ISSN 1870-4069

Implementation of a Security Model for Malware Based on Artificial Immune System

Santiago Yip Ortuño[1], José Alberto Hernández Aguilar[1], Carlos Alberto Ochoa Ortiz[2]

[1] Autonomous University of Morelos State, Cuernavaca, Morelos, Mexico

[2] Autonomous University of Ciudad Juárez, Chihuahua, Mexico

alberto.ochoa@uacj.mx
totoy.yip@gmail.com, jose_hernandez@uaem.mx (corresponding author)

Abstract. This research discusses intrusion detection systems based on computer networks and a model for the detection of malware using artificial immune system (AIS). The SIA has three main theories: the clonal selection, negative selection and network theory. This work used the ClonalG algorithm developed by Castro & Timmis (2002) [5] and implemented in Weka 3.6.4 for the intrusions detection in the KDD 1999 database. Preliminary results indicate good results, since was obtained 77.92% accuracy in the classification of threats using CLONALG algorithm, and 92.69% of accuracy by using CLONALG and feature selection of a total of 494,021 processed registers.

Keywords: Artificial Immune System, ClonalG, Intrusion Detection System, Security model.

1 Introduction

1.1 Information Technologies and Security

The growth of information technology has generated a change in the world; currently the use of technology is an obligation to achieve competition, both for companies and organizations of all kinds. However, despite all the benefits that these technologies offer there are many threats, which? Among all those who manipulate ICTs, which eleven materialized cause irreparable damage, ranging from damage to the image of an entity or person, millionaire losses and even loss of freedom or endangering human lives. Some of the best-known cases include Sony Pictures, Home Depot [13], Celebgate, Stuxnet [12] and Ransomware [14].

For these reasons technologies have emerged that help reduce the risks of using these, within these tools are the intrusion detection systems [4]. This research proposes the

evaluation of the intrusion detection system with a bioinspired heuristic known as Artificial Immune System (AIS) [2].

Problem at Hands

Can an artificial immune system (AIS) detect and report intrusions in an institution?

Hypothesis

H. Through an artificial immune system is possible to detect and report intrusions in an institution.

H0. Through an artificial immune system is not possible to detect and report intrusions in an institution.

Structure of the document, in Section 2 we discuss the theoretical framework, in section three the methodology used for this research as well as the description of the algorithm used (CLONALG), the preprocessing of the database, and the results of an experiment using the KDD cup 1999 database [18]. Finally, we present the conclusions, future work and our conclusions.

2 Review of Literature

2.1 Information Technologies and Security

Whitman and Herbert (2011) in his work "Principles of information security" the define information security as: "The Protection of Information and its critical elements including the systems and the hardware that use stored and transmitted the information" [15].

For that reason, the security of the information includes all those mechanisms, controls, devices, best practices, etc., that ensure the 3 basic aspects of the information [8]:

- Availability: The information is available when it is required.
- Integrity: The information should not suffer any type of alteration; the modifications to be made shall be solely done by processes or mechanisms known in the treatment of it.
- Confidentiality: the information will be available to the persons entitled to it.

At the same time, according to [15] an intrusion is defined as:

"The satisfactory accesses to an information system in order to disrupt, modify, remove or damage the information or integrity of the same".

2.2 Intrusion Detection System

Due to the exponential growth of the threats have emerged different types of tools that allow us to their detection and mitigation, among these are the physical and logical tools; within the physical tools there are: Firewalls, content filtering, intrusion prevention system (IPS), among others; within the logic are: Antivirus, Intrusion Detection Systems (IDS), etc. Among of the tools the IDS have had enough popularity to the companies why? An IDS consists of procedures that react to detect patterns of intrusion, this includes all those actions taken by an organization when an intrusion is detected [15] due is difficult to know what an attacker will do as explained in [24]. For this reason, the security of the information includes all those mechanisms, controls, devices, good Practices, etc. that ensure the 3 basic aspects of the information [8] which as mentioned in [25] represents a challenge for cyber law in Mexico and abroad.

