An Artificial Immune System using Combination of ANN for Detector Construction and Learning Operator for Clonal Selection

Irfan Iqbal

Abstract— Idea of un-known virus detection using Artificial Immune System was introduced at Fourth International Workshop on Synthesis and Simulation of Living Systems. This paper proposes a technique for detecting unknown viruses by using two AIS techniques being combined. It has been realized from the study that combination of the techniques may lead to the better accuracy and efficiency.

Keywords— ANN (Artificial Neural Network), AIS (Artificial Immune System), Virus Detection, Clonal Selection

1. INTRODUCTION

Modern anti-virus programs scan for Virus Signatures to find out some virus-pattern in the ordinary programs. Stealth actions of the modern viruses make difficult for the antivirus software to detect the virus. *Self-Modification, Encryption with a Variable Key, Polymorphic Code and Metamorphic Code* are more dangerous techniques used by the viruses [12]. Using these techniques virus modifies its signature on each infection instance. In this way each infected file contains a variant of the same virus. This is where a known virus becomes *unknown* to ordinary virus detection tools and they fail to detect and remove such viruses. At this stage arises a need to develop some novel techniques which may provide some solution to such problems.

The basic concept of the Biological Immune System (BIS) is based on the capability of the antibodies to discriminate between the self (cells of the own body) and the non-self (foreign substance or antigens). BIS generates a large variety of diverse detectors (B-Lymphocytes and T-Lymphocytes) for the complete and successful detection of the antigens [3].

BIS provides the foundation stone for the Artificial Immune System (AIS). Steps involved in the AIS are:

• *Irfan Iq bal is is currently working as research assistant and Docent in Computer Science Department in Qassim University, Saudi Arabia, E-mail: e.eqbal@qu.edu.sa.*

detectors generation, detectors maturation, detection process, detectors cloning and maturation, immune memory creation. Detectors are randomly created. A selection process selects the suitable detectors and all the undesired detectors are discarded.

Fitness criteria of some detector are its capability to strongly discriminate between self and non-self. Detectors are compared against the self-files and only those detectors survive which have different structure from those of the self-files.

Mature detectors circulate in the veins of the computer system. It results into quick infection prevention. These detectors also undergo a mutation process as well to combat with the variants [3].

2. BACKGROUND AND RELATED WORK

In 1994 and 1996, IBM's Thomas J. Watson Research Center (in Yorktown Heights, New York), the Anti-Virus science and technology group initiated a work on automatic virus detection using an Artificial Immune System (AIS). Proposed system was able to detect and eradicate unknown viruses [1, 2]. Then after in 1997, the same group proposed an immune system for generation of unknown pathogen prescription. Staying within the domain of the AIS, researchers used different intelligent techniques to implement their ideas. These techniques include: Negative Selection, Multi Agent, Chromosome based AIS, Danger Theory, Clonal Selection, Evolutionary Methods, Neural Networks, FSM Hidden Markov Models and Apoptosis [3-11].

3. RESEARCH QUESTIONS

RQ1. Does the combination of "ANN for Detector Construction" and "Clonal Selection with Learning Operator for suitable Detector Selection" improve the results?
RQ1.1. what will be the effect of proposed changes on Unstable Detectors Elimination?
RQ1.2. what will be the effect of proposed changes on Malicious Code Detection?

4. RESEARCH DESIGN

We conducted mixed method study [14] which consist of a Qualitative and a Quantitative part. For quantitative part of research methodology experiment is conducted and to ensure the quality of the outcomes we tried to go along contemporary research in the field of artificial immune system. The results obtained from experiment performed in quantitative part are used to measure the performance of the system and to answer the research questions.

4.1 QUALITATIVE APPROACH

In qualitative research a survey [14] is used for Semi- Structured Interview along with observation to maintain quality of the study. All participating researchers observed each subject individually and the overall system ration. Interviews to selected industry people in the virus detection Research and Development are made.

In this qualitative study subjects were asked to rate from 1-5 against different characteristics of the system.

4.1.1. DATA COLLECTION

A set of semi-structured interviews [14] are conducted by the researchers. During the interviews, participants are encouraged by cross questioning to get good reliable data. The data collected from the interviews is used to maintain the quality of the study. The data collected is the rating of the individuals against following characteristics;

- Accuracy of the proposed system
- Accuracy compared with the isolated techniques.
- Efficiency with respect to time used by the system.
- Stability of the system in terms of error rates and crashs.

