

**RELATION OF FIREWALLS, MACHINE LEARNING METHODS
AND HIGH-PERFORMANCE COMPUTING SYSTEMS**

Erkan OZHAN¹, Erdem UCAR²

Corlu Vocational School, Namik Kemal University, Tekirdag, 59860, Turkey¹

eozyhan@nku.edu.tr

Department of Computer Engineering, Trakya University, Edirne, 22030, Turkey²

erdemucar@trakya.edu.tr

Abstract

In this study to ensure network security which is one of the important elements were investigated firewalls and consistency of written rules. For this purpose were sampled to investigate data mining, machine learning, and how to use high-performance computing systems. In the first part of the study are explained firewalls operating principles, types and importance and then are indicated the scope and application of data mining science. As in the third section is highlighted the definition and the advantages of machine learning and are introduced Weka software from of machine learning tools. The fourth part of the study are examined using areas and advantages of high-performance computing systems. The fifth part of the study about the relationship between the branches of science relate with subject of study which is shown a sample analysis and application. For the analysis used in the firewall logs to big data volume and this data which data mining, machine learning and is studied how the examined using high performance computing system. As in the last part explained the importance of advantages to provide of interdisciplinary systematic connection.

Key words: Firewalls, machine learning, data mining, parallel computing

1. INTRODUCTION

Firewalls are the primary devices functioning as a bridge between the present-day internet and intranet infrastructures. Although it has different types in terms of function and structure, it basically serves for protecting the information security of an institution or an organization.

Firewalls, one of the primary components of network and information system security, are deployed for the purpose of preventing suspicious or unauthorized access in internet-based companies. The firewalls are the most common security mechanisms employed to provide security for the private networks in many companies and institutions (Hu, Ahn & Kulkarni May-June 2012).

The ability of a computer to connect to an another computer from any place makes a highly advantageous situation, and it is of great fun for individual users at home. Yet, it proves a nightmare for institutional security directors. That is, the release of financial information, product development plans, and commercial secrets can give rise to terrible outcomes for most of the companies. (Tanenbaum 2003).

Machine learning is the programming of computers using the sample data and past experience for the purpose of optimizing a criterion. Human beings collect information and need data and experience as well before finding a direct solution to a problem. In a similar way, these data are transferred to the software for machine learning (Alpaydın 2010).

Data mining, on the other hand, is the analysis of the data in order that the unexpected relationships between the huge data can found and made easy to understand and useful. The relationships and summaries obtained through data-mining searching methods are generally named model or pattern (Hand, Mannila & Smyth 2001).

As to high performance computing, it is the general name for techniques employed to accelerate the operation in the fields such as scientific applications involving complicated calculations (Balladini et al. 2011).

2. FIREWALLS

Firewalls function in two parts which are software-based and hardware-based. Despite having types as hardware and software, they all own a rule repository. The firewalls directs the incoming and outgoing network traffic according to these rules available in their memories, hence it carries great importance that these rules be written correctly.

Recently, the firewall technology has matured in a way that it can coordinate with other firewall and attack-detection systems. These systems can scan viruses in e-mails and harmful codes in web pages. Accordingly, firewalls has become a Standard hardware for internet connections. Home users having dial-up or cable/DSL internet access via a commercial internet service provider also use personal firewalls for their own connections (Wack, Cutler & Pole 2002).

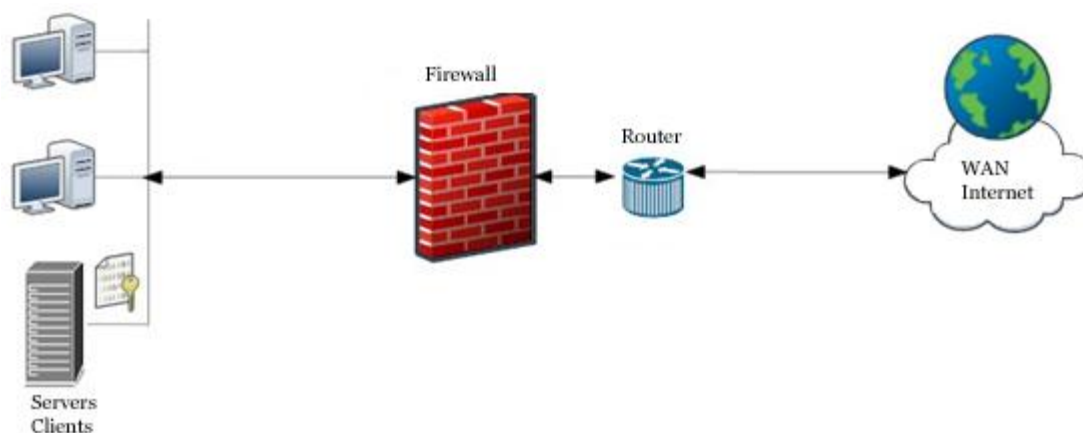


Figure 1. The location of the firewall in the network

1.1 Packet Filter Firewalls

This type of firewalls allows or denies the packet inspecting the title information of the incoming or outgoing packets and they work on the network layer. When millions of packets incoming and outgoing in the network taken into consideration, it can be seen that hardware requirements for this type firewalls are considerably high (Shinder et al. 2003).