2.3 Artificial Immune System

The biological immune system is a collection of molecules with highly evolved procedures that allow the identification and elimination of any substance foreign to the body that protects [2, 3], the artificial immune system is a simulation of the biological functioning of the immune system to perform specific tasks [5, 6].

The artificial immune system [10-11] has three main theories:

Clonal Selection:

- According to the work reported by [1], the clonal selection theory was proposed by Frank Macfarlane Burnet in 1959 in his work "The clonal selection theory of acquired immunity", the main idea of this theory states that cells are able to recognize antigens will be those that will proliferate.

Negative Selection

- Kim & Bentley (2001) in their work “An Evaluation of Negative Selection in an Artificial Immune System”, quoted that the theory of negative selection was proposed by Stephanie Forrest in the year 1994 in his work "Self - no self-Discrimination in a Computer", where the antibodies generated by the immune system reacts only against the antigens by omitting any action against its own cells [16].

Network Theory

- According to [1] the theory of networks in the artificial immune systems was proposed by Niels K. Jerne in 1974 in his work: "Toward a Network Theory of the Immune System", in establishing a network of antibodies that recognize antigens.

3 Methodology

The proposed methodology for this research is called KDD (knowledge data discovery); is based on the works of [7, 17] we briefly explain it in the following figure.

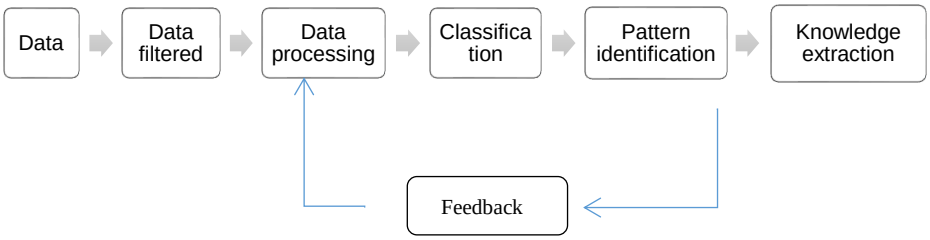


Fig. 1. Cycle generic data mining [7].

1) The test databases to be used in this research are: 1) the available on the Internet for the KDD CUP 1999 tournament [18] (10% of 4 million records).
2) The data will be processed in Weka ClassAlgoss for validation, and presented in tables and graphs.
3) We use SIA algorithms available in the WEKA 3.6.4. [19], the algorithm used in this research is based on the work of [5].
4) According to [1] were run existing SIA algorithms in intrusion detection. For this research is proposed ClonalG:

ClonalG Algorithm [5]

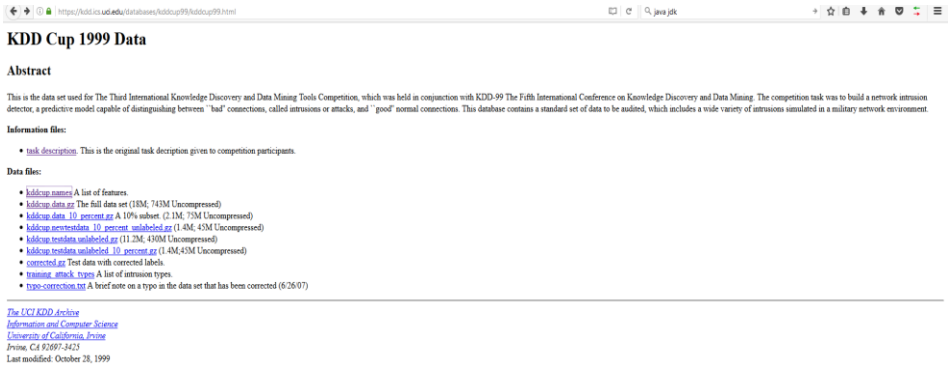
1. Beginning: Initial population created in a random way (P)
2. Antigen presentation: for each pattern, do:

   a. Fitness function: present to the P population and determine its affinity for each element in the P population;
   b. Clonal Selection: select n1 elements with highest fitness function of P and generate clones of these individuals in a proportional way according affinity to antigen: highest the fitness, the greatest number of copies and vice versa;
   c. Maturity of fitness: mutate all these copies with a ratio inversely proportional regarding its fitness function according initial pattern: More fitness, mutation ratio is lesser and vice versa. Add these mutated individuals to P and select the best individual to preserve him as the memory "m" of antigen.
   d. Dynamic goal: replace n2 number of individuals with less fitness function (randomly generated) by the newest;

3. Cycle: Repeat the step 2 until certain criterion is obtained.
5 The classification of the data shall be made using the ClonalG algorithm.
6) The results obtained will be compared among themselves and with other experiments reported in the literature.

Experiment

For the testing of this research was used to the database of the KDD cup 1999, in the competition that year of the KDD is proposed to build a model for the detection with the ability to distinguish between "good" and "bad" connections, the database consists of about 4 million records with 42 attributes each, see Figures 2-3.



KDD Cup 1999 Data

Abstract

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Information files:

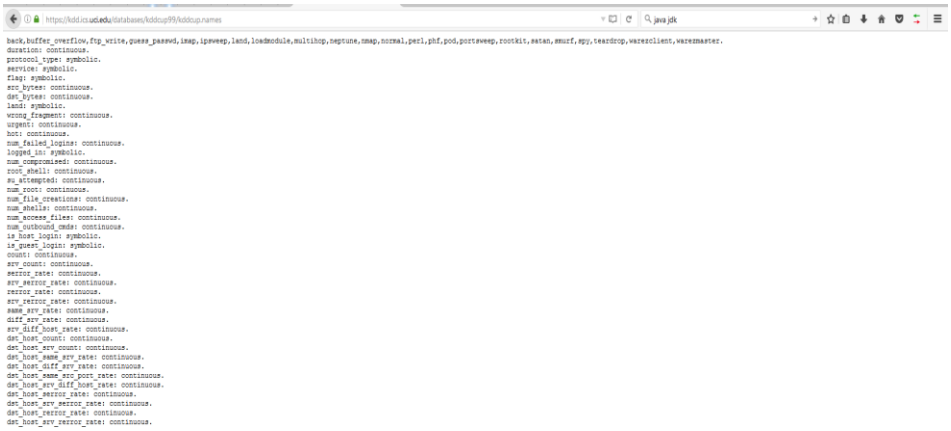
- task description. This is the original task decription given to competition participants.

Data files:

- kddcup.names A list of features.
- kddcup.data.gz The full data set (18M; 743M Uncompressed)
- kddcup.data_10_percent.gz A 10% subset. (2.1M; 75M Uncompressed)
- kddcup.newtestdata_10_percent_unlabeled.gz (1.4M; 45M Uncompressed)
- kddcup.testdata.unlabeled.gz (11.2M; 430M Uncompressed)
- kddcup.testdata.unlabeled_10_percent.gz (1.4M;45M Uncompressed)
- corrected.gz Test data with corrected labels.
- training_attack_types A list of intrusion types.
- typo-correction.txt A brief note on a typo in the data set that has been corrected (6/26/07)

The UCI KDD Archive
Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
Last modified: October 28, 1999

