IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Hybrid Email Spam Detection Method Using Negative Selection and Genetic Algorithms

Mohammad Reza Abdolahnezhad, Touraj Banirostam

Department of Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran

Abstract: In this paper, a new model was proposed to cope with the trend of email spam that improves the generation of a detector in the improved and standard negative selection algorithm (NSA) with the use of stochastic distribution to model the data point using genetic algorithms. The theoretical analysis and the experimental result show that the performance of proposed method is higher than the improved and standard NSA, which the accuracy of the proposed model is 91.90%, while the improved NSA model is 85.27%, and the standard NSA model is 62.75%.

Keywords: Negative selection algorithm, genetic algorithm, spam email, spam detectors generation.

I. INTRODUCTION

Artificial Immune Systems (AIS), motivated by the natural immune systems, are an emerging kind of soft computing approaches [1]. With the features of pattern recognition, anomaly detection, data analysis, and machine learning, the AIS have recently gained considerable research interest from diverse communities [2]. A weight is assigned to the detector which was decremented or incremented when observing the expression in the spam message. 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. 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 [4-7], 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 [8] first proposed the r-chunk matching rules which is an improvement over the r-contiguous matching rule originally proposed by Forest et al. [9].

An improved NSA by introducing a novel training is proposed in [10]. 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 [11]. 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 [12]. In [13] and [14], a new improved model that combines NSA with PSO algorithm 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. Also in [5], genetic optimized spam detection using AIS and spam detection using Continuous Learning Approach ANN is proposed which are good enough to be used as an anti-spam effective detection methods to fight spam.

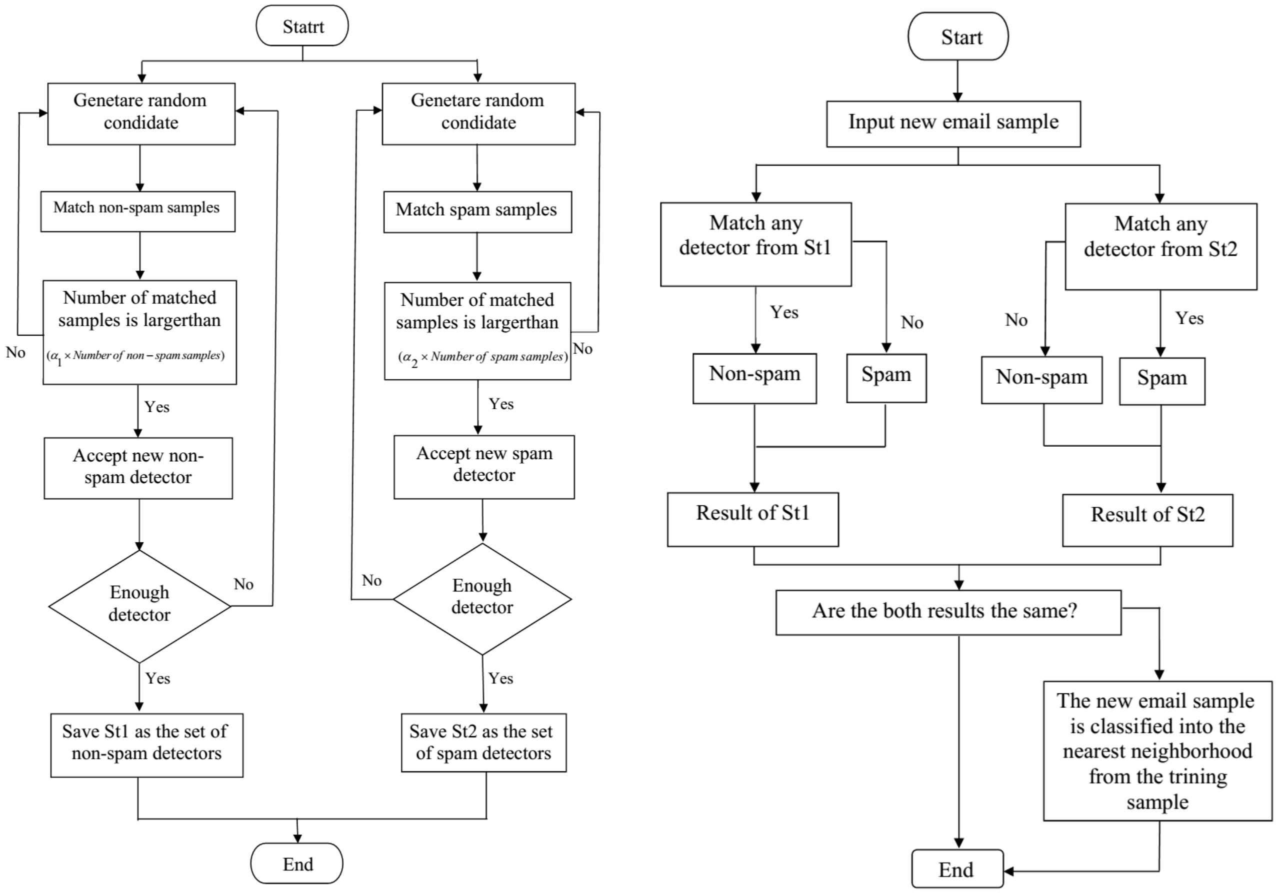
The NSA method used in email spam detection compares each email message with spam data before generating detectors while our proposed system proposed, refer to as NSAII, an email detection system that is designed based on both of spam and non-spam spaces in the NSA. In this paper, we propose an improved solution for email spam detection inspired by the AIS by the adoption of spam detection generation techniques with NSA and genetic algorithm (GA), refer to as GA-NSAII. The GA was implemented to generate detectors for training of NSA to cover the spam space instead of the original random generation of detector use by NSA.