Packet filter firewalls can fall short when the security needs of the network to be protected become complicated. Packet filtering rules do not own the ability to test the network against the new types of attacks. If complicated filtering rules and their arrangements are required, it may become difficult to administer the system (Ucan & Osman 2006).

1.2 Application-Proxy Gateway Firewalls

These firewalls have far more advantages when compared to packet filter firewalls. Application-Proxy Gateway firewalls are very strong in inspecting the network address and port-based network connections. Besides this, they make it possible to employ commands special for the application within the network traffic. Another advantage they pose is that they can adapt to the situation regardless of the type of the identity confirmation which is found suitable for the institutional infrastructure. Whereas packet filter firewalls perform identity confirmation according to the network layer address, they have the ability to confirm users' identities directly (Wack, Cutler & Pole 2002).

This type of firewalls have a flexible structure in terms of closing the port connection points users do not need and opening them when needed. They work on the 7th layer, which is application layer of OSI reference model. Therefore, they are the firewalls which are frequently preferred.

1.3 Dual-Homed Gateway Firewalls

Dual-homed gateway firewalls function between two interfaces connected to different networks with two network interface cards. On one side of the network, generally untrusted connection requests come and these firewalls do not allow the access of the unwanted requests coming from an untrusted network into the trusted part of the network. In addition, IP packets are not directed (Tipton & Krause (eds.) 2004).

In the architectural approach of this type of firewalls, two different network interface cards (NIC) are employed instead of controlling the traffic at one point. These interface cards make the traffic between the internal and external networks pass through this firewall. NAT (Network Address Translation) is generally used in this architectural application (Whitman & Mattord 2012).

1.4 Screened Host Firewalls

These firewalls have a more flexible structure than Dual-Homed Gateway Firewalls. However, this situation may pose some disadvantages. In the Dual-homed Gateway Firewalls, an application gateway is placed in the sub network the status of which is protected. This situation gives rise to combination of the application gateway and the router. Filtering and screening of the router control the access in a protocol-based way (Ucan & Osman 2006).

2. DATA MINING

Knowledge Discovery in Databases is modeling the huge databases with the equipments which automatically search and analyze. Knowledge discovery can be understood from great and complicated clusters of data and is a process which organizes determining the useful information. Data-mining forms the nucleus of database knowledge discovery process which consists of algorithm deductions which allow data analysis, model development, and finding previously unknown patterns. Model data are used to solve the events which require predictions with the analysis of the new data (Maimon & Rokach (eds.) 2010).

Data-mining is described as the process of discovering data models. The process should be automatic or (generally) semi-automatic and the model is to bring some meaningful advantages. The data are always available in great amounts (Larose 2006).

Data mining has plenty of applications in a set of fields including marketing, sales, web and e-commerce, law, production and recently national and cyber security. For instance, Data mining can provide a company with the predictions which will help develop its marketing skills and targets through analyzing the past trends available (Abedin et al. 2010).

Firewall logs can also be made meaningful with the Data mining methods. This situation helps increase the security by detecting a lot of situations which the system administrator cannot see even though they cause security violations.

3. MACHINE LEARNING

Engineering techniques may fall short in solutions to some problems. For instance, no model exists for increasing the motivation. Machine learning is needed here. Machine learning generally makes use of test cluster to test its algorithms and a data table and educational data are required to achieve this. Initially, the program is fed with education samples. This operation can be repeated by the operator. The predicted values of the target variable according to the educational data are compared in relation to the algorithm employed and the appropriate algorithm is determined. A great number of methods in which test and education data can be used are available (Harrington 2012).

Firewall logs can be used as test and education data. Nonetheless, before this is, the data obtained should be extracted and put into an analyzable form. Despite there being a lot of means for this, text editors will be needed as well as program languages for rapid and effective extraction.

Machine learning methods can be employed to find and reveal the relationships which are difficult to see among the large heaps of data (Nilsson 1996).

