A HYBRID INTRUSION DETECTION SYSTEM BASED ON DIFFERENTMACHINELEARNING ALGORITHMS

Kayvan Atefi[1], Saadiah Yahya[2], Ahmad Yusri Dak[3], and Arash Atefi[4]

[1, 2, 3] *Faculty of Computer and Mathematic science,*
University Technology Mara (UiTM),Shah Alam, Malaysia
k1.educational@gmail.com
saadiah@tmsk.uitm.edu.my
yusri@tmsk.uitm.edu.my
[4]*Electrical Engineering Department, Science and Research Branch,*
Islamic Azad University, Kermanshah, Iran, arash_atefi@yahoo.com

ABSTRACT. Recently, Networks have developed quickly during the last many years, and attacks on network infrastructure presently are main threats against network and information security. With quickly growing unauthorized activities in network Intrusion Detection as a part of defense is extremely necessary because traditional firewall techniques cannot provide complete protection against intrusion. There are numerous study in intrusion detection system (IDS) especially with Genetic algorithms (GA) and Support Vector Machine (SVM) but most of them did not get the potential of hybrid SVM using GA. Hence this study aims to hybrid GA and forbids with high accuracy. The paper illustrates the benefit of hybrid SVM via GA also the paper has proven that by enhancing SVM with GA can reduce false alarms and mean square error (MSE) in detecting intrusion.

Keywords. Intrusion Detection System, IDS, Genetic Algorithm, GA, Support Vector Machine, SVM

INTRODUCTION

Nowadays, with the network development, secured information communication has becoming more vulnerable because of threats from unknown sources and therefore the requirement for secured information assumes greater importance (S. Aneetha, 2012). Attacks on network infrastructure presently are main threats against network and information security. With rapid growth of unauthorized activities in network, Intrusion Detection (ID) as a component of defense is very necessary because traditional firewall techniques cannot provide complete protection against intrusion(Casella, 1998).

According to Aneetha (2012), for secured communication, intelligent analysis needs to be carried out on large volume of online data generated from the network devices. One of the popular ways of securing the network against any external threats has been that of putting in place effective IDS. An IDS is a security control or countermeasure that has the capability to detect misuse and abuse of, and unauthorized access to, network resources. An IDS, in most cases, is a dedicated device that monitors network traffic and detects malicious traffic or anomalies based on multiple criteria (David Burns, 2011).

This paper proposes enhanced anomaly detection based on GA and SVM that increased Accuracy to detect intrusion while using hybrid model (GA and SVM) rather than primary algorithms (SVM). Moreover is to get good percentage of alarms in terms of: False positive,

True positive, False Negative and True Negative when researcher use hybrid model. Recently hybrid of GA and SVM is very novel methods, so that several researches have been working on the combination of GA and SVM to enhance the performance of classification for SVM.

BACKGROUND OF STUDY

The Attack Types and Phases

There are three types of network attacks as follows:

- Reconnaissance
- Access
- Denial of service (DoS)

The first type of attack is to determine the goal of the attack. The second type of attack is to gain access to a system network. Meanwhile, the third and final type of attack is the actual intrusion or attack around the network assets. In the third phase, attack, would come with a DoS or an access attack. Moreover, in a Denial-of-Service (DoS) attack, an adversary attempts to disrupt, corrupt or destroy a network. It reduces or eliminates the network's capacity to perform its expected function (Mohammad Sadeghi, 2012).

Intrusion Detection System (IDS)

Recently, networks have developed quickly during the last many years, and thus possess the techniques that we defend individuals' networks. Typically, IDS happen to be used as a security control or countermeasure to monitor, identify, and inform any unauthorized use, abuse, or misuse of knowledge systems or network assets (David Burns, 2011).

According to Hugo Gascon (2011), Network Intrusion Detection Systems (NIDS) play a simple role on security policy deployment and help organizations in safeguarding their assets from network attacks. NIDS was introduced as a strategy to monitor and identify attacks on vulnerable services. Although ID becoming a comprehensive and promising research area, where anomaly recognition techniques happen to be developed to cope with unknown weaknesses, misuse recognition approaches according to signatures the rules written from intrusion trails would be the current standard in real scenarios. David Burns (2011) stated an IDS is really a security control or countermeasure which has the capacity to identify misuse and abuse of, and unauthorized use of network assets. An IDS, generally, is really a devoted device that monitors network traffic and picks up malicious traffic or anomalies according to multiple criteria. Figure 1 shows how an IDS is usually used.