The rest of this paper is organized as follows. In Section II, we introduce our improved NSA model. We express some criterions for evaluation results in Section III. In Section VI, we present our proposed improved model with GA. We discuss along with comparative simulation results is presented in Section VI; and finally in Section V, we conclude the paper.

II. IMPROVED NEGATIVE SELECTION ALGORITHM

The NSA, was proposed by Forest et al. [9], has been used widely for applications in the construction of the AIS [8]. 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 [8], while the testing phase evaluates the trained algorithm.

DOI 10.17148/IJARCCE.2016.5401


Statrt
Genetare random condidate
Match non-spam samples
Number of matched samples is largerthan
($\alpha_1 \times$ Number of non − spam samples)
No
Yes
Accept new non-spam detector
Enough detector
No
Yes
Save St1 as the set of non-spam detectors
Genetare random condidate
Match spam samples
Number of matched samples is largerthan
($\alpha_2 \times$ Number of spam samples)
No
Yes
Accept new spam detector
No
Enough detector
Yes
Save St2 as the set of spam detectors
End
Start
Input new email sample
Match any detector from St1
Yes
No
Non-spam
Spam
Match any detector from St2
No
Yes
Non-spam
Spam
Result of St1
Result of St2
Are the both results the same?
The new email sample is classified into the nearest neighborhood from the trining sample
End

(a) (b)

Fig. 1. a) Detector generation of improved negative selection algorithm; b) Testing of improved negative selection algorithm.

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. 1(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, 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 neighbourhood from the training sample. Our proposed testing phase for NSA-II is shown in Fig. 1(a).

III.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; as follow:

1. Sensitivity (SN):

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

$$SN = \frac{TP}{TP + FN} \qquad (1)$$

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

$$PPV = \frac{TP}{TP + FP} \qquad (2)$$

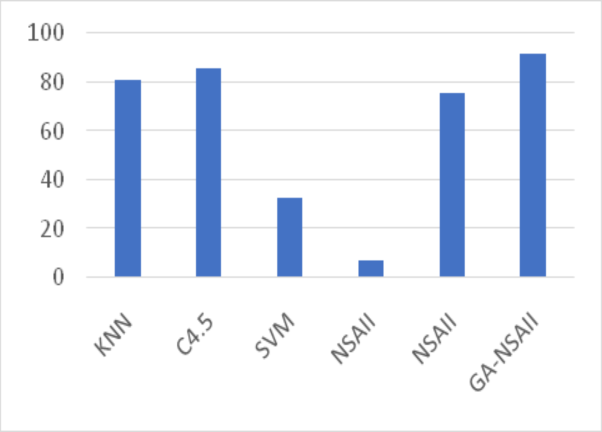


Fig.2. Performance of the intensity criterion.