Similarly, there are a large number of programs which perform analyses through the use of machine learning techniques. In the study, GNU machine learning software General Public License, which was developed by Waikato University and has been distributed over the internet has been used.

3.1 *Weka Software*

WEKA software is a software containing the machine learning algorithms which can be easily applied to the data clusters. It was developed at Waikato University, New Zealand and its name was derived from Waikato Environment for Knowledge Analysis. Weka is also a name given to bird which was found on New Zealand Islands and has become extinct. Weka has been developed with Java programming language and can be downloaded for free with the General Public License. (Witten, Frank & Hall 2011).

Weka interface involves methods such as regression, classification, clustering, collocation rules, and attribute choice for the standard data mining problems. Data recognition is a highly important part of Data mining and a lot of data are visualized and subjected to preprocessing. All the algorithms and methods can be applied at this interface with a database query or to the data coming from a single file (Maimon & Rokach (eds.) 2010).

Weka software can receive the data in “.arff, .csv” formats so can it transfer them from a URL or a database. In the study, Weka 3.7.9 and 3.2.3 version which was redesigned by Celis and Musicant so that it could work in a parallel way have been employed. As Weka is a open source code software, extensions have been released by a lot of developers. Of these extensions, Wekaserver has been added to 3.7.9 version and has been utilized. This extension was developed by Pentaho and it is open to everybody. All Weka versions and extensions use Java programming language.

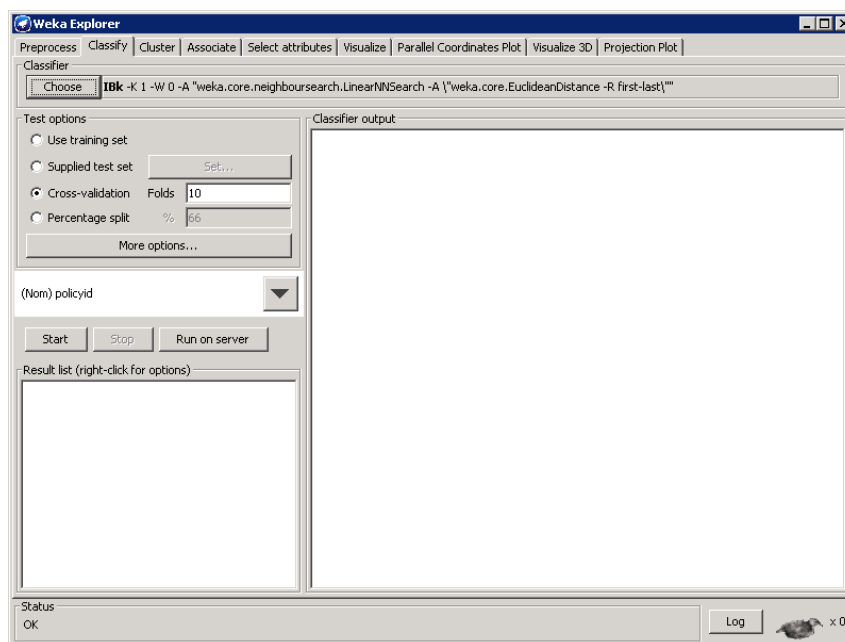


Figure 2. Weka Software Interface

4. HIGH PERFORMANCE COMPUTING (HPC)

Over the last ten years, parallel calculating techniques have turned into high performance computing. At present the use of high performance parallel calculation and parallel application is increasing. These parallel applications are being used in a wide range in search and industry and they are offering user-friendly administration and development means (Foster et al. (eds.) 2011).

There are a lot of pieces of software making use of parallel calculation techniques. In this study, Windows HPC Server software, which was developed by Microsoft, has been employed. Windows HPC server uses two interfaces which are Enterprise and Private. A computer functioning as a bridge connects to Enterprise and Private. There are two interface cards available on this computer. Headnode is the common point to which the computers to be used in high performance computing will connect besides it is a bridge between the two interfaces. Each connection point is called node and the other computers to be used in high performance computing should be loaded with HPC Server. Yet, the other computers are registered to the Headnode computer as members at the last stage of the installation.

After the connection is achieved, the core and RAM information of the other computers can be seen with the aid of the HPC Cluster Manager software of the Headnode computer. The computers connected do not have to have the same features; however, their 64-bit support is of great importance in increasing the success level of the computing.

The applications to be run for the use of HPC systems should also support parallel working. Furthermore, it may be possible to use the programs lacking bus support with the additional software.

HPC systems are employed in the applications calling for the processing of high amount of data and in carrying out the endless calculations. The use of the HPC systems poses various advantages. One of

these advantages is time. This situation has been stated with a sample analysis in the conclusion part of this study.