Fig. 2. KDD cup data base [18].



```
back,buffer_overflow,ftp_write,guess_passwd,imap,ipsweep,land,loadmodule,multihop,neptune,nmap,normal,perl,phf,pod,portsweep,rootkit,satan,smurf,spy,teardrop,warezclient,warezmaster.
duration: continuous.
protocol_type: symbolic.
service: symbolic.
flag: symbolic.
src_bytes: continuous.
dst_bytes: continuous.
land: symbolic.
wrong_fragment: continuous.
urgent: continuous.
hot: continuous.
num_failed_logins: continuous.
logged_in: symbolic.
num_compromised: continuous.
root_shell: continuous.
su_attempted: continuous.
num_root: continuous.
num_file_creations: continuous.
num_shells: continuous.
num_access_files: continuous.
num_outbound_cmds: continuous.
is_host_login: symbolic.
is_guest_login: symbolic.
count: continuous.
srv_count: continuous.
serror_rate: continuous.
srv_serror_rate: continuous.
rerror_rate: continuous.
srv_rerror_rate: continuous.
same_srv_rate: continuous.
diff_srv_rate: continuous.
srv_diff_host_rate: continuous.
dst_host_count: continuous.
dst_host_srv_count: continuous.
dst_host_same_srv_rate: continuous.
dst_host_diff_srv_rate: continuous.
dst_host_same_src_port_rate: continuous.
dst_host_srv_diff_host_rate: continuous.
dst_host_serror_rate: continuous.
dst_host_srv_serror_rate: continuous.
dst_host_rerror_rate: continuous.
dst_host_srv_rerror_rate: continuous.
```

Fig. 3. Available attributes in KDD cup data base.

Figure 2 shows the web page where was downloaded the database, in figure 3 are displayed the names of the 42 attributes of the database.

Subsequently, the database was imported through the Excel tool, as shown in figure 4.

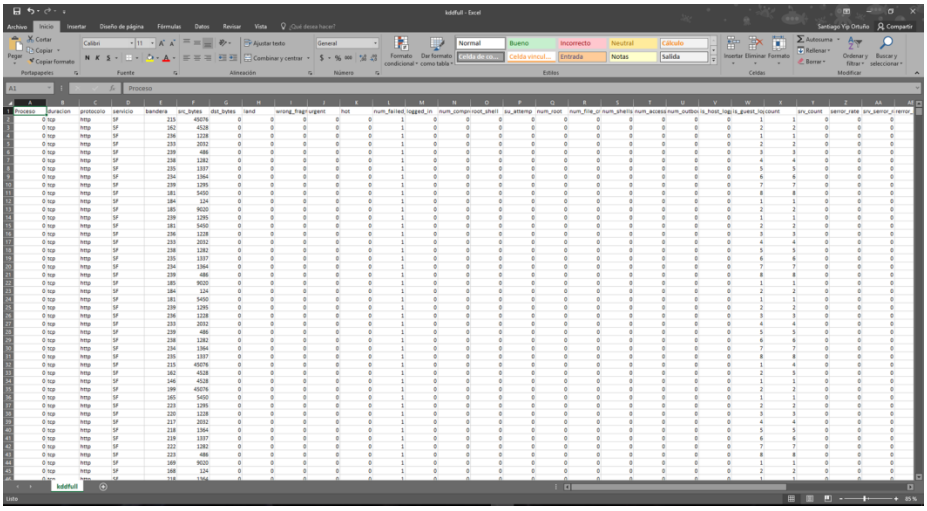


Fig. 4. Cleaning of the information in Excel.

Once preprocessed the BD, we proceeded to the analysis in Weka 3.6.4 running in Linux platform, initially we started the ranking of all the 42 attributes achieving an accuracy of 77%, with an execution time of 91.8 seconds per model see figures 5-6.

