Real-Valued Negative Selection Algorithms: Ensuring Data Integrity Through Anomaly Detection

Rihab Salah Khairy, Rozaida Ghazali and Ayodele Lasisi

Abstract The Real-Valued Negative Selection algorithms which are the focal point of this work generate their detector set based on the points of self data. Self data is regarded as the normal behavioural pattern of the monitored system. An anomaly in data alters the confidentiality and integrity of its content thereby causing a defect for making useful and accurate decisions. Therefore, to correctly detect such an anomaly, this study applies the real-valued negative selection with; fixed-sized detectors (RNSA) and variable-sized detectors (V-Detector) for classification and detection of anomalies. Classifier algorithms of Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are used for benchmarking the performances of the real-valued negative selection algorithms. Experimental results illustrate that RNSA and V-Detector algorithms are suitable for the detection of anomalies, with the SVM and KNN producing significant efficiency rates. It was also gathered that V-Detector yielded superior performances with relation to the other algorithms.

Keywords Artificial immune system · Real-valued negative selection algorithm · Variable detector · Anomaly detection

R.S. Khairy · R. Ghazali · A. Lasisi (✉)
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja,
Batu Pahat, Johor, Malaysia
e-mail: lasisiayodele@yahoo.com

R.S. Khairy
e-mail: rehabsalah08@gmail.com

R. Ghazali
e-mail: rozaida@uthm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_3

1 Introduction

In modern day life, anomaly is one of the major cause of great losses. A number of anomaly detection techniques are proposed in handling issues related with protecting the integrity of data. These techniques are adequately applied to fault tolerance, robotic control, network intrusion detection, bioinformatics. In general, the problem of anomaly detection can be seen as a two or more class classification problem. Given an element from a given problem space, the system should classify it as normal or abnormal [1, 2]. However, this is a very general characterization since it can correspond to different problems depending on the specific context where it is interpreted. Therefore, from a statistical point of view, the problem can be seen as that of outlier detection which is referred to as an observation deviating from other observations and triggering uncertainty as to how it was generated [3].

Many modern techniques exist in literature that are based on Artificial Intelligence, Neural Network, Bayesian Network, Fuzzy logic, K-nearest Neighbour algorithm, Support Vector Machine, Decision Tree, Fuzzy Logic Based System, Sequence Alignment, Genetic Programming etc., and has evolved in detecting various anomalies [4].

The field of Artificial Immune Systems which began in the early 1990s serve as alternative and efficient algorithms for detecting anomalies to the already existing methods. They were inspired by the Biological Immune System (BIS) which is robust, decentralized, error tolerant, and adaptive in nature. The immune system is highly complicated and appears to be precisely tuned to the problem of detecting and eliminating infections [5]. There are a number of AIS models used in pattern recognition, fault detection, computer security, and a variety of other applications in the field of science and engineering. These AIS models tend to mimic the biological processes of negative selection, immune network, clonal selection, and dendritic cell/danger theory. These models emphasize on designing and applying computational algorithms and techniques using simplified models of various immunological processes and functionalities [6]. The negative selection algorithms which utilizes real-valued representations namely RNSA and V-Detector, as surveyed in [7] shall be applied for anomaly detection.

Hence, in this study, a performance analysis resting on the proficiency of real-valued negative selections; RNSA and V-Detector algorithms for anomaly detection are explored and examined. The structure of the paper is highlighted as follows: Sect. 2 describes the artificial immune system, its inspiration and some of its models. Negative Selection Algorithm and insight into the real-valued negative selection algorithms are discussed in Sect. 3. Experimental simulations, results and analysis are reflected in Sect. 4. The contribution of the study concludes with Sect. 5.

2 Artificial Immune System

The concept and theory of Artificial Immune System will be incomplete without the mention of the source of inspiration in bringing its algorithms into being, referred to the Biological Immune System (BIS). The body has different mechanisms to protect itself (*self* cells) from harmful foreign materials. One of these mechanisms is the natural immune system and its main purpose is to detect and destroy any unwanted foreign cells (*non-self* cells) that could be harmful to the body. These *non-self* cells are known as *antigens* and the natural immune system produces *antibodies* to bind to these antigens.

