PERFORMANCE ANALYSIS-BASED PARAMETER TUNING OF CLONAL SELECTION ALGORITHM FOR ANOMALY DETECTION (C007)

[1]Sule Aishat Aladi, [2]Oyefolahan Ishaq Oyebisi., [3]Muhammed Bashir Abdullahi

School of Information & Communication Technology

Federal University of Technology

Minna, Nigeria.

[1]*suleaishat990@gmail.com*, [2] *o.ishaq@futminna.edu.ng*, [3]el.bashir02@futminna.edu.ng

Abstract — Intrusion detection has become paramount in the field of network security owing to the fact that network data are being compromised on a daily basis. To this effect, several algorithms have been made available to detect intrusion in the network environment. The Clonal Selection Algorithm (CSA) is one of such algorithms for intrusion detection. Often times, the detection capability of this algorithm are limited by incorrect settings of the parameters involved. Thus, tuning some of the parameters involved in CSA is sacrosanct in determining the performance analysis of the algorithm. Hence, this paper is aimed at tuning the parameters of CSA and analysing its performance for anomaly-based intrusion detection using KDDcup'99 dataset as benchmark for the evaluation. The findings showed that CSA is a good intrusion detection algorithm.

Keywords: *Intrusion Detection; Artificial Immune System; Clonal Selection Algorithm KDDcup'99 Dataset.*

1.0 INTRODUCTION

Artificial Immune System (AIS) algorithms have received much attention from researchers in the past years owing to its increased popularity in solving computational problems. This algorithms is composed of four major algorithms: artificial immune Networks (aiNet); danger theory and Dentritic Cell Algorithm (DCA); Clonal Selection Algorithm (CSA) and Negative Selection Algorithm (NSA) according to Dasgupta, Yu & Nino (2011). Their application areas are in intrusion detection, fault detection, numerical function optimization, image processing, bio-informatics, robotics, web mining etc.

A category of the AIS algorithm which is inspired by the clonal selection theory is the clonal selection algorithm that produces candidate solutions by means of selection, cloning and mutation process. Diverse problems are being solved by this algorithm and it has been reported by Ulutas & Kulturel-Konak (2011) to perform better in cases such as pattern recognition, and function optimization as compared to its counterpart genetic algorithm and neural network.

Literatures regarding CSA have not witnessed its performance evaluation by

tuning the parameters. Hence, this paper is geared towards tuning some parameters of the CSA algorithm to determine their effect on the performance of the algorithm under such performance metrics as true positive rate, false positive rate, precision, recall, F-Measure and ROC Area.

2.0 LITERATURE REVIEW

2.1 Intrusion Detection

With the increase in system complexity, the traditional intrusion detection system such as firewall finds it difficult to provide the system with the needed security. This has given rise to the development of more efficient mechanism for providing the needed protection as a second layer defense. Intrusions are malicious activities that compromise system security; the process of detecting such activities is termed intrusion detection (Farnia, 2017). Intrusion detection systems monitor network traffic for possible malicious activities, raising an alarm when there is any compromise relating to confidentiality, integrity and/or availability of a system resource. There are two classification of intrusion detection approach: misuse detection and anomaly detection

2.1.1 Misuse detection

The misuse detection which is also referred to as the signature-based approach has a predefined rules or patterns (signature) in the database in which identified packets are compared with. Whilst a signature is a pattern or string that corresponds to a known threat or attack; these attack signatures pass specific activity or traffic that is based on known intrusive activity (Liao, Lin, Lin & Tung, 2013). The process usually involves comparing patterns against captured activities for identifying possible attacks. This technique is simple and efficient in the processing of audit data. However, the false positive is minimal in this approach.

2.1.2 Anomaly detection

In network intrusion detection where this work is based, anomaly detection is able to detect attacks that are unknown previously without the need for any programming of the system to signatures of attacks that can possibly occur. Unlike the misuse approach, the anomaly based IDS uses rules or heuristics rather than signature or patterns and is able to detect any compromising activities that deviate from normal system operations. Here, normal profiles are compared with observed events in order to recognize possible intrusions (Liao, Lin, Lin & Tung, 2013).

