ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 4, April 2016

Improved Negative Selection Algorithm for Email Spam Detection Application

Mohammad Reza Abdolahnezhad, Touraj Banirostam

***Abstract*— Email is a convenient means of communication throughout the entire world today. For email spam detection, previous algorithm compares each email message with spam data before generating detectors. In this paper, we propose an email detection system based on an improvement in the negative selection algorithm (NSA). To cope with the trend of email spam, a novel model that improves the random generation of a detector in NSA with the use of both of spam and non-spam spaces. The theoretical analysis and the experimental result show that the detection performance of our improved proposed NSA is higher than the conventional NSA.**

***Index Terms*—Negative selection algorithm, spam and non-spam email, spam detectors generation.**

I. Introduction

Artificial Immune System (AIS) is a new mechanism implemented for the control of email spam [1], [2], it uses pattern matching in representing detectors as regular expression in the analysis of message. A weight is assigned to the detector which was decremented or incremented when observing the expression in the spam message with the classification of the message based on the threshold sum of the weight of matching detectors. The system is meant to be corrected by either increasing or decreasing all the matching detector weights with a 1000 detector generated from spam-assassin heuristic and personal corpus. The results were acceptable on the basis of the few number of detectors used. A comparison of the two techniques to determine message classification using spam-assassin corpus with 1000 detectors was also proposed in [3]. This approach is like the previous techniques but the difference is the increment of weight where there is recognition of pattern in the spam messages. The weighting of features complicates the performance of the matching process.

The implementation of different pattern recognition scheme inspired by the biological immune system in order to identify uncommon situations like the email spam [4-6], unfortunately, has not been able to produce outstanding result.

It is quiet desirable to determine quantitatively the coverage of certain negative selection algorithm (NSA) or make a conclusion on how detectors are distributed and their coverage in the spam space. For the binary matching rules commonly used in NSA, in [7] first proposed the r-chunk matching rules which is an improvement over the r-contiguous matching rule originally proposed by Forest et al. [8].

An improved NSA by introducing a novel training is proposed in [9]. With consideration for spam and non-spam class as a source of information, a data compression model operating at raw message level was proposed in [10]. Also, with the assumption of a black list that is made up of words that are related to the spam messages, a hidden Markov model (HMM) was applied to the problem of finding observed words in [11]. In [12], a new improved model that combines NSA with particle swarm optimization (PSO) has been proposed and implemented which PSO implementation with local outlier factor (LOF) as fitness function no doubt improved the detector generation phase of NSA. Then, the proposed improved model serves as a better replacement to NSA model.

Implementation of different spam detection methods based on machine learning techniques was proposed to solve the problem of numerous email spam ravaging the system. The NSA method used in email spam detection compares each email message with spam data before generating detectors while our proposed system proposed, refer to as *NSA-II*, an email detection system that is designed based on both of spam and non-spam spaces in the NSA.

The rest of this paper is organized as follows. In Section II, we introduce the original negative selection model. In Section III, we present our main contribution, that is, the proposed improved model and its constituent framework are discussed. We express some criterions for evaluation results and discussions along with comparative simulation results is presented in Section VI; and finally in Section V, we conclude the paper.