The Biological Immune System mainly consists of lymphoid organs that create lymphocytes. The two most familiar lymphocytes are the T-cell and B-cell formed in the bone marrow. Both T-cell and B-cell have receptors on their surfaces to bind with the antigen [8]. The immune system is a natural resistance to diseases using sophisticated adaptive mechanisms intended either to destroy the invaders or to neutralize their effects. The BIS can be classified according to functionality into two different layers of defence which are innate and adaptive. The innate immunity is the first line of defense and its non-specific. It is categorized as non-specific because does not concentrate on a particular type of pathogen. When an invasion bypass the innate immunity, the adaptive immunity line of defense is called into action. Adaptive immunity is specific as it targets, matches a particular pathogen, and stores in memory the structure of that pathogen for faster detection and elimination if encountered again [9].

Meanwhile, the artificial immune systems, techniques new to the scene of biological inspired computation and artificial intelligence, are based on metaphor and abstraction from theoretical and empirical knowledge of the mammalian immune system. Brownlee [10] stated that "a robust biological process is critical to combating of disease in the body. Furthermore, the immune system is known to be distributed in terms of control, parallel in terms of operation, and adaptive in terms of function, all of which are features desirable for solving complex or intractable problems faced in the field of artificial intelligence".

There are a number of AIS models used in pattern recognition, fault detection, computer security, and a variety of other applications in the field of science and engineering [6]. Most of these models emphasize on designing and applying computational algorithms and techniques using simplified models of various immunological processes and functionalities [11, 12]. Also, AIS has gained increasing interest among researchers in the development of immune-based models and techniques to solve diverse complex computational or engineering problems [13]. Researchers have explored the main features of the AIS mechanisms and exploited them in many application areas. Based on their aspects, some AIS techniques have been found to be more suitable for certain application areas compared to other AIS approaches. It has been found that negative selection models and algorithms are widely used in fault detection and computer security applications utilizing the self/non-self-recognition aspect. Alternatively, the artificial immune network approaches are used in clustering, classification, data analysis and data mining applications. The clonal selection models are used mostly for optimization problems [14]. The Danger Theory Project/Dendritic Cell Algorithm concludes the major AIS approaches that exist in literature, and they are targeted at anomaly detection and computer security applications based on the identification of danger rather than differentiating between self/non-self as highlighted by negative selection algorithm [15].

3 Negative Selection Algorithm

To guard the *self* cells and also eradicate unknown antigens (*non-self* cells), the Negative Selection Algorithm (NSA) inspired by the biological negative selection is equipped with the *self-non-self* discrimination process [16]. The T-cells are involved in the negative selection process, and starts from within the thymus at an immature state. For the T-cells to acquire maturation, they undergo a pseudo-random genetic rearrangement and are exposed to the *self* cells in the host. The T-cells that react to the *self* cells are eliminated via a process called *apoptosis* while those without reaction are granted permission to leave the thymus and circulate all around the body to detect and destroy *non-self* antigens. The result of such a mechanism is that while on the one hand the (released) matured T-cells kill the *non-self* antigens; they are, on the other hand, non-reactive to the *self* (body) cells. Thus, Negative Selection Algorithm (NSA) may be viewed as a mechanism to discriminate the *self* from *non-self* [17]. There exist two types of NSA based on the data representation, which are the string (or binary) negative selection algorithm, and the real-valued negative selection algorithm.

In illumination of the concept of NSA, the research group lead by Stephanie Forrest proposed the immune negative selection algorithm [18]. This first implementation initially used a binary representation for the elements in the self/non-self space. The main idea of the algorithm is to generate a set of detectors which do not harm *self* and distinguish the *non-self* (unauthorized user, virus, etc.) from *self* (authorized users, protected data files, etc.). This algorithm consists of two processes: censoring and monitoring. The censoring phase caters for the generation of mature detectors. Subsequently, the system being protected is monitored for changes by the detectors generated in the censoring stage. The real-valued negative selection algorithms, the focus of this study, are discussed in the subordinate sections.

3.1 Real-Valued Negative Selection with Fixed Detectors

The Negative Selection Algorithm proposition [18] suffers greatly from time complexity as it is exponential to the size of the matching window (the number of bits used to compare two binary strings). In order to tackle these problems, Gonzalez et al. [19] proposed a negative selection algorithm that uses real-valued representation of the self/non-self space. This algorithm, called Real-Valued Negative Selection Algorithm (RNSA) tries to alleviate the scaling issues of binary negative selection algorithms, while it uses various schemes to speed up the detector generation process.

The real-valued negative selection algorithm using fixed sized detectors is based on a pre-specified number of detectors [19]. This is not the best approach, and obviously provides no guarantee that the non-self space is completely covered. However, by selecting a large enough value for the number of detectors, the algorithm is expected to provide adequate results. The input to the algorithm is a set of self samples represented by n-dimensional points (vectors). The algorithm tries to evolve a complement set of points called *antibodies* or *detectors* that cover the non-self space. This is accomplished by an iterative process that updates the position of the detector driven by two goals: (1) Move the detector away from the self points, and (2) Keep the detectors separated in order to maximize the covering of the non-self space.