2.2 Clonal Selection Algorithm

In 1954, immunologist Niels Jerne puts forward her original idea of clonal selection theory which explains how B and T lymphocytes improve their response to antigens. Later in 1958, Joshua Lederberg and Sir Guster reviewed that only one antibody is always produced by the B cell which forms the first evidence for clonal selection theory. The clonal selection theory states that the occurrence of a clonal expansion of the original lymphocytes is triggered by the activation of the original lymphocytes by binding to the antigen and that any clone of the activated lymphocyte with antigen receptors specific to molecules of the body of the organism during the development of the lymphocyte is eliminated. The clonal selection theory forms the basis on which CSA was introduced by Castro & Zuben (2000). The algorithm was later known as CLONALG

(Cai, Gong, Ma & Jiao, 2015) implementing the affinity maturation of immune response and the clonal selection principle.

The clonal selection algorithm is highlighted below (Ulker & Ulker, 2012):

Step 1: Generate a set of antibodies (generally created

in a random manner) which are the current candidate solutions of a problem.

Step 2: Calculate the affinity values of each candidate

solutions.

Step 3: Sort the antibodies starting from the lowest

affinity. Lowest affinity means that a

better matching between antibody and antigen.

Step 4: Clone the better matching antibodies more

with some predefined ratio.

Step 5: Mutate the antibodies with some predefined

ratio. This ratio is obtained in a way that

better matching clones mutated less and weakly matching clones mutated much more in order to reach the optimal solution.

Step 6: Calculate the new affinity values of each

antibody.

Step 7: Repeat Steps 3 through 6 while the minimum

error criterion is not met.

The CSA is useful for recognition of antigen, propagation and discrimination of cell into the memory cell (Fathima, 2017).

2.3 Performance Metrics

The metrics used for the performance analysis of anomaly detection algorithms are: true positive, false positive, false negative, true negative, precision, recall, receiver operating characteristic (ROC) score, and F-measure. It is true positive (TP) when a true and predicted class of the observation is positive. When an instance that is negative is classified as positive, then it is termed a false positive (FP). Similarly, when a negative observation is classified as negative, then it is named a true positive (TN). Finally, if a positive instance is classified as negative, then it is called a false negative (FN).

In the area of anomaly-based intrusion detection, FN shows the attacks that are not detected by the intrusion detection system and FP shows the false alarm rate. Consequently, TP shows the rate of detecting attacks, and TN shows the rate of accepted non-attack observations.

Recall, also known as *sensitivity* or *TP rate*, is the percentage of detected positive instances. When the algorithm detects all positive instances, the recall value will be equal to one (Ting, 2011). This is depicted formally in equation 1.

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

Precision describes the success of an algorithm in detecting real positive observation as depicted in equation 2, (Ting, 2011).

$$precision = \frac{TP}{TP + FP} \quad (2)$$

F-measure is an evaluation model on which the weighted harmonic mean of recall and precision is calculated, as shown in equation 3, F-measure is a compromise between precision and recall. A value close to one indicates that the classifier is proper to use, whereas an F-measure value close to zero, indicates that the classifier has failed in detecting the intrusion, detecting non-attack observations or both.

$$F - Measure = \frac{2}{{}^{1}/_{precision} + {}^{1}/_{recall}} \quad (3)$$

ROC is the most widely used measure to compare the performance of different algorithms. ROC curves are graphical plots which show the trade-off between false positive (FP) and true positive (TN) rates (Diaz, Lopez & Sermiento, 2016). AUC is a portion of a unit square that has a value between 0 and 1. This is depicted in equation 4.

$$AUC = \frac{FPR * TPR}{2} + \frac{(1 - FPR)(1 + TPR)}{2} \quad (4)$$

Where FPR is False Positive Rate and TPR is True Positive Rate respectively. The ROC curve is a 2D plot that shows the TP rate on the Y axis versus the FP rate on the X axis, and they are plotted in a unit square called ROC space.