4.1.2. DATA ANALYSIS

Data is analyzed by drawing graph and diagrams. Any pattern found in the data is checked for relevance with the results obtained in the experiment.

4.1.3. DATA VALIDATION

The data is validated to remove the different threats that can influence the study. It is made sure that data is interpreted correctly by using interview recordings and cross examination by the researchers.

4.2 QUANTITATIVE APPROACH

The experiment performed consists of following process [13].

- Definition
- Planning
- Operation
- Analysis

4.2.1 DEFINITION

The experiment is performed to measure the accuracy of the system for detecting the unknown viruses.

4.2.1.1 OBJECTIVE

Main aim of our work is to use a combination of two AIS techniques (*"ANN for Detector Construction"* and "*Clonal Selection with Learning Operator*') and analyzing the results.

4.2.1.2 PURPOSE

The purpose of the study is to see the effect of proposed system on virus detection and to obtain the answers to research questions.

4.2.1.3 QUALITY FOCUS

The quality focus of the study is the improvement in the detection rate for proposed system.

4.2.1.4 PERSPECTIVE

The study will benefit the researcher in the field of the artificial immune system and specially those striving to detect unknown viruses. It will also benefit the computer users who like new ways to detect the viruses.

4.2.1.5 CONTEXT

The experiment was performed in a controlled lab environment consisting of nine computers. All computers were of the same specification and fresh operating system was installed. Researchers observed the operation of the systems and switched their positions to make sure each computer got same attention. Each group of computers had different algorithm running on it. Since this technique is combination of two algorithms so, one group of systems executed the new hybrid algorithm.

4.2.2 PLANNING

4.2.2.1. CONTEXT SELECTION

The experiment was performed in a lab setting so, it was not real time and can be said as 'Offline study' [13]. It was controlled by the observation as part of qualitative study.

4.2.2.2 HYPOTHESIS FORMULATION

*Null Hypothesis (H**0**):* There is no improvement in the detection rate of the proposed antivirus system.

*Alternative hypothesis (H**1**):* The proposed antivirus system is improved version of its ancestors.

4.2.2.3 VARIABLE SELECTION

The independent variables of the system are as follows;

- Number of True Positives (TP)
- Number of True Negatives (TN)
- Number of False Positives (FP)
- Number of False Negatives (FN)

The dependent variable of the system is the accuracy of the system, which is given as follows;

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

4.2.2.4 SUBJECT SELECTION

Computers are used as subjects of the experiment. Keeping in view that all the subjects are of same specification, a random sampling is used to select the subjects. After selection of subjects, random selection will be made to divide them into three groups. Each group will run unique algorithm.

4.2.2.5 EXPERIMENT DESIGN

4.2.2.5.1 DESIGN PRINCIPLE

The design of the system is using Balancing [13] as a technique. Each group have same no of computers and all computers are of same specifications. It is also made sure that all systems have fresh operating systems installed on them and have no hardware problem.

4.2.2.5.2 DESIGN TYPE

The design type of the experiment is 'One factor with more than two treatments' and is completely randomized [13].

4.2.2.6 INSTRUMENTATION

The instruments of the experiment performed are as follows;

- Computers
- Infected Data
- Non Infected Data
- Algorithms

Human resource for conducting the experiment is the participating researchers.

4.2.2.7 VALIDITY EVALUATION

There was risk of some validity threats which are mitigated accordingly as mentioned below:

4.2.2.7.1 INTERNAL VALIDITY

- *History threat* to internal validity can influence the experimentation process. This threat could be of the form that system being used was previously virus infected. It is tackled by making sure fresh operating system is installed on each system before it begins. It is also being made sure that no parallel activity be performed during the experiment.
- Maturation threat can take place if systems are executed without rest provided. It is managed by restarting each system every day and also restoring the backed up registry to reflect previous position.
- *Selection threat* can also alter the results and is mitigated by using same set of virus population being tested for each algorithm.
- Instrumentation threat can also be a hurdle in getting the true results and is mitigated by making sure all computer systems are of same specification and swapped after each iteration.
- *Diffusion or imitation of treatments* will not take place since computers are not connected to network.