5. MATERIAL AND METHODS

Each of the disciplines which are subjects of this study has allowed the operations required for the realization of the analysis. The meaningful information, which seems difficult to obtain on its own can be obtained in this way.

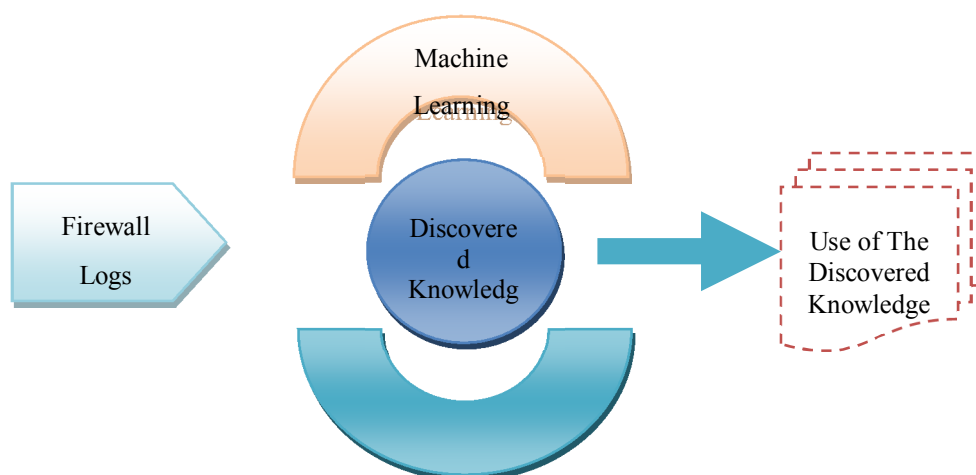


Figure 3. The relationship among Firewall-Machine Learning-HPC

In this study, firewall logs have been employed for the experimental analysis. These logs have been subjected to a set of operations to be organized in raw form in the .csv format. It has been seen that as there is a great amount of data, it is impossible to transfer these data to the Weka software by keeping them in a single file. Thus, MySQL database has been used. In addition, a need to use the additional text editors for the organization of the data.

In the installation of the HPC Server, which has been employed in the study, 13 computers of two different types in terms of Core and RAM have been used. These computers have been administered and observed through HPC Cluster Manager software.

Weka-parallel 3.2.3 and Weka 3.7.9 x64-bit version were loaded to the each of the 13 computers used in the study. Besides, WekaServer extension of Weka 3.7.9 was loaded to each computer as well. Since Weka software is java-based, Java Jre 64-bit version was used.

After meeting the hardware and software requirements, the data of 5 million lines and 17 columns have been analyzed firstly in a single computer and then analyzed with Weka 3.2.3 and WekaServer software of Weka 3.7.9. The time of the analyses has been recorded and then converted to minutes and transferred to the graphics.

6. RESULTS

In this study, it has been seen that the use of high performance computing systems in the analysis of the consistency of firewall rules with Weka software is highly time-saving. This situation is indicated in Figure 3.