2.4 Previous Studies

The performance of algorithms are analysed in a number of ways most times comparing the performance of one with another. A study conducted by Ehsan, Hossein & Alireza (2018) using a version of the negative selection algorithm known as Real-valued Negative Selection Algorithm (RNSA) for intrusion detection varied two parameters of the algorithm: the normal radius and the anomaly radius. At each run, different values of the two parameters were used for intrusion detection. After 20 runs, the value 0.2 and 0.2 for normal radius and anomaly radius respectively, provided the optimum performance. The parameters of CSA were varied in the work of Chaudhary & Kumar (2018) and it was found that there was no significant effect in their result except test tolerance which was varied from 0.6 to 1.0; as the tolerance value increases, accuracy and specificity increases while sensitivity decreases. Furthermore, Chan, Prakash, Tibrewal & Tiwari (2013) showed in their work how the accuracy of the clonal selection algorithm for classification (CSCA) is affected by the number of antibodies. In their experiments, they varied

the number of antibodies between 0 and 100 and accuracy was found to decrease as the number of antibodies tends towards 100.

3.0 METHODOLOGY

3.1 Data Acquisition

The dataset utilized in these experiments is the KDDcup'99 dataset downloaded from the UCI repository (http://kdd.ics.uci.edu). The dataset is known to be the most widely used dataset and the only publicly available dataset for anomaly-based intrusion detection since 1999 (Zekrifa, 2012). The 10% of the whole dataset was used due to its large nature. The 10% consist of 494021 connection records with 41 features and a label of either normal or an attack. The 10% was split into two: 70% for training the model and 30% for testing.

3.2 Parameter Tuning

The parameters involved in the algorithm experiment are:

I. Antibody pool size (N): This describes the total number of antibodies maintained in the memory pool and remainder pool.

II. Clonal factor (beta): This parameter is used to scale the number of clones created by the selected best antibodies

III. Selection pool size (n): This describes the total number of best antibodies selected on each iteration, for cloning and mutation.

IV. Number of generations (G): this describes the total number of times that all antigens are exposed to the system.

V. Remainder pool ratio: This is the percentage of the total antibody pool size allocated for the remainder pool.

This paper considers tuning three parameters that have very high effect on the performance of the algorithm which are N, beta, and number of generations. The simulation environment used is Weka platform using the weka.classifiers.immune.clonalg.CLONALG software developed by Castro & Zuben (2002).

3.2.1 Test case 1

Experiments were conducted by varying the three parameters aforementioned one at a time while keeping others constant. The initial values for each parameter are: N = 30, Beta = 0.1, n = 20, number of generations = 10, remainder pool ratio = 0.1 and total replacement = 0. Subsequently, the N value is varied incrementally by 5; Beta by 0.1; and number of generation by 10 on each iteration. Remainder pool ratio and selection pool size were not varied because they have little effect on the performance of the algorithm.

3.2.2 Test case 2

The values that generated the best results in test case 1 were selected and used to carry out another experiments to see the behaviour of the system whether it performed better or worse.

4.0 RESULTS AND DISCUSSION

4.1 Case 1

The result obtained from tuning the antibody pool size is depicted in figure 1 and it shows that as the number of antibody pool size increases, TPR increases and FPR decreases.

Optimum performance of the algorithm is achieved at N = 70.