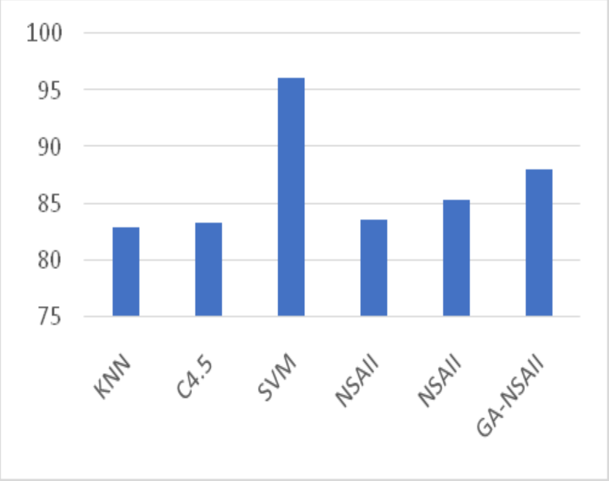


Fig 3.Performance of the positive prediction value criterion

3. F-measure (F1):
The F-measure combines both positive predictive value and sensitivity, as follow

$$F1 = 2 \times \frac{PPV \times SN}{PPV + SN} \qquad (3)$$

4. Correlation coefficient (CC):
The correlation coefficient is used as a measure of the quality of binary classification in machine learning, as follow

$$CC = \frac{TP \times TN - FP \times FN}{(TP + FN)(TP + FP)(TN + FP)(TN + FN)} \qquad (4)$$

where TN is the number of true negative and FP is the number of false positive.

5. Accuracy (Acc):
The accuracy measures the percentage of samples correctly classified, as follow

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \qquad (5)$$

IV. IMPROVED GENETIC NEGATIVE SELECTION ALGORITHM

GA is a meta-heuristic population-based evolutionary algorithm that belongs to the larger class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution such as selection, crossover, and mutation.

A. Generation of the initial population:
In this problem, binary encoding is used to represent a feasible solution. Each chromosome can be represented as a binary string. Similar to other swarm intelligence algorithms, the initial population of GA is generated randomly, in such a way that each gene within the each chromosome has equal probability to be “0” or “1”.

B. Fitness evaluation:
After updating the population in the every iteration, the fitness of the each chromosome is evaluated according to the proposed fitness function as

$$\begin{aligned} Cost_i = {} & w_1 \times (1 - ACC_i) + w_2 \times (1 - SN_i) \\ & + w_3 \times (1 - SP_i) + w_4 \times (1 - PPV_i) \\ & + w_5 \times (1 - NPV_i) + w_6 \times \frac{L_i}{N} \end{aligned} \qquad (6)$$

After evaluation of the fitness values for all chromosomes, all chromosomes are sorted from the best to the worst. Then, some of the best chromosomes are selected based on a selection strategy (e.g. roulette wheel election, rank selection, or elitism selection), as the parents for updating the population. In this paper elitism selection strategy was used.99

C. Population updating:
In order to generate an offspring, two parents are randomly chosen among the best individuals (parents) of the population, and crossover operator is performed on them. In this paper, uniform crossover operator is used. In uniform crossover, each gene in the offspring is copied from the same gene from one of the two parents, with the same probability. After generation of each offspring, each gene can be mutated with the probability of P_m .The binary swap mutation was used, in which, the value of the selected gene is complemented. We used an adaptive mutation strategy in order to improve the investigation of the search space. In this way, the mutation probability P_m is considered to be decreased linearly from P_m^{max} to P_m^{min} during execute algorithm, as

$$P_m = P_m^{max} + \frac{I_c}{I_{max}} \left(P_m^{min} - P_m^{max} \right) \qquad (7)$$

where I_c and I_{max} are the current iteration, and the defined maximum number of iterations for GA, respectively. The larger P_m, the more genes would be complemented via mutation operator.

DOI 10.17148/IJARCCE.2016.5401

V. RESULTS AND DISCUSSIONS

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 [15]. The features are represented as 58-dimensional vectors.

Fig. 2 investigates the sensitivity criterion for different methods. It is shown that a higher performance is achieved by the proposed method than the standard NSA, the SVM, the C4.5 and the KNN models.