3.2 V-Detector Negative Selection Algorithm

The first implementation of the real-valued negative selection algorithm [19] generated detectors in which the distance threshold (or radius) was constant throughout the entire detector set. However, the detector features can reasonably be extended to overcome this limitation. Ji and Dasgupta [20] proposed a new scheme of detector generation and matching mechanisms for negative selection algorithms which introduced detectors with variable properties. The algorithm includes a new variable parameter, which is the radius of each detector. The threshold used by the distance matching rule defines the radius of the detectors; the choice of variability becomes paramount as the non-self hemisphere to be covered by detectors exhibit an option to be changeable with respect to its size.

The V-Detector and the RNSA share similar characteristics when the detection phase is concerned. The significant difference is with the detector threshold utilized for the unknown data detection. This is made possible as each detector is now assigned radius which differs from the RNSA having a constant radius for all the detectors. An unknown data is classified as non-self if the minimum distance to any detector is less than the detector variable radius, and else, it is classified as self.

4 Experimental Results and Analysis

Experiments are performed to provide empirical evidence on the comparative study of the real-valued negative selection algorithms; RNSA and V-Detector for anomaly detection problems, with two different anomaly detection techniques;

SVM and KNN. The MATrix LABoratory (MATLAB) is used for the implementation of the algorithms. Datasets have been retrieved from the UCI Machine Learning Repository and Knowledge Extraction based on Evolutionary Learning (KEEL), and they are: Iris plant data, Balance-Scale data, Lenses data, and Hayes-Roth dataset. The data partition for RNSA and V-Detector is based on the class distributions in the datasets. In order to pass the datasets as input for execution in MATLAB, for a two class dataset, the normal class is employed as the training data (100 %) and considered as *self* while the other class as *non-self.* In the case of datasets with three classes as registered in Iris, Balance-Scale, Lenses, and Hayes-Roth datasets, one of the classes is selected as the *self* and the remaining classes as *non-self.* This procedure is repeated for all the classes, which simply means that each of the class is employed as self for training, with others as non-self. For testing, all the data elements are used in classifying either as self or non-self. In all the experiments, 100 % of the training data is used and have an execution of 20 runs each, with the average values recorded. The Euclidean distance in (1) is used to measure the affinities between the detectors and real-valued coordinates. The parameters for RNSA are: detection radius $r_d = 0.1$, adaptation rate $\eta_o = 0.005$, age of the detector $t = 15$, and decay rate $\tau = 15$. Also, for V-Detector, the parameters are: self radius $r_s = 0.1$, estimated coverage $c_0 = 99.98$ % and Maximum Number of Detectors $T_{max} = 1000$.

$$D = \sqrt{\sum_{i=1}^{n} (d_i - x_i)^2} \tag{1}$$

where $d = \{d_1, d_2, ..., d_n\}$ are the detectors, $x = \{x_1, x_2, ..., x_n\}$ are the real-valued coordinates, and D is the distance.

4.1 Performance Evaluation

The target outcome of the simulations is to know the ability of algorithms that can perform best with two evaluations performance in consideration. They are the Detection Rate (DR) and False Alarm Rate (FAR) depicted in (2) and (3). If the algorithms perform well and meet the targets, it can then be applied to new data to detect anomalies in the future.

$$DR = \frac{TP}{TP + FN} \tag{2}$$

$$FAR = \frac{FP}{FP + TN} \tag{3}$$

where *TP* represent the number of *non-self* elements identified as *non-self*; *TN* represent the number of *self* elements identified as *self*; *FP* translate to the number of *self* elements identified as *non-self*; *FN* translate to the number of *non-self* elements identified as *self*.

4.2 Simulation Results

The simulation experiments are carried out using MATLAB (R2011b) on Petium4 Core i5 CPU. Results after series of experiments are tabulated, graphed and discussed.

