This process is called false positive and false negative. The false positive occurs when the email or data that are needed to create a detector are classified as self while emails or data that are supposed to be discarded are recognized as non-self. This scenario (false negative and false positive) is calculated using classification accuracy which is the main measure of performance. Classification accuracy deals with false positive, false negative and accuracy whose formulae are used to compare different classifier performance [11]. False positive is the percentage of non-self data classified as self while false negative is the percentage of self data which are classified as non-self and accuracy is calculated by the formulae below.

$$\text{Accuracy} = ((TP+TN)/(S+NS)) \text{ X } 100 \quad (1)$$

Where TP represent True Positive; TN represent True Negative; S represent Spam and NS represent Non Spam.

The figure below demonstrates how false positive and false negative are calculated. The first row depict the total non-spam that is divided to true positive and false positive. These rows contain total dataset which are non spam and some are wrongly classified as spam (FP) while others are assigned correctly as non-spam while the opposite is the case with the second row.

Non-spam	spam
True positive (TP)	False negative (FP)
False negative (FN)	True negative (TN)

Figure 1. False positive and false negative

None of the anti spam solution that has been proposed on false positive and false negative approach perfection [12]. Though the result of spam is reduced but not completely. More so, the false positive is more problematic and important than the false negative. Therefore, most researchers and developers are trying to completely get rid of possible false positive mistakes.

IV. CLASSIFICATION TECHNIQUES BASED ON NEGATIVE SELECTION ALGORITHM

The process of classification based on Negative Selection Algorithm is in stages and are summarized as follows:

Our classification techniques comprises of the pre-processing face which encompasses the transformation code, e-mail vector extractor, text body and header separation. The best way we reflect the message characteristics is through extraction which is the process of choosing a vector set. [8]. The size of attachment and text, the word in message and also HTML code are some of the component of the vector. This component varies in importance and can represent the e-mail's characteristics.

The Training generates set of self detector due to most represented self. Best suitable detector can be realized after experiment by adjusting the value at all time. This is due to the center and radius present in the detector. Newer e-mails are been compare with existing detector in other to decide if to add the new e-mails to the detector or not. Though, it is disregarded if the self is closer to the center of any of the detector. This is so as to allow the detector to have a wider range of coverage area instead of having much aggregation.

Also, as a result of the indeterminate length of the e-mail vector, it becomes a tax calculating affinity of two variable length vectors which is defined as below.

$$\text{Affinity } (x,y) = \sum \text{match } (x,y)/l, \; / \; x \; /= l \quad (2)$$

x = Affinity of e-mail x
y = Affinity of e-mail y
L = Denote shortest length of x and y
n and I ensures the affinity is between the range 0 or 1

$$\text{Match } (x,y) = \begin{cases} 1, & \sum_{i=0}^{n} \text{match}(x \rightarrow x_{i\,,} y)/l, / \; x \; /= l \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$\text{Match } (x,y) = \begin{cases} 1, & \sum_{i=0}^{n} \text{match}(y \rightarrow y_{i\,,} x)/l, / \; y \; /= l \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Where, the Matching of x is the total match value between memory-self x And the Matching of y is the total match value between memory- non-self y

The dataset is given as:

$$S= \{<x, y, \text{affinity}>|x, y \in \bigcup_{i-1}^{n} H, \text{affinity} \in N\} \; a \quad (5)$$

Affinity is the ability of the self to match with both self and non-self. This exists when memory-self and memo every-non self are generated. The set is divided in to self training set and non-self training set.

The trained detectors is used to classify in coming e-mail by obtaining feature vector after pre-processing when a new e-mail arrives and both e-mail and detectors affinity are calculated. When we have affinity that is greater than threshold, it is said to be self; otherwise it's a non-self. Then we have intervention manually by the users re-training and error correction. With these, the accuracy and early timing of the detector is certain.

In our optimized e-mail classification method, evaluating the classification method is best achieved by the reduction of false rate. This aspect of reducing the false rate of e-mail classification is a vital aspect of our research work.