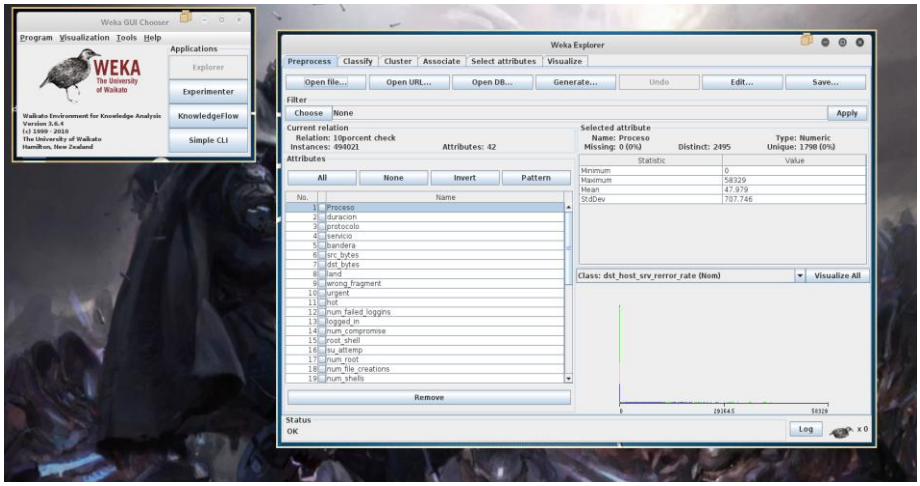


Fig. 5. Load data in Weka.

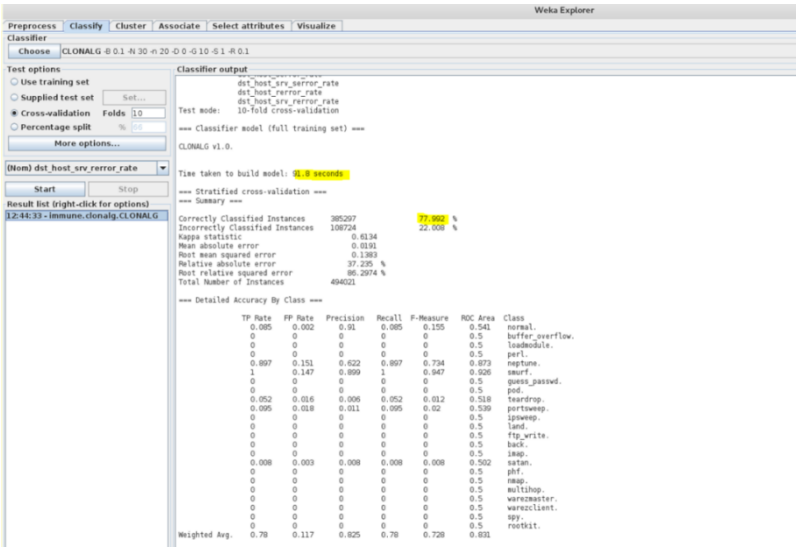


Fig. 6. Classification using CLONAL G, cross-validation 10 Folds.

In order to optimize the execution time and improve the accuracy, a selection feature process was applied, the filter that reached a better accuracy was the GeneticSearch CfsSubsetEval, which allows discarding 34 attributes and improving the accuracy of 77% to 92%, with a running time of 34.18 seconds per model, see Figures 7-8.

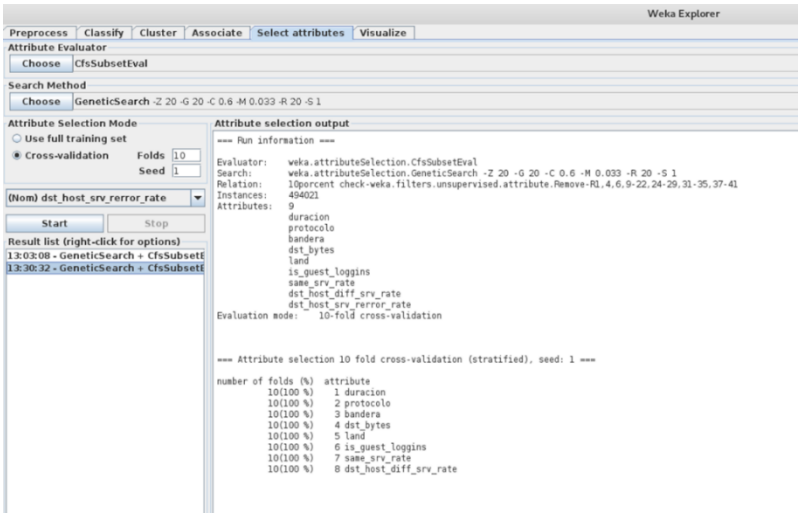


Fig. 7. Feature selection process.