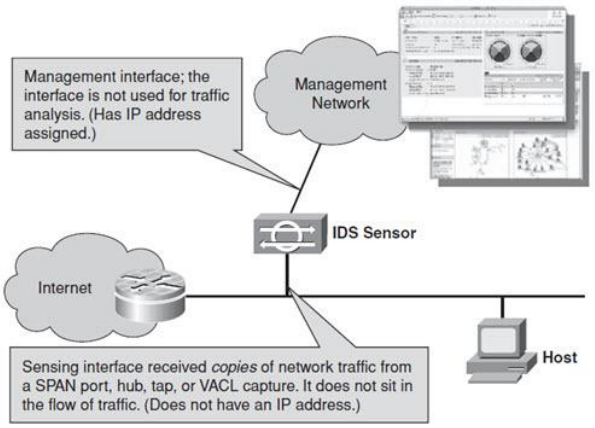

Figure 1. Intrusion Detection System

Tzeyoung Max Wu (2009), presented ID is the act of discovering undesirable traffic on a network or a device. An IDS could be a part of installed software or a physical appliance that monitors network traffic to be able to identify undesirable activity and occasions such as illegal and malicious traffic, traffic that violates security policy, and traffic that violates acceptable use guidelines. Many IDS tools will also store a detected event inside a log to become examined later or will mix occasions along with other data to create choices regarding guidelines or damage control.

Types of Intrusion Detection

According to Sandip Sonawane (2012), There are several types of IDS and the choice of which one to use depends on the overall risks to the organization and the resources available. All the classifications of IDSs were made through the resources they have monitored. Based on Sandip Sonawane (2012) and Yuebin Bail (2003), IDSs are split into two primary types of groups:

- host-based (HIDS)
- network-based (NIDS)

A HIDS resides on the particular host and search for signs of attacks with that host. An NIDS resides on the separate system that watch network traffic, searching for signs of attacks that traverse in the area of the network. The present trend in ID would be to mix both host based and network based information to build up hybrid systems(Sandip Sonawane, 2012).

Intrusion Detection Techniques

According to Earl Carter (2006) and Yuebin Bail (2003), the techniques for the intrusion detection can be divided into two categories:

- Anomaly Intrusion Detection
- Misuse Intrusion Detection

They were classified based on approaches like Statistics, Data mining, Neural Network Based and Self Organizing Maps Based approaches etc. Based on Earl Carter (2006), Anomaly Intrusion Detection the process works using the definition "anomalies aren't normal". It attempts to see whether deviation in the established normal usage designs could be flagged as intrusion. Anomaly recognition technique assumes that the intrusive activities are anomalous and in the other hand Earl Carter (2006), stated that misuse detection is the most common approach used in the commercial IDS. Misuse Intrusion Detection uses the pattern of known attacks or weak spots of the system to complement and identity the attacks. So there will be some methods to represent the attack in the form of pattern or an attack signature to ensure that even the same version of attack could be detected.

Types of Algorithms

- GA (Genetic Algorithms)

Genetic algorithm is really a group of computational model according to the concepts of evolution and natural selection. GA converts the issue right into a model by utilizing chromosomes like data structure and evolves the chromosomes using selection, recombination and mutation operator (Sharmila Devi, 2012)(cisco press, 2012)(Whitley, 1992).

GA starts with a random selected population of chromosomes which signifies the issue to become solved. An assessment function can be used to calculate the "goodness" of every chromosome. The operation begins with a preliminary population of a random produced chromosomes population developed for several generation and each time quality of the

individual progressively enhanced. Three fundamental GA operator are put on every individual i.e. selection, crossover and mutation (Sharmila Devi, 2012).

First of all numerous individual are selected according to user defined fitness function, the rest are discarded. Then the numerous individual are combined with one another. Each pair produces one offspring by using crossover operator. Finally a particular quantity of individual are selected and mutation operator is applied randomly. Figure 2 Shows Genetic algorithm processes (Sharmila Devi, 2012).

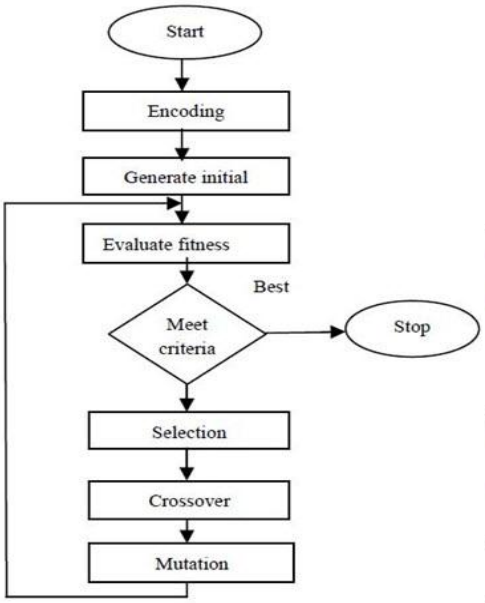

Figure 2. Genetic Algorithm Process

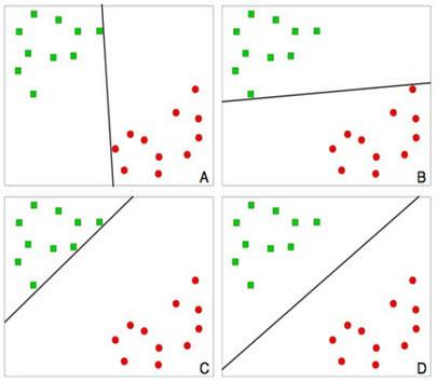

Figure 3. Different Variants of Linear Separation of Two Sets in SVM

Based on Syng-Yup Ohn & Park (2004), GA is among the most effective tools for searching in large search spaces also it imposes couple of mathematical constraints the same shape as the function of optimization. Moreover Syng-Yup Ohn & Park (2004) and Dong Seong Kim (2005) mentioned that GA strategy is used to get the optimal group of features along with the optimal parameters for any kernel function.GA produces enhanced recognition models that contains some features and parameters through the iterative procedure for reproduction.