4.2.2.7.2 CONCLUSION VALIDITY

The reliability of the measure can influence the experiment results so; it is managed by giving the correct input and repeating the experiment multiple times to get mean value which will also reduce the error rate.

4.2.2.7.3 CONSTRUCT VALIDITY

- The *Inadequate preoperational explication of constructs* can not affect the system since construct is made clear which is to measure the overall accuracy of the system during the experiment.
- *Mono-operation bias* is not present since independent variables are more than one.
- Mono-method bias *can not affect the results since experiment are repeated no of times and are cross checked.*
- *Interaction of different treatments* can affect the system since systems are swapped after each round of experiment and are assigned different algorithm.

4.2.2.7.4 EXTERNAL VALIDITY

The threats to external validity are eliminated by making the controlled and realistic lab environment. Generalizablity is attained by applying another test with the random virus being injected into the system and performance monitored accordingly.

4.2.3 OPERATION

All required resources including computers, data and algorithms are prepared for the experiment. Experiment is conducted in the isolated lab environment in working hours without the network connectivity to obtain the controlled output. Collected data is validated to ensure removal of threats. Data from nine computers organized in three groups is collected and is made sure that any incorrectness in the data must not occur due to software or hardware issues.

4.2.4 ANALYSIS AND INTERPRETATION

The data collected is being analyzed by statistical methods. Hypothesis testing is being used to analyze and interpret the data. Parametric test ANOVA [13] (Analysis of Variance) is suitable for design type selected earlier. For non-parametric analysis Chi-2 [13] is being used to determine the normal distribution.

4.2.4.1 DESCRIPTIVE STATISTICS

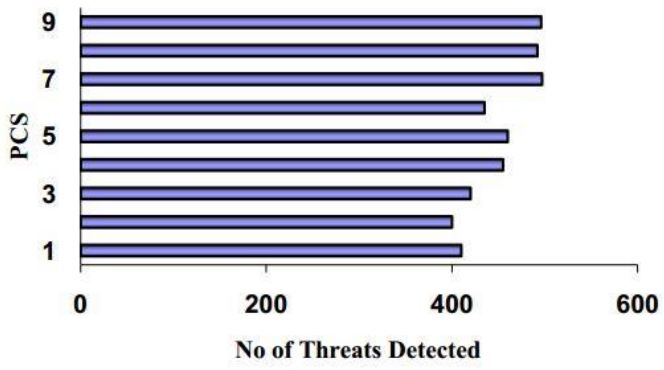
Data is being divided into training and test sets. Test set consists of two sub groups non-self and self. N-test-set represents the non self group and S-test-set represents the self test group. The results are described in Table 3. The ANNDC means Artificial neural network for detector construction, CSA means Clonal selection algorithm and Hybrid the proposed technique being used.

Table 3
Experiment Results

	N-test set	S-test set	Mean Threats (TP)	Self (TN)	Wrong Decision FP+FN	Accuracy TP+TN/ (TP+TN) +(FP+FN)
ANNDC	500	500	410	300	290	71.0%
CSA	500	500	450	400	150	85.5%
Hybrid	500	500	495	437	68	93.2%

The no of threats detected by each subject is presented in the Figure 1. The subjects 1-3 are representing Artificial neural network for detector construction algorithm, subjects 4-7are representing Clonal selection algorithm and 7-9 are representing the proposed technique.

Figure 1



4.2.4.2 HYPOTHESIS TESTING

The results are analysed by parametric test ANOVA and are presented in the table 4.

Table 4
Calculation for ANOVA test

	Sum of Squares	df	Mean Square	Fisher Value (F0)
Between Groups:	10,850.0	2	5,425.0	86.619
Within Groups:	375.782	6	62.630	
Total:	11,225.78	8		

From Table 4 it is concluded that we can reject the null hypothesis for ANOVA since $F_{0.025,2,6} = 7.26$ which is greater than F0. In this case alternative hypothesis is true that means all expected means are not equal. Also since p value is very low so results are highly significant.

The results are also analyzed by non parametric testing method Chi-2. This test is used to compare if measurements from two or more groups come from the same distribution. The results from Chi-2 test are summarized in table 5.