Therefore, this research not only depends on the recall rate. A new classification method is been design in this paper with the use of two detector set. The self and non-self detector. They are combined together to form an effective detector thereby reducing false rate were self is considered as non-self and then discarded.

The proposed method consists of suspicious spam extractor and a suspicious spam detector as shown above. In the suspicious spam extractor; both the self detector set and the non-self detector set are generated from the dataset, this is taking from the spam and non-spam programs of the training data set with the help of the suspicious spam extractor. After securing suspicious program (self and non-self) in the suspicious self detector using the suspicious self extractor, an r matching rules will be initiated and computed between the suspicious program and the memory self and memory non-self will be established resulting in to a new memory detector.

V. EXPERIMENTS

We use the machine learning repository from the center for machine learning and intelligent system for classifying e-mail as self and non-self. The 'spam base' last column indicate if the e-mail was considered spam (1) or non-spam (0). The data set used in these techniques 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 data set and the testing data set divided in the ratio 60% to 40%. From the training data set, the self and non-self-undergo a preprocessing stage, finally generating about 100 self and non-self detectors. We select an e-mail from the training data set for word segmentation, and generate a set of self detector by training, which consist of word list, number of detected e-mail and number of matched e-mail. As well we calculate the affinity on property wordlist. If affinity is greater than threshold, we add one to the number of matched e-mail. Also number of e-mail already detected will be increased, irrespective whether the affinity is less than or greater than the threshold.

Consequently, a set of non-self detector is generated similar to that of self detector. We select an e-mail from the training data set for word segmentation. In case of the affinity, the object to be compared is the self detector that was earlier generated instead of the non-self. If any value is said to be greater than threshold, we add the generated detector to the non-self detector set. We compare a non-self detector with self detector in other to find detectors which are similar to self but rather belong to non-self. Other experiment for testing by using the trained detectors generated is carried out after training.

VI. RESULTS AND ANALYSIS

Result obtained from the improved classification techniques and the formal techniques are represented as below.

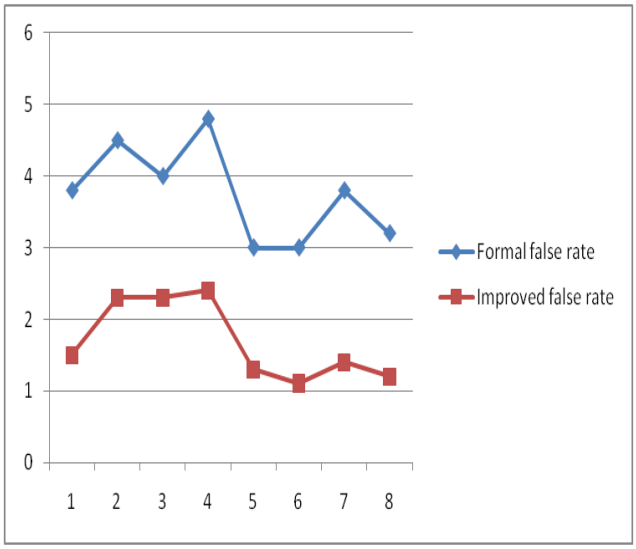

Figure 2. Classification False rate.