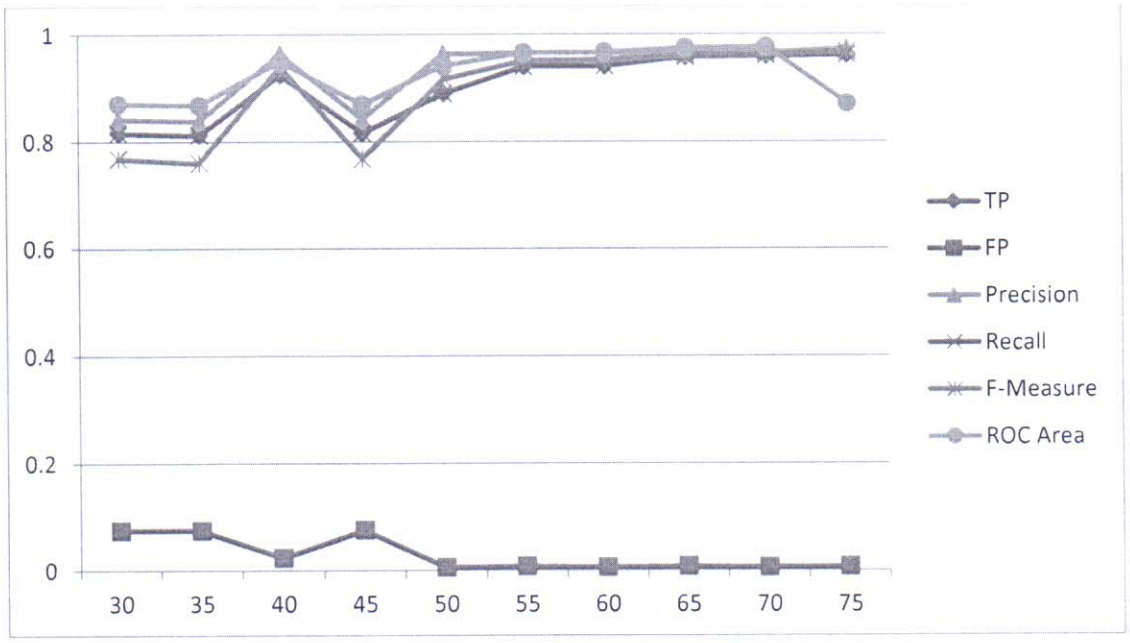

Figure 1: Results obtained from tuning antibody pool size (N).

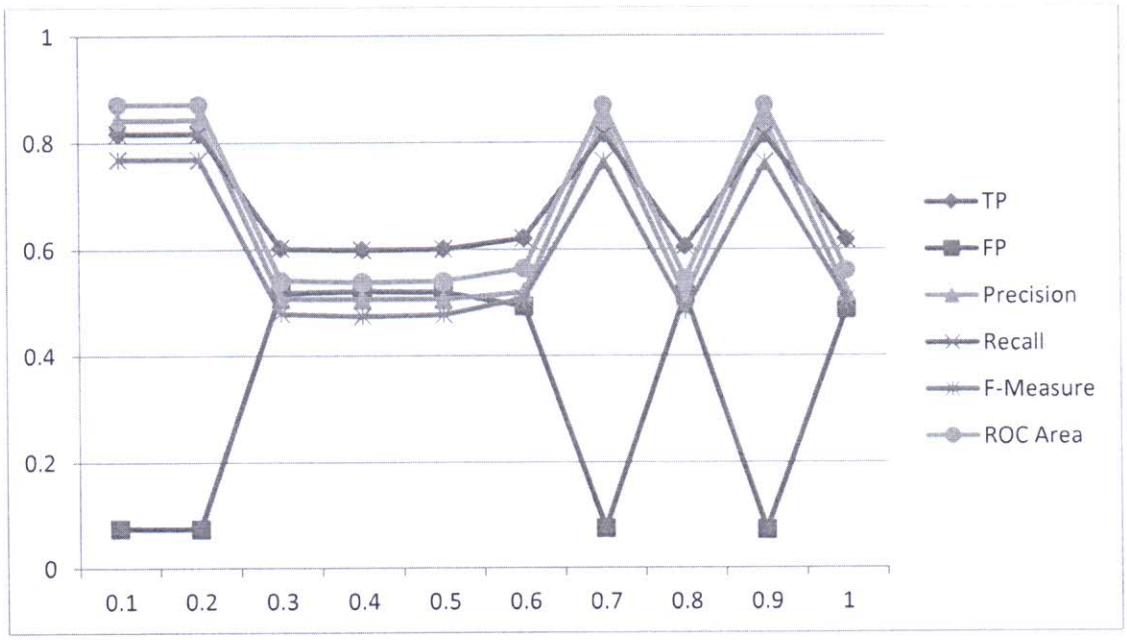

Figure 2: Results obtained from tuning clonal factor (Beta).

The result obtained from tuning the clonal factor parameter is depicted in figure 2. The effect of clonal factor on the algorithm is not stable. At 0.2, it maintained same value of true positive rate while decreasing the false positive rate; along the line, the effect was seen much as it was decreasing true positive rate and increasing false positive rate.