II. The Original Negative Selection Algorithm

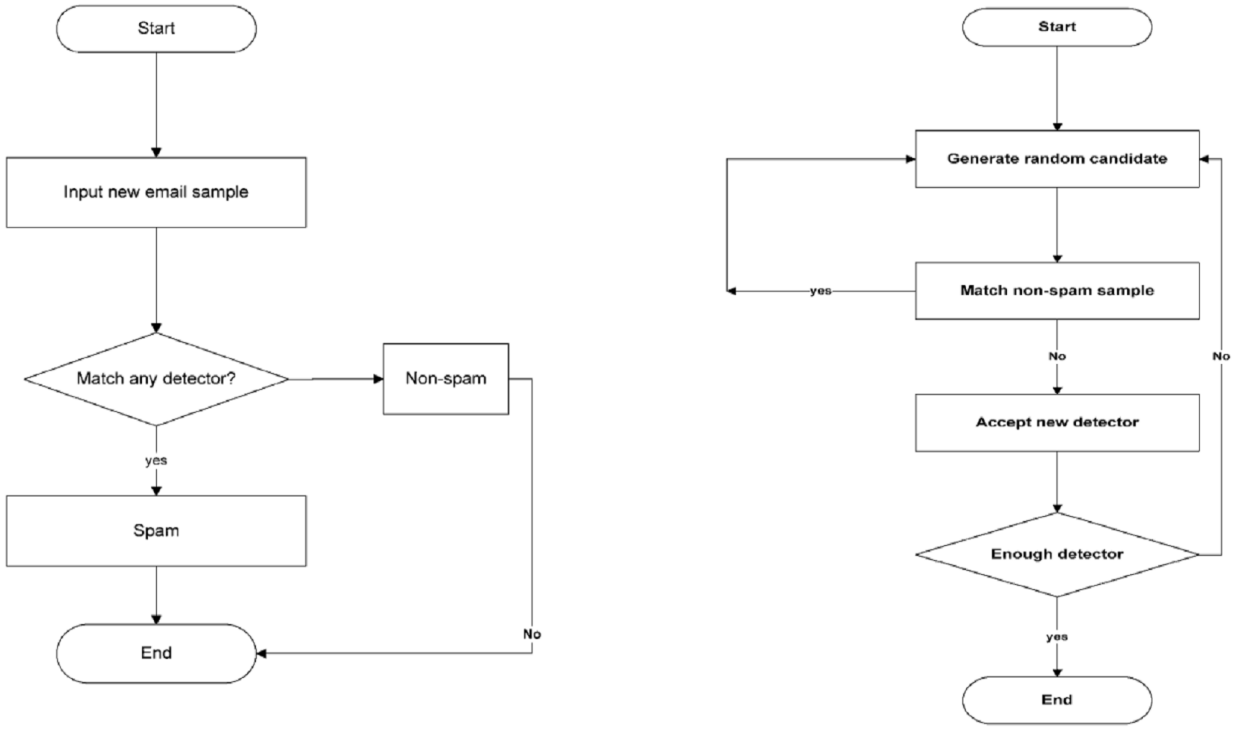
The NSA, was proposed by Forest et al. [8], has been used widely for applications in the construction of the AIS [7]. The algorithm comprises of the *data representation* phase, the *training phase* and the *testing phase*. Data are denoted in a binary or in a real valued representation, in the data representation phase. The training phase of the algorithm refer to as the *detector generation phase*, randomly produce detector with binary or real valued data. Hence, the detectors are consequently used to train the algorithm [7], while the testing phase evaluates the trained algorithm. Fig. 1 illustrate the training and testing phase of NSA.

The main concept of the NSA was meant to generate a set of candidate detectors C, such that $\forall xi \in C$ and $\forall z_p \in S$,

Manuscript received March, 2016.

***Mohammad Reza Abdolahnezhad and Touraj Banirostam*,** *Department of Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran.*

ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 4, April 2016



(a) **(b)**

Fig. 1. a) Detector generation of NSA; b) Testing of NSA [12].

$f_{MATCH}(x_i, z_p) < r$ where x_i is a detector, z_p is a pattern and $f_{MATCH}(x_i, z_p)$ is the affinity matching function.

The proposed spam base dataset for the research is in real valued representation. The real value NSA is encoded in real value for classifying non-spam and spam. The candidate detectors are randomly generated and then compared to the non-spam patterns. Candidate detectors that do not match any pattern of the non-spam set are accepted as feasible detectors. Candidate detectors that match the pattern the non-spam set are discarded as undesirable detectors. After the generation of detectors in the spam space, the generated detectors can then monitor the status of the system. If some other test patterns match at least one of the detectors in the system, it is assumed to be spam which is abnormal to the system; but if the test pattern does not match any of the generated detectors in the spam space, it is assumed to be non-spam. The non-spam sample in a real value NSA is represented in N-dimensional points and a non-spam radius Rs, as training dataset. One detector is denoted as $d_j = (\mathrm{C}_j, R_j^d)$ where C_j is the detector center and R_j is the detector radius. The Euclidean distance is used as the matching measurement. The distance between non-spam sample X_i and the detector d_j can be expressed as [12]

$$L(\mathrm{X}_i, \mathrm{d}_i) = \sqrt{\left(x_{i,1}, -\mathrm{C}_{j,1}\right)^2 + \ldots + \left(x_{i,N}, -\mathrm{C}_{j,N}\right)^2} \quad (1)$$

Since $L(X_i, d_j)$ is compared with the non-spam space threshold Rs, obtaining the match value of α, as [12]

$$\alpha = L(\mathrm{X}_i, \mathrm{d}_i) - rRs \quad (2)$$

The detector d_j fails to match the non-spam sample X_i if $\alpha > 0$; therefore 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 $R_d^j = \min(\alpha)$, if $\alpha \le 0$. Also, if detector d_j matches the non-spam sample, the detector will be eliminated. The generation of detectors continues until the number of detectors needed to cover the spam space is attained. After the generation of detectors in the spam space, the detectors are then used to monitor the system status. If the testing dataset matches any detector in the spam space, it is labelled as spam but if the testing dataset set does not match any detector in the spam space, it is labelled as non-spam.