The results shown in Table 1 illustrate the effectiveness of the anomaly detection techniques on Iris and Balance-Scale datasets. The RNSA and V-Detector generated detection rates of 93.93 % and 98.73 % respectively for the Iris data, with 97.38 % and 100 % for the Balance-Scale data. Their false alarm rates are at lowest minimum with RNSA accounting for a higher rate at 2.09 %. The high accuracy rates reveal that the RNSA and V-Detector are equipped with the capabilities of detecting anomalies. The SVM and KNN generated good detection rates with both reaching their highest rates at 97.30 % and 96.70 % for Iris data, and for Balance-Scale data, rates of 91.70 % and 80 % are produced. Higher false alarm rates are attributed to SVM and KNN at 2.20 % and 11.60 % respectively. With respect to all the algorithms, V-Detector proved to be superior and the graph representation translated in Fig. 1.

The results obtained for the detection rate based on Hayes-Roth dataset varied proportionally to real-valued negative selection algorithms, SVM and KNN algorithms as shown in Table 2 and diagrammatically displayed in Fig. 2. The SVM gained superiority over the V-Detector with detection rate of 86.90 % as against 85.12 % for V-Detector. Same could not be reported for the false alarm rate as the table turned against SVM by yielding a higher positive rate at 8.90 %. A 4.11 % false positive rate is attributed to V-Detector, which coincidentally is the rate for RNSA. For RNSA and KNN, detection rates of 82.65 % and 81.30 % are generated respectively. KNN gave the highest positive rate of 11.90 %.

Table 1 Performances analysis for iris and balance-scale

Algorithms	Iris		Balance-scale	
	Detection rate (%)	False alarm rate (%)	Detection rate (%)	False alarm rate (%)
RNSA	93.93	0.90	97.38	2.09
V-detector	98.73	0.00	100	0.00
SVM	97.30	1.30	91.70	2.20
KNN	96.70	1.70	80.00	11.60

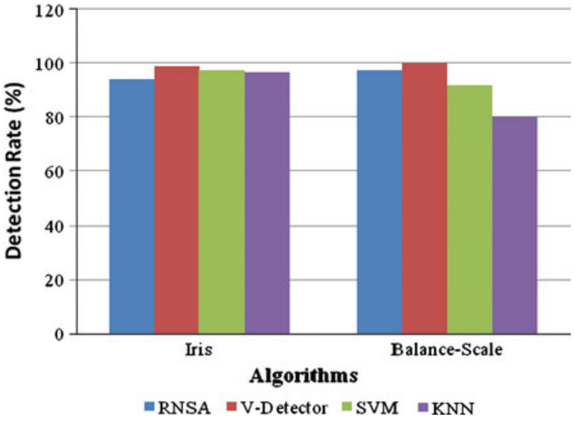

Fig. 1 The detection rates on iris and balance-scale

Table 2 Performances analysis for Lenses and Hayes-Roth

Algorithms	Lenses		Hayes-Roth	
	Detection rate (%)	False alarm rate (%)	Detection rate (%)	False alarm rate (%)
RNSA	99.92	0.00	82.65	4.11
V-detector	100	5.61	85.12	4.11
SVM	83.30	18.60	86.90	8.90
KNN	66.70	32.50	81.30	11.90

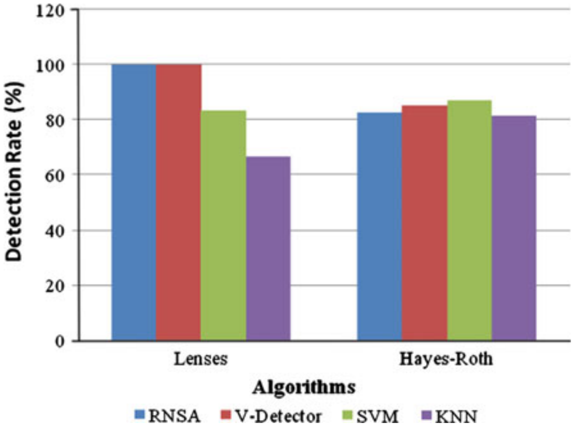

Fig. 2 The detection rates on lenses and Hayes-Roth

Consequently, the performance results for the Lenses data reached the climax of 100 % with V-Detector, followed by RNSA with 99.92 %, SVM with 83.30 %, and lastly KNN yielding 66.70 %. False alarm rates of 0.00 %, 5.61 %, 18.60 % and 32.50 % are recorded by RNSA, V-Detector, SVM and KNN respectively. Overall, the V-Detector surpassed all other algorithms performance wise.