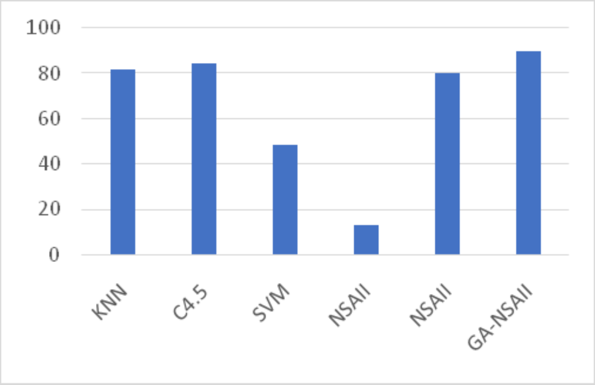


Fig 4 Performance of the F1 criterion

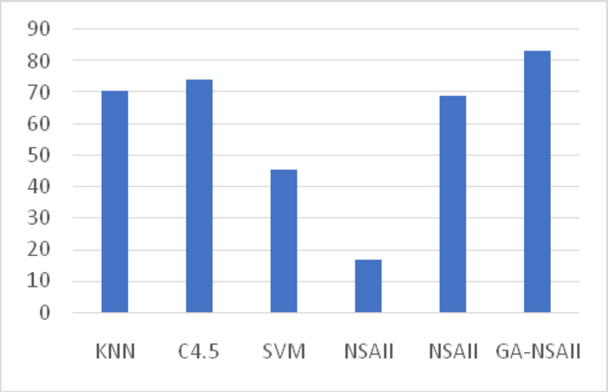


Fig. 5. Performance of the correlation coefficient criterion.

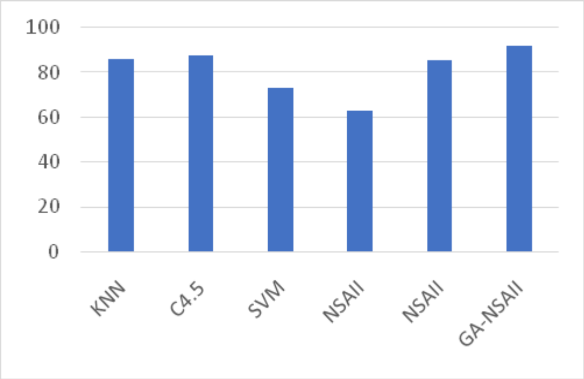


Fig. 6. Performance of the accuracy criterion

The positive prediction value, the F-1 measure, the correlation coefficient and the accuracy criterions analysis in Fig. 3, Fig. 4, Fig. 5 and Fig. 6, respectively. In general, the proposed model outperforms the standard NSA and PSO models. Fig. 6 shows that the difference in performance between the proposed GA-NSA model with the other models are very significant, the best accuracy of the proposed model is 91.90%, while the improved NSAmodel is 85.27%, the standard NSA model is 62.75%, the SVM model is 87.612%, the C4.5 model is 85.87%, and the KNN model is 68.86%, respectively.

The comparison between proposed improved GA-NSA, the improved NSA, the standard NSA, the SVM, the C4.5 and the KNN models using validation of an unseen data is summarized in Table I. The proposed model shows an improved performance when compared with others.

TABLE I. Performance of GA improved NSA, improved NSA, NSA, SVM, C4.5 and KNN methods

	SN	**PPV**	**F1**	**CC**	**Acc**
KNN	80.79	82.96	81.86	70.31	85.87
C4.5	85.75	83.34	84.53	74.22	87.612
SVM	32.55	96.13	48.63	45.53	72.86
NSA	7.019	83.60	12.95	16.71	62.75
NSAII	75.63	85.39	80.21	68.88	85.27
GA-NSAII	**91.94**	**88.06**	**89.96**	**83.24**	**91.90**

VI. CONCLUSION

In this paper, we combine improved NSA with GA. Since efficient and effective robust algorithm determined by the detector generation phase of NSA, the NSAII method works as a better replacement to the standard NSA method. GA implementation with LOF as fitness function no doubt improved the detector generation phase of NSAII and NSA. Performance and accuracy investigation has shown that the proposed method is capable to detect email spam better than the improved NSA and the standard NSA methods. The proposed improved model serves as a better replacement to the other models.

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), 2003.

[4] S. Afaneh, R.A. Zitar and A. Al-Hamami, “Virus detection using clonal selection algorithm with Genetic Algorithm (VDC algorithm)”, Appl. Soft Comput, vol. 13, no. 1, pp.239-246, 2013.

[5] 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.

[6] 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.

DOI 10.17148/IJARCCE.2016.5401

[7] 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.

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

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

[10] 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

[11] 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.

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

[13] I. Idris, and A. Selamat, "Improved email spam detection model with negative selection algorithm and particle swarm optimization", Appl. Soft Comput., vol. 22, pp.11-27, 2014.

[14] I. Idris, A. Selamat, N.T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system", Engineering Applications of Artificial Intelligence, vol. 39, pp.33-44, 2015.

[15] 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

DOI 10.17148/IJARCCE.2016.5401