III. The Proposed Improved Negative Selection Algorithm Model

Classic NSA has a lot of problems that lowers its effectiveness in spam detection system: **i.** The main problem is that classic NSA isn't basically consider spam patterns in the training phase; **ii.** The classic NSA has a mechanism which too cautious; and, **iii.** Since the detectors are generated only by taking non-spam patterns, detectors acceptable are away from all non-spam patterns at least as minimum radius. Then, we the improved NSA proposed.

In the improved NSA are considered two various detector types. Actually in the training phase, two set detectors are generated; St1 as the set of spam detectors and St2 as the set of non-spam detectors. Each detector of St1 or St2 set is acceptable if it detect at least α_1 or α_2 percentage of spam or non-spam patterns, respectively. Our proposed training phase is shown in Fig. 2(a).

In testing phase, the detector output for St1 and St2 are separately determined. If one of the detectors set St1 make known new pattern, new email knows as a spam pattern. Otherwise, it considers as a non-spam pattern. Similarly,

ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 4, April 2016

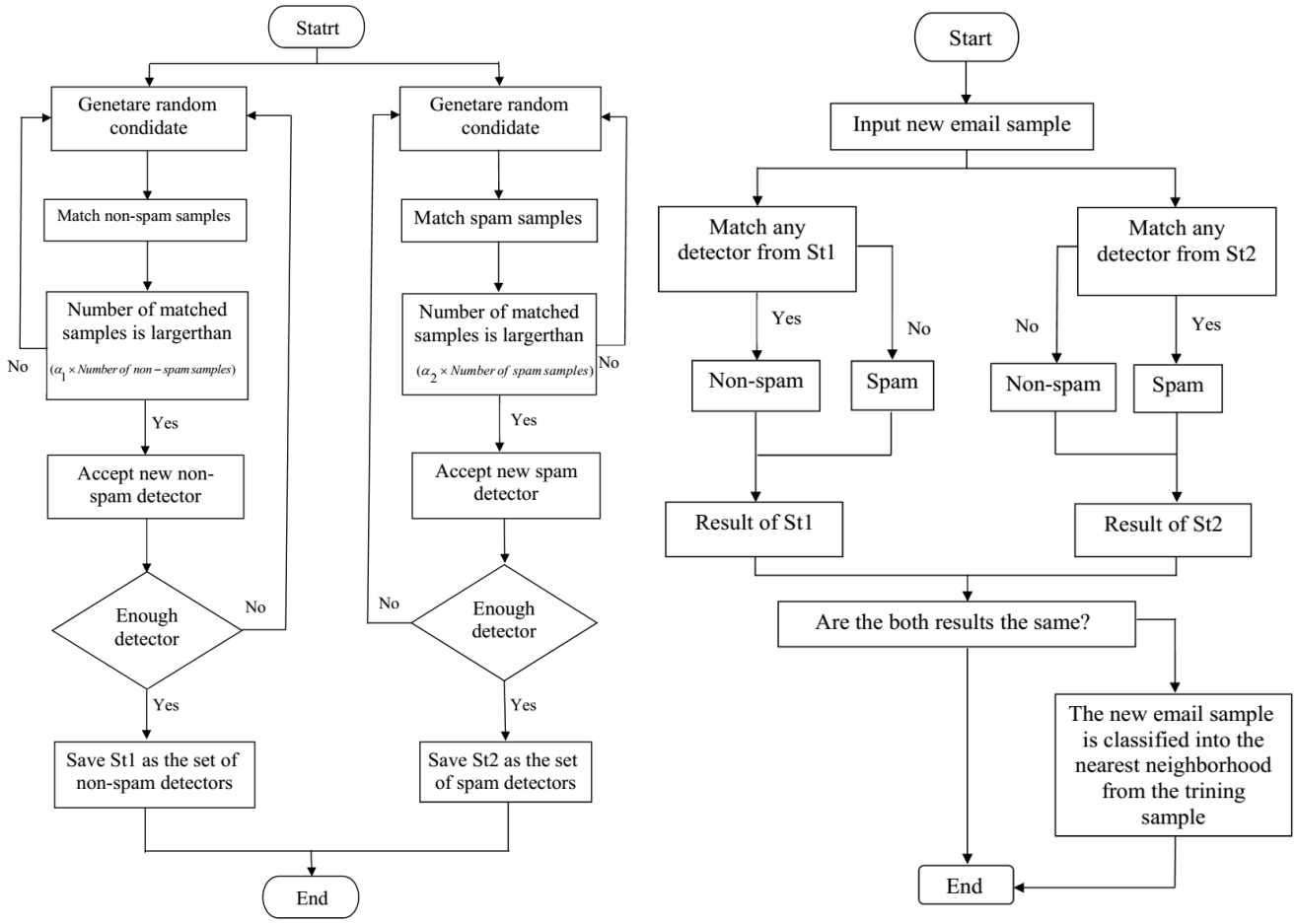


Fig. 2. a) Detector generation of improved NSA; b) Testing of improved NSA.

new email knows as a non-spam pattern if one of the detectors set St2 make known new pattern; otherwise, it considers as a spam pattern. If the both results aren't same, the new email sample is classified into the nearest neighborhood from the training sample. Our proposed testing phase for NSA-II is shown in Fig. 2(a).

IV. Results and Discussions

A. Criteria for performance evaluation