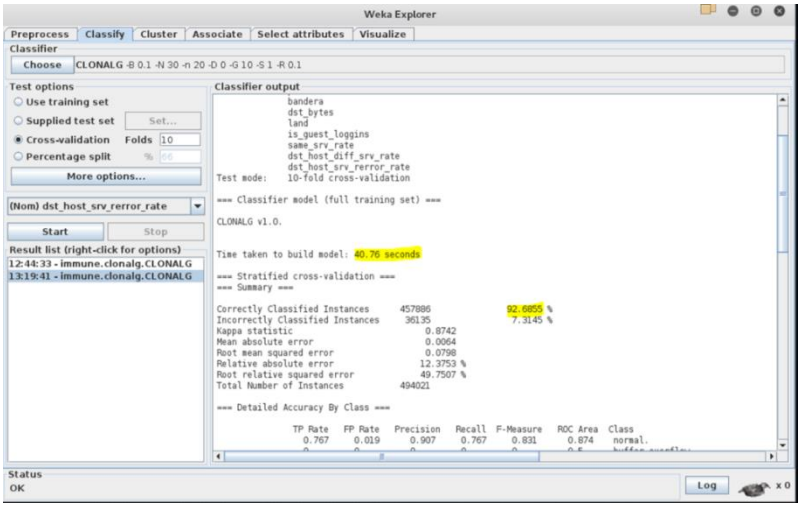


Fig. 8. Data Base Clonal G KDD cup with feature selection.

4 Results and Discussion

In this section are summarized - in a table- a comparative analysis of the experiments carried out in this research, in columns 3 and 4 are shown precision and time for classification process by using CLONALG on a full set of database fields; in columns 4 and 5 Precision and time by using CLONALG and Feature selection process is presented. Columns 6 and 7 show improvements in precision and precision in time.

Table 1. Comparative Analysis for the two models of classification generated.

Database 494021 Records	Precision Clonal G	Time CG (s)	Precision CG + FS	Time CG+FS (s)	Improvement Precision	Improvement Time (s)
KDD cup 1999	77.92%	91.8	92.69%	34.18 Add to	14.77%	51.04

As table 1 show by using feature selection, processing time decrease and precision increase substantially. The use of artificial immune system through the use of the ClonalG algorithm for classification of traffic in a network yielded very good results 92.69% of accuracy. Yan and Yu (2006) in his work "AINIDS: an immune-based network intrusion detection system" obtained 88 per cent of accuracy in classification [20]. Xiaojie Jinquan Zeng et al. (2009) in their work "A self-adaptive negative selection algorithm used for anomaly detection" obtained 88% of correct classification [21]. Itzhak Levin (2000) in the

KDD-99 Classifier Learning Contest [18] obtained 92% of classification [22]. Our model provides a better performance than those related works.

5 Conclusions and Future Work

Artificial Immune System is a promising technology to identify malware and intrusions in computer networks, therefore our hypothesis is true. Artificial Immune System is very important due its adaptive nature regarding other available technologies.

Preliminary results show AIS provides an acceptable technology to identify intrusions regarding similar technologies. Feature selection reduces substantially the time to carry out classification process and increase precision.

The results presented here are not in any way final, they represent only the first steps for the generation of a robust model for intrusion detection system. We want to increase the capabilities of this approach to detect intrusion in smart grids. We would like to compare different SIA algorithms available in literature and with respect traditional algorithms like J48 or a priori. As future work we plan to apply the methodology described above with actual instances of an institute that experience IDS and include comparisons with other well-known techniques like Bayesian Networks [23]. We would like to implement a prototype system in R Language as described in [9].