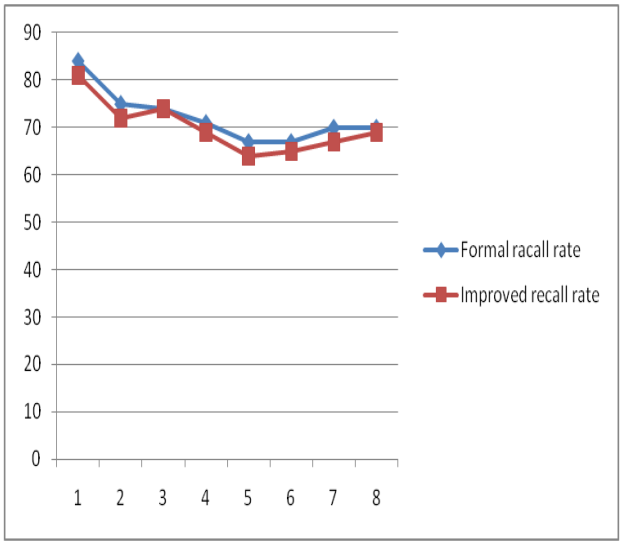

Figure 3. Classification of Recall rate

In Figure 2 above, the result of the traditional (formal) false rate against our improved classification. For our improved false rate classification, our best rate is 1.2%, average is at 1.6% while the worst false rate is at 2.4%. The best rate for the existing techniques is at 3.0% while its average is at 3.8% and the worst false rate is 4.8%. In Figure3, the recall rate is lower than the existing technique when compared, the recall rate for our new technique is 81% at its best, average is 69% while it is worst at 64%. For existing technique, its best recall rate is 84%, average is71% while its worst recall rate is 67%. We see how both figure 2 and figure 3 analyses both performance by

comparison of the improved technique and the formal technique. Figure 3 shows that our recall rate of the improved classification is lower than the existing classification and also the false rate of the improved classification is lower than the existing classification.

VII. CONCLUSION

An improved e-mail classification techniques based on Negative Selection Algorithm is proposed in this paper, the essence is to reduce false positive and create efficiency in spam detectors. We make use of spam and non-spam which are represented as self and non-self in our training data set. This process is very effective in reducing the false rate but its drawback is the reduction of the call rate.

REFERENCES

[1] J.R. Al-Enezi, M.F. Abbod , and S. Alsharhan, Artificial Immune System – Models, Algorithm and Application. IJRRAS, 2010.

[2] Ahmed, K.: An Overview of Content-based Spam Filtering Techniques. Informatica 31, 269–277 (2007).

[3] Graham, P.: Better Bayesian Filtering. In: Proceedings of 2003 Spam Conference (2003).

[4] Oda, T., White, T.: Increasing the Accuracy of a Spam-detecting Artificial Immune System.

In: Proceedings of the Congress on Evolutionary Computation, Canberra, Australia, vol. 1, pp. 390–396 (2003).

[5] Andrew, S., Freitas, A., Timmis, J.: AISEC: An Artificial Immune System for E-mail Classification. In: Proceedings of the IEEE Congress on Evolutionary Computation, Canberra, Australia, pp. 131–139 (2003).

[6] De Castro, L.N., Timmis, J.: Artificial Immune Systems: a New Computational Intelligence Approach. Springer, London (2002).

[7] H. Drucker, D. Wu, and V. N. Vapnik. (September 1999). "Support vector machines for spam categorization," IEEE Trans. Neural Netw., vol. 10, pp. 10481054.

[8] Burim, S., Ohm, S.: Artificial Immunity-based Feature Extraction for Spam Detection. In: Eighth ACIS Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Qingdao, P.R. China, pp. 359–364 (2006).

[9] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. Spyropoulos. (2000). An evaluation of naive bayesian anti-spam ¯ltering. In Proceedings of the Workshop on Machine Learning in the New Information Age: 11th European Conference on Machine Learning, pages 9-17.

[10] Asvial, M.; Sirat, D.; Susatyo, B. (2008). Design and analysis of anti spamming SMS to prevent criminal deception and billing froud: Case Telkom Flexi, Management of Innovation and Technology, 2008. ICMIT 2008. 4th IEEE International Conference on Digital Object Identifier: 10.1109/ICMIT.2008.4654491 Publication Year: 2008 , Page(s): 928 – 933.

[11] Zhang, Y., et al., Immunity-based model for malicious code detection. 2010: Changsha. p. 399-406.

[12] Minh Tran and G. Armitage, Evaluating The Use of Spam-triggered TCP/IP Rate Control To Protect SMTP Servers. Australian Telecommunications Networks & Applications Conference 2004 (ATNAC2004), Sydney, Australia. , December 8-10 2004.