Different measures can be used to evaluate and compare performance and accuracy of NSA and improved NS methods. Then, statistical quality measure can be employed in machine learning and data mining journals. They are *Sensitivity (SN)*, *Positive prediction value (PPV)*, *Negative prediction value (NPV)*, *F-measure (F1)*, *Accuracy (ACC)* and *Correlation coefficient (CC)*.

1) Sensitivity (SN):

The sensitivity measures the proportion of positive pattern that are correctly recognized as positive, as follow

$$SN = \frac{TP}{TP + FN} \tag{3}$$

where *TP* is the number of true positive and *FN* is the number of false negative.

2) Positive prediction value (PPV):

The positive prediction value of a test gives a measurement of the percentage of true positives to the overall number of patterns that are recognized to be positive. It measures the probability of a positively predicted pattern as positive, as follow

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

where *FP* is the number of false positive.

3) Negative prediction value (NPV):

The negative prediction value of a test also gives the measurement of percentage of true negative to the overall number of patterns recognized to be negative, as follow

$$NPV = \frac{TN}{TN + FN} \tag{5}$$

where *TN* is the number of true negative.

4) Accuracy (Acc):

The accuracy measures the percentage of samples correctly classified, as follow

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

5) F-measure (F1):

The F-measure combines both positive predictive value and sensitivity, as follow

$$F1 = 2 \times \frac{PPV \times SN}{PPV + SN} \tag{7}$$

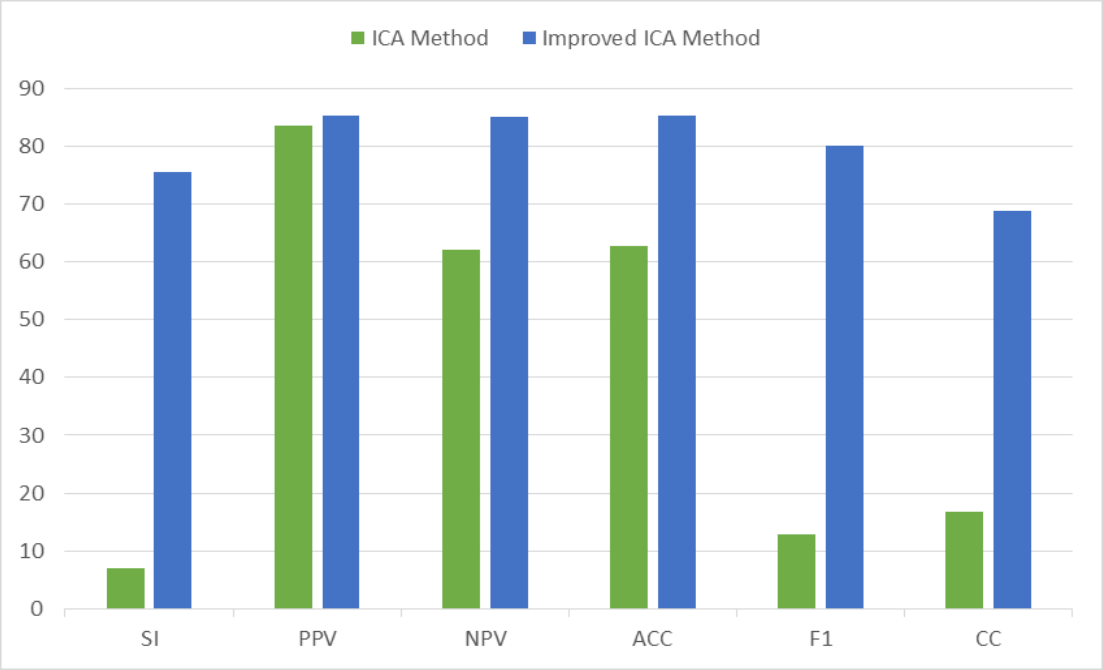


Fig. 3. Performance of improved NSA and NSA methods.