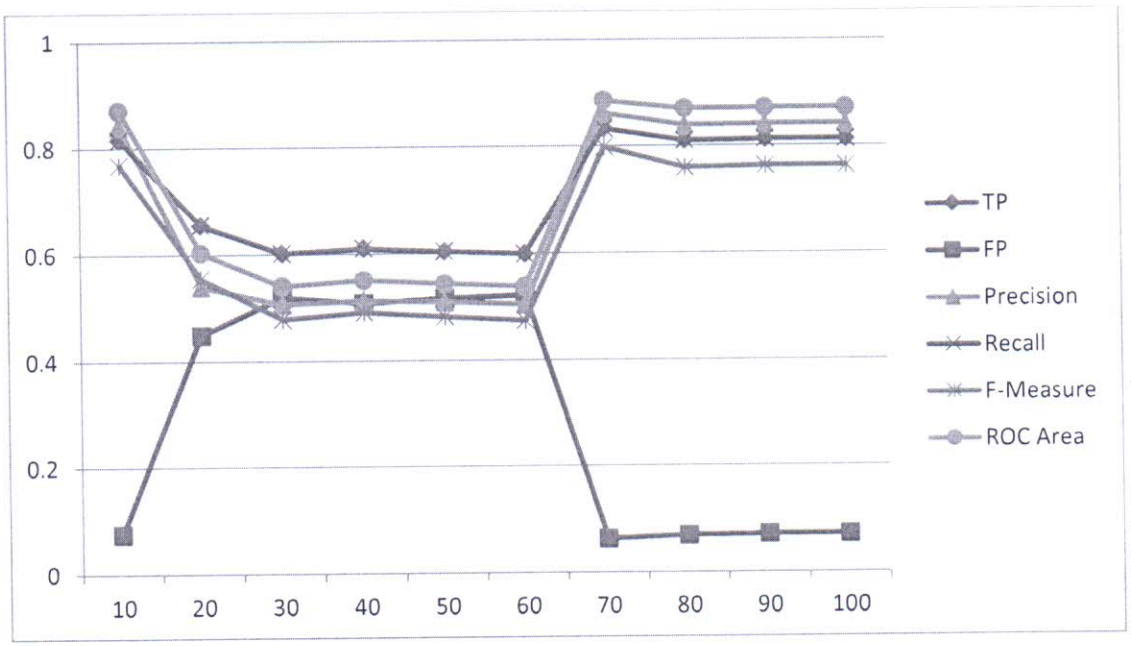

Figure 3: Results obtained from tuning Number of generations (G)

The behaviour of the system is not stable with increase in the number of generation; its effect is seen much in false positive rate as shown in figure 3.

4.2 Case 2

The antibody pool size decreases the performance of the algorithm as it tends towards 100 for a selection pool size of 20. Therefore, the antibody pool size of 70 which achieved the highest true positive rate and lowest false alarm rate was selected for test case 2 experiment. Similarly, the clonal factor of 0.2 generated the highest true positive rate and lowest false positive rate and was selected as the clonal factor value of test case 2. In the same vein, the number of generations of 10 was selected.

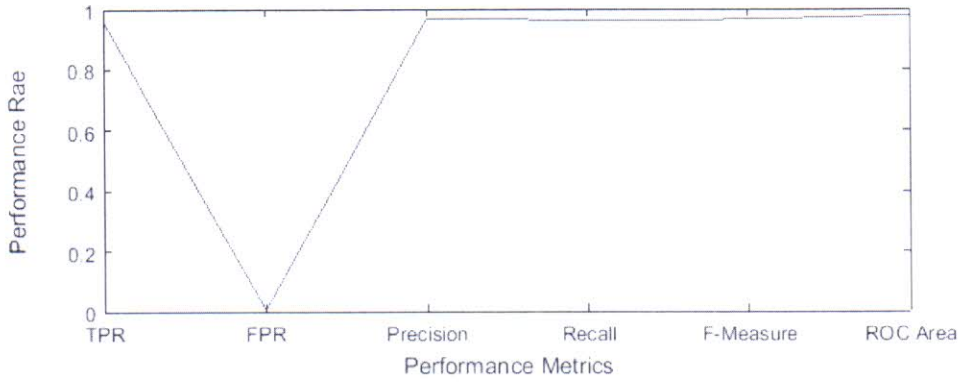

Figure 4: Results obtained from selecting best parameter values from N, Beta and G