References

1. Al-Enezi, J. R., Abbod, M. F., lsharhan, S.: Artificial Immune Systems - Models, Algorithms and Applications. Academic Research Publishing Agency (2010)
2. Bachmayer, S.: Artificial Immune Systems. Department of Computer Science, University of Helsinki (2008)
3. Dasgupta, D., Ji, Z., González, F.: Artificial immune system (AIS) research in the last five years. IEEE Congress on Evolutionary Computation, 1, pp. 123–130 (2003)
4. Dario-Duke, N., Chavarro-Porras, J. C., Moreno-Laverde, R.: Smart security. Scientia Et Technica , 1(35) (2007)
5. Castro, L. N., Timmis, J.: Artificial Immune Systems: A New Computational Intelligence Approach. Springer Science & Business Media (2002)
6. Farmer, J. D., Packard, N. H., Perelson, A. S.: The immune system, Adaptation, and Machine Learning. Elsevier Science Publishers B.V., pp. 197–204 (1986)
7. Dave, J., Jian, P., Micheline, K.: Data mining: concepts and techniques. Elsevier (2011)
8. ISO: The portal of ISO 27001 in Spanish. Obtained from what is a http://www.iso27000. es/sgsi.html ISMS? (2012)
9. Torgo, L., Torgo, L.: Data mining with R: learning with case studies. Boca Raton, FL:: Chapman & Hall/CRC (2011)
10. Zum-Herrenhaus, M., Schommer, C.: Security Analysis in Internet Traffic Through Artificial Immune Systems. INTERREG IIIC/e-Bird, Workshop Trustworthy Software, pp. 1–9 (2006)
11. Zum-Herrenhaus, M., Schommer, C.: Healthy-security analysis in internet traffic through artificial immune systems. ArXiv preprint arXiv:0805.0909 (2008)

12. Symantec: El gusano Stuxnet. Obtained from: El gusano Stuxnet, Available at: http://www.symantec.com/es/mx/page.jsp?id=stuxnet (2010)
13. Neal, D.: Home Depot confirms 53 million email addresses stolen in recent hack. Available at: http://www.v3.co.uk/v3-uk/news/2380100/home-depot-confirms-53-million-email-addresses-stolen-in-recent-hack (2014)
14. Kaspersky, K. S. N.: Report: Ransomware in 2014-2016 (2016)
15. Whitman, M. E., Herbert, M. J.: Principles of information security. Cengage Learning (2011)
16. Kim, J., Bentley, P. J.: An Evaluation of Negative Selection in an Artificial Immune System. In: Proceedings of GECCO, pp. 1330–1337 (2001)
17. Hernández-Aguilar, J. A., Burlak, G., Lara, B.: Diseño e Implementación de un Sistema de Evaluación Remota con Seguridad Avanzada para Universidades Utilizando Minería de Datos. Computación y Sistemas, 13(4), pp. 463–473 (2010)
18. Cup, K. D. D.: Dataset: Available at: http://kdd. ics. uci.edu/databases/kddcup99/kddcup99. html 72 (1999)
19. Weka. Available at: https://sourceforge.net/projects/weka/files/weka-3-6/3.6.4/ (2017)
20. Yan, Q., Yu, J.: AINIDS: an immune-based network intrusion detection system. In: Defense and Security Symposium, International Society for Optics and Photonics, pp. 62410U–62410U (2006)
21. Jinquan, Z., Xiaojie, L., Tao, L., Caiming, L., Lingxi, P., Feixian, S.: A self-adaptive negative selection algorithm used for anomaly detection. Progress in natural Science, 19(2), pp. 261–266 (2009)
22. Rojas-Gonzalez, I., García-Gallardo, J.: Bayesian Network Application on Information Security. I. P. Nacional, Ed., Research in Computing Science, 51, pp. 87–98 (2010)
23. Pimentel, J. C. L., Monroy, R.: Formal support to security protocol development: a survey. Computación y Sistemas, 12(1), pp. 89–108 (2008)
24. Argüelles-Arellano, M. D. C.: Challenges of Cyber Law in Mexico. Computación y Sistemas, 20(4), pp. 827–831 (2016)