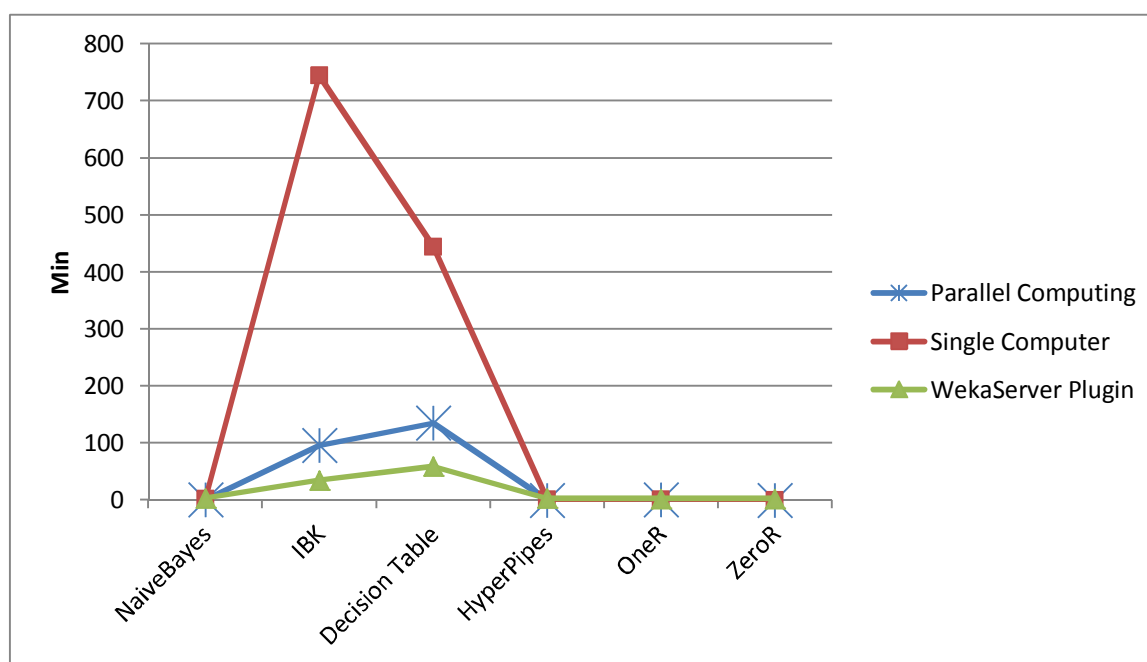


Figure 4. Analysis Time Results

Also, as seen in the figure 3, two different types of parallel calculation techniques have been compared with each other and it has been found that Weka has performed the analyses in a shorter period of time in relation to the other parallel calculation technique.

And another result that we have obtained from this study is that the searches can be concluded more rapidly and more effectively through forming an interdisciplinary bridge.

REFERENCES

- Abedin, M, Nessa, S, Khan, L, Al-Shaer, E & Awad, M 2010, 'Analysis of firewall policy rules using traffic mining techniques', *International Journal of Internet Protocol Technology*, vol 5, no. 1/2, pp. 3-22.
- Alpaydin, E 2010, *Introduction to Machine Learning*, 2nd edn, MIT Press, Cambridge, Massachusetts, London, England.
- Balladini, J, Suppi, R, Rexachs, D & Luque, E 2011, 'Impact of parallel programming models and CPUs clock frequency on energy consumption of HPC systems', *Computer Systems and Applications (AICCSA)*, IEEE, Sharm El-Sheikh, Egypt.

Journal of International Scientific Publications:
Materials, Methods & Technologies, Volume 7, Part 1
ISSN 1313-2539, Published at: <http://www.scientific-publications.net>

Foster, I, Gentzsch, W, Grandinetti, L, Joubert, GR (eds.) 2011, *High Performance Computing*, IOS Press eBook Academic Collection Trial, EBSCOhost, viewed 22 April 2013., Amsterdam.

Hand, D, Mannila, H & Smyth, P 2001, *Principles of Data Mining*, A Bradford Book The MIT Press, Cambridge, Massachusetts London England.

Harrington, P 2012, *Machine Learning in Action*, Manning Publications Co., Shelter Island NY.

Hu, H, Ahn, G-J & Kulkarni, K May-June 2012, 'Detecting and Resolving Firewall Policy Anomalies', *IEEE Transactions On Dependable And Secure Computing*, vol 9, no. 3, pp. 318-331.

Larose, DT 2006, *Data Mining Methods And Models*, A John Wiley & Sons, Inc Publication, New Jersey.

Maimon, O, Rokach, L (eds.) 2010, *Data mining and Knowledge Discovery Handbook*, 2nd edn, Springer, London.

Maimon, O, Rokach, L (eds.) 2010, *Data Mining and Knowledge Discovery Handbook*, 2nd edn, Springer Science+Business Media, London.

Nilsson, NJ 1996, 'Introduction to Machine Learning', TextBook, Robotics Laboratory Department of Computer Science, Stanford University, Stanford.

Shinder, TW, Amon, C, Shimonski, RJ & Shinder, DL 2003, *Best Damn Firewall Book Period*, Syngress Publishing, Rockland, MA, USA.

Tanenbaum, AS 2003, *Computer Networks*, 4th edn, Prentice Hall PTR, New Jersey.

Tipton, HF, Krause, M (eds.) 2004, *Information Security Management Handbook*, 5th edn, Auerbach Publications, Washington, D.C.

Ucan, ON & Osman, O 2006, *Computer Networks and Network Security*, 1st edn, Nobel Yayin Dagitim, Istanbul.

Wack, J, Cutler, K & Pole, J 2002, 'Guidelines on Firewalls and Firewall Policy', *NIST SP Computer Security*, vol 800, no. 41, pp. 1-64.

Whitman, ME & Mattord, HJ 2012, *Principles of Information Security*, 4th edn, Course Technology CENGAGE Learning, Boston, MA USA.

Witten, IH, Frank, E & Hall, MA 2011, *Data Mining Practical Machine Learning Tools and Techniques*, 3rd edn, Morgan Kaufmann Publishers is an imprint of Elsevier, Burlington, USA.