Table 1: Results obtained from selecting best parameter values from N, Beta and G

TPR	FPR	Precision	Recall	F-Measure	ROC Area
0.961	0.005	0.968	0.961	0.964	0.978

The system achieved a higher true positive rate and a lower false positive rate as compared to other parameter settings in test case 1 with a true positive rate of 0.961 and a false positive rate of 0.005. The result of the experiment is depicted in figure 4 and summarized in table 1.

5.0 CONCLUSION

The parameters tuned include the antibody pool size, the clonal factor and the number of generations. TPR was seen to increase with increase in N value but later decreased as N tends towards 100 whilst FPR was seen to decrease with increase in N. The effect of Beta on TPR is not stable, it increases at one point and decreases at another point. Same was seen on FPR. The effect of tuning G on FPR and TPR was also not stable. However, a higher TPR and lower FPR were achieved with the following parameter value: N = 70; n = 20; Beta = 0.2; G = 10; remainder pool ratio = 0.1 and total replacement = 0.

The analysis of the performance of CSA has been done in a way that broadens our knowledge on the performance of CSA in intrusion detection and the effect of CSA parameters on each performance metrics investigated. Therefore, it can be concluded that CSA performed considerably well in the detection of intrusion. Future research can look into optimizing the parameters to improve on the detection rate.

REFERENCES

Castro, L. N. D., & Zuben, F. J. V. (2000). The clonal selection algorithm with engineering applications. *In* Genetic and Evolutionary Computation Conference, Las Vegas, Nevada.

Chan, F. T., Prakash, A., Tibrewal, R. K., & Tiwari, M. K. (2013). Clonal selection approach for network intrusion detection. In *Proceedings of the 3rd International Conference on Intelligent Computational Systems, ICICS.*

Cai, Q., Gong, M., Ma, L., & Jiao, L. (2015). A novel clonal selection algorithm for community detection in complex networks. *Computational Intelligence*, *31*(3), 442-464.

Chaudhary, P., & Kumar, K. (2018) Artificial Immune System: Algorithms And Applications Review.

Dasgupta, D., Yu, S., & Nino, F. (2011). Recent advances in artificial immune systems: models and applications. *Applied Soft Computing*, *11*(2), 1574-1587.

Diaz, M., López, S., & Sarmiento, R. (2016). A new comparison of hyperspectral anomaly detection algorithms for real-time applications. In *High-Performance Computing in Geoscience and Remote Sensing VI* (Vol. 10007).

Ehsan, F., Hossein, S., & Alireza, N.,(2018). The Real-Valued Negative Selection Algorithm (RNSA): A Matlab Simulation.

Farnia, F. (2017). *Low-Rate False Alarm Anomaly-Based Intrusion Detection System with One-Class SVM* (Doctoral dissertation, École Polytechnique de Montréal).

Fathima H. (2017) Anomaly Detection in Wireless Sensor Networks using Immune based Bio-inspired Mechanism

Castro, L. N. D., & Zuben, F. J. V (2002) Learning and Optimization Using the Clonal Selection Principle. IEEE Transactions on Evolutionary Computation, Special Issue on

Artificial Immune Systems. 2002; 6(3): 239-251.

Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), 16-24.

Ting K. M. (2011), "Precision and recall." in *Encyclopedia of machine learning*. Springer, 2011, pp. 781–781.

Ulutas, B. H., & Kulturel-Konak, S. (2011). A review of clonal selection algorithm and its applications. *Artificial Intelligence Review*, *36*(2), 117-138.

Ulker, E. D., & Ulker, S. (2012). Comparison study for clonal selection algorithm and genetic algorithm. *arXiv preprint arXiv:1209.2717*.

Zekrifa, D. M. S. (2014). *Hybrid Intrusion Detection System* (Doctoral dissertation, School of Information Technology & Mathematical Sciences).