6) Correlation coefficient (CC):

The correlation coefficient is used as a measure of the quality of binary classification in machine learning, as follow

$$CC = \frac{TP \times TN - FP \times FN}{(TP + FN)(TP + FP)(TN + FP)(TN + FN)} \quad (8)$$

TABLE 1. Performance of improved NSA and NSA methods.

	ICA Method	ICA-II Method
SI	7.019959	75.63661
PPV	83.60656	85.39239
NPV	62.0399	85.213
ACC	62.75469	85.27574
F1	12.95	80.21
CC	16.71709	68.88093

B. Simulation Results

Here, we evaluate the performance of our improved NSA method. The corpus bench-mark is established from spam base dataset which is an acquisition from email spam message. This dataset is made up of 4601 messages and 1813 (39%) of the messages are marked to be spam messages and 2788 (61%) are identified as non-spam and was acquired by [13]. The features are represented as 58-dimensional vectors.

Fig. 3 and Table 1 investigate sensitivity, positive prediction value, negative prediction value, F-measure, accuracy and correlation coefficient criterions for ICA and improved ICA methods. It is shown that our improved NSA can achieve a higher performance than that achieved by the conventional methods. It shows that compared to the conventional NSA, our improved NSA algorithm improves positive prediction value, negative prediction value, accuracy and correlation coefficient criterions about 2%, 37%, 35% and 312%, respectively.

V. Conclusion

In this paper, improved NSA, has been proposed and implemented. Since efficient and effective robust algorithm determined by the detector generation phase of NSA, the NSA-II method works as a better replacement to the conventional NSA method. Also, performance and accuracy investigation has shown that the NSA-II method is capable to detect email spam better than the conventional NSA method.

References

[1] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "A survey and experimental evaluation of image spam filtering techniques", *Pattern Recognition Letters*, vol 32, no. 10, pp.1436-1446, 2011.

[2] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey", *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3634-3642, 2015.

[3] T. Oda, T. White, "Increasing the accuracy of a spam-detecting artificial immune system", *in: The 2003 Congress on Evolutionary Computation (CEC)*, 2003.

[4] A.H. Mohammad, R.A. Zitar, "Application of genetic optimized artificial immune system and neural networks in spam detection", *Appl. Soft Comput.*, vol. 11, no. 4, pp. 3827–3845, 2011.

[5] A. Visconti, H. Tahayori, "Artificial immune system based on interval type-2 fuzzy set paradigm", *Appl. Soft Comput.*, vol. 11, no. 6, pp. 4055–4063, 2011.

[6] N. Pérez-Díaz, D. Ruano-OrdáS, F. Fdez-Riverola, and J. R. MéNdez, "SDAI: an integral evaluation methodology for content based spam filtering models", *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12487–12500, 2012.

[7] J. Balthrop, S. Forrest, M.R. Glickman, "Revisiting LISYS: parameters and normal behavior", *in: Proceedings of the Congress on Evolutionary Computing*, 2002.

[8] S. Forrest, A.S. Perelson, "Self Nonself Discrimination in Computer", 1994.

[9] M. Gong, J. Zhang, J. Ma, and L. Jiao, "An efficient negative selection algorithm with further training for anomaly detection", *Knowl.-Based Syst.*, vol. 30, pp. 185–191, 2012

[10] A. Bratko, B. Filipič, G. V. Cormack, T. R. Lynam, and Zupan "Spam filtering using statistical data compression models", *J. Mach. Learn. Res.*, vol. 7, 2673–2698, 2006.

[11] J. Gordillo, E. Conde, "An HMM for detecting spam mail", *Expert Syst. Appl.*, vol. 33, no. 3, pp. 667–682, 2007.

[12] I. Idris, and A. Selamat, "Improved email spam detection model with negative selection algorithm and particle swarm optimization", *Applied Soft Computing*, vol. 22, pp.11-27, 2014.
[13] M. Hopkins, E. Reeber, G. Forman, and J. Suermondt, UCI Machine Learning Repository: Spambase Data Set, Hewlett-Packard Labs, 1999, https://archive.ics.uci.edu/ml/datasets /Spambase.