Table 5
Calculation for Chi test
Expected values E_{ij} are displayed in parenthesis

	Group 1	Group 2	Group 3	Combined
ANNDC	400 (409.09)	420 (415.14)	410 (405.76)	R1=1230
CSA	455 (449.003)	460 (455.645)	435 (445.35)	R2=1350
Hybrid	497 (493.90)	492 (501.21)	496 (489.88)	R3=1485
Total	C1=1352	C2=1372	C3=1341	N=4065

Based on the data the test statistics are calculated to X2=0.93. The number of degree of freedom is (r-1)*(k-1)= 2*2= 4. Since $X^2 < X^2_{0.05,2} = 9.49$ it is impossible to reject the null hypothesis.

5. CONCLUSION

The paper proposes a hybrid technique to solve the problem of virus detection using artificial immune system. It is suggested that if combination of these techniques are used detection rate of the system is expected to increase. Data from the nine computer subject is collected and then validated to filter the threats.

A research design is presented by selecting appropriate

methodologies for qualitative as well as quantitative part. Hypotheses are formulated in research design and are further analyzed in the analysis section.

The results obtained by the operation of the experiment are analyzed both by descriptive statistics and hypothesis testing. The experiment and the qualitative part of the study give notion of improvement achieved by the application of this proposed technique. This addresses the research question raised earlier. It can be concluded from the work that combination of two AIS techniques can improve the detection rate of the system.

The possible future work of this could be the application of different clonal selection algorithm to improve detector selection process. The merger of different AIS techniques could also be a prospective area to consider while developing virus detection systems using AIS.

ACKNOWLEDGEMENT

My thanks and sincere appreciation goes to Professor Dr. Jobair bin Suleman Alharbi and professor Dr. Ali Hasan Husien Alahmadi, the finest teachers and mentors we could possibly want. Dr. Jobair provided continuous encouragement throughout my research work.

REFERENCES

[1] S. Hedberg, "Combating computer viruses: IBM's new computer immune system," *IEEE Parallel & Distributed Technology: Systems & Applications*, vol. 4, pp. 9-11, 1996.

[2] J. O. Kephart, "A biologically inspired immune system for computers," Cambridge, MA, USA, 1994, pp. 1309.

[3] S. Bezobrazov and V. Golovko, "Neural networks for artificial immune systems: LVQ for detectors construction," Dortmund, Germany, 2007, pp. 180-184.

[4] K. S. Edge, *et al.*, "A retrovirus inspired algorithm for virus detection optimization," Seattle, WA, United states, 2006, pp. 103-110.

[5] H. Fu, *et al.*, "Multi-agents artificial immune system (MAAIS) inspired by danger theory for anomaly detection," Harbin, Heilongjiang, China, 2007, pp. 570573.

[6] A. Iqbal and M. A. Maarof, "An antigen presenting cell modeling for danger model of Artificial Immune System," Leeds, UK, 2004, pp. 43-4.

[7] S. Kwee-Bo, *et al.*, "Realization of a self-recognition algorithm based on the biological immune system," *Artificial Life and Robotics*, vol. 7, pp. 32-9, 2003.

[8] M. Samy, *et al.*, "An immunological approach to computer viruses detection," *Computing and Information Systems*, vol. 12, pp. 1-12, 2008.

[9] M. M. Saudi, *et al.*, "Defending virus infection through extrinsic apoptosis," Piscataway, NJ, USA, 2008, p. 5 pp.

[10] Y. Ying and H. Chao-Zhen, "A clonal selection algorithm by using learning operator," Piscataway, NJ, USA, 2004, pp. 2924-9.

[11] W. Zejun, *et al.*, "A chromosome-based evaluation model for computer defense immune systems," Piscataway, NJ, USA, 2003, pp. 1363-9.

[12] *P. Szor,* The Art of Computer Virus Research and Defense*, Boston: Addison-Wesley, 2005.*

[13] Wohlin, C., Runeson, P., Host, M., Ohlsson, C. M., Regnell, B., and Wesslen, A. Experimentation in Software Engineering: An Introduction. Kluwer Academic Publishers,Boston, Dordrecht, London, 2000.

[14] Cresswell, W. J. Research Design: Qualitative, Quantitative and Mixed Method Approaches. Second Edition. Thousand Oaks: Sage, 2003.