5 Conclusion

The need for ensuring integrity and confidentiality in data has prompted computer scientists and researchers in proffering ways and avenues to adequately secure information. This stem from anomalies or abnormality, and therefore detection improvement requires continuous efforts in many fields, including Artificial Immune System (AIS). For several data, AIS classifiers have proven their ability in classifying successfully those data by revealing the abnormalities therein. As such, this research focuses on using Real-Valued Negative Selection algorithms with focus on fixed detector (RNSA) and variable detector (V-Detector) in classifying different datasets. Two benchmarked algorithms; SVM and KNN are used for comparison and simulated on datasets acquired from standard databases. Their performances are validated with two measuring criteria: detection rate, and false alarm rate, then a comparison carried out based on their performances on the datasets.

Overall, RNSA and V-Detector performed well on the datasets, and with compatible results from the benchmarking algorithms. Meanwhile, V-Detector was the best in terms of detection rate and false alarm rate. Consequently, it can be inferred that the V-Detector yielded more accurate results, provided that the choice of parameters are properly determined and thus affirm the real-valued negative selection algorithms suitability for detecting abnormalities.

Acknowledgements This work is supported by the Office for Research, Innovation, Commercialization, and Consultancy Management (ORICC), Universiti Tun Hussein Onn Malaysia (UTHM), and Ministry of Higher Education (MOHE) Malaysia under the Fundamental Research Grant Scheme (FRGS) Vote No. 1235.

References

1. Patcha, A., Park, J.-M.: An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput. Networks. **51**, 3448–3470 (2007)
2. Lasisi, A., Ghazali, R., Herawan, T.: Comparative Performance Analysis of Negative Selection Algorithm with Immune and Classification Algorithms. Recent Advances on Soft Computing and Data Mining. pp. 441–452. Springer, Berlin (2014)
3. Gonzalez, F.: A study of artificial immune systems applied to anomaly detection (2003)
4. Tripathi, K.K., Ragha, L.: Hybrid approach for credit card fraud detection. Int. J. Soft Comput. Eng. **3**, 2231–2307 (2013)
5. Tuo, J., Ren, S., Liu, W., Li, X., Li, B., Lei, L.: Artificial immune system for fraud detection. Systems, Man and Cybernetics, 2004 IEEE International Conference on. pp. 1407–1411 (2004)
6. Aziz, A.S.A., Salama, M.A., Hassanien, A.E., Hanafi, S.E.-O.: Artificial Immune System Inspired Intrusion Detection System Using Genetic Algorithm. Inform. **36**, 347–357 (2012)
7. Dasgupta, D., Yu, S., Nino, F.: Recent advances in artificial immune systems: models and applications. Appl. Soft Comput. **11**, 1574–1587 (2011)

8. Andrews, P.S.: An investigation of a methodology for the development of artificial immune systems: a case-study in immune receptor degeneracy. University of York, Department of Computer Science, York (2008)
9. Lasisi, A., Ghazali, R., Herawan, T.: Negative selection algorithm: a survey on the epistemology of generating detectors. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Lecture Notes in Electrical Engineering. pp. 167–176 (2014)
10. Brownlee, J.: Artificial immune recognition system (airs)-a review and analysis. Swinburne Univ. Technol. Melbourne, Aust. Tech. Rep. (2005)
11. De Castro, L.N., Timmis, J.: Artificial immune systems: a new computational intelligence approach. Springer, Berlin (2002)
12. Dasgupta, D.: Advances in artificial immune systems. Comput. Intell. Mag. IEEE. **1**, 40–49 (2006)
13. Al-Enezi, J.: Artificial immune systems based committee machine for classification application (2012)
14. Al-Enezi, J.R., Abbod, M.F., Al-Sharhan, S.: Advancement in artificial immune systems: a perspective of models, algorithms and applications. GCC Conference and Exhibition, 2009 5th IEEE. pp. 1–6 (2009)
15. Greensmith, J.: The dendritic cell algorithm. Doctoral dissertation, Nottingham Trent University (2007)
16. Chakraverty, S.: Mathematics of Uncertainty Modeling in the Analysis of Engineering and Science Problems. IGI Global (2014)
17. Dasgupta, D., Nino, F.: Immunological computation: theory and applications. CRC Press (2008)
18. Forrest, S., Perelson, A.S., Allen, L., Cherukuri, R.: Self-nonself discrimination in a computer. In: Proceedings of 1994 IEEE Computer Society Symposium on Research in Security and Privacy. pp. 202–212 (1994)
19. Gonzalez, F., Dasgupta, D., Kozma, R.: Combining negative selection and classification techniques for anomaly detection. In: Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02. pp. 705–710 (2002)
20. Ji, Z., Dasgupta, D.: Real-valued negative selection algorithm with variable-sized detectors. Genetic and Evolutionary Computation–GECCO 2004. pp. 287–298 (2004)