- SVM (Support Vector Machine)

Guggenberger (2008), stated that for think about a typical classification problem, some input vectors (feature vectors) plus some labels receive. The goal of the classification issue is to calculate labels of recent input vectors to ensure that the mistake rate from the classification is minimal. You will find many algorithms to resolve such type of problems. A number of them require the input information is linearly separable. However for many programs this assumption isn't appropriate. And even when the idea holds, more often than not you will find many possible solutions for hyper plane. Figure 3 illustrates this.

In 1965 Vapnik introduced a mathematical method to solve this type of an optimization problem. The foundation of his approach may be the projection from the low-dimensional training data inside a greater dimensional feature space, because within this greater dimensional feature space it is simpler to split up the input data. Furthermore through this

projection it is possible, that the training data, which could not be separated linearly within the low-dimensional feature space, could be separated linearly within the high-dimensional space.

F. Dataset

The 1998 DARPA Intrusion Detection Evaluation Program was prepared and handled by Lincoln Labs, MIT. The aim ended up being to survey and evaluates data of research in ID. A typical group of data to be audited, including a multitude variety of intrusion simulated inside a military network environment, was provided. The 1999 KDD Intrusion Detection contest utilizes a form of this dataset(Salvatore J. Stolfo, 1998).

According to Mahbod Tavallaee(2009), throughout the final decade, anomaly recognition has attracted the interest of numerous scientists to beat the weakness of signature-based IDSs in discovering novel attacks, and KDDCUP'99 may be the mostly broadly used data set looking for the evaluation of those systems.

Since 1999, KDD'99 continues to be probably the most extremely used data looking for the evaluation of anomaly recognition techniques. This data set is prepared by Stolfo et al. and it is built in line with the data taken in DARPA'98 IDS evaluation program. DARPA'98 is all about 4 gb of compressed raw (binary) tcp dump data of seven days of network traffic, which may be processed into about 5 million connection records, each about 100 bytes. The two weeks of data test have around two million connection records. KDD training dataset includes roughly 4,900,000 single connection vectors because both versions consists of 41 features and it is called either normal or perhaps an attack, with exactly one specific attack type(Mahbod Tavallaee,2009).

SYSTEM DESIGN

Research design is devised in order to keep track of the study progress. Firstly, researcher identified the problem and selected it to be reviewed in the literature section. Then the simulation tools and configuration processes are considered. Finally after getting the results and collecting the data some extra tools for analyzing the results are used in this study.

One of the important purposes of system design is to address practical needs for the system by technical solution. The proposed approach combines several processes together to synthesize a new variation of the embedding and extracting algorithms. In constructing part the appropriate algorithms for IDS will be chosen and tried to investigate IDS based on different algorithms in networks. The last activity in the developmental design is to develop the prototype based on the hybrid algorithms to detect intrusion. The proposed system architecture shown in Fig.4 has two major functions, namely, SVM and GA.

The KDD99 dataset is taken as training data and this data is prepressed by SVM and GA is applied as optimization to the dataset and the trained NID Model has been produced. The training dataset is used to create the rules and the generated signatures are stored for pattern matching. In the testing phase the test dataset has been pattern matched with the IDS model and finally the data are classified into normal attack or abnormal attack.

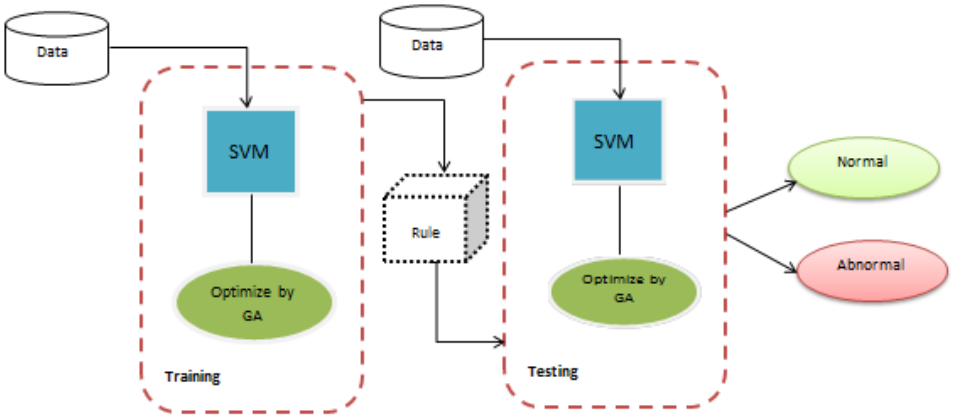

Figure 4. Proposed Architecture of Hybrid System

IMPLEMENTATION DETAILS

In this research, GA and SVM were selected. In additional the researcher use KDD99 Data set for this research that used plenty of data with forty features, which can be used for IDS test. After finishing the preliminary study, the researcher designed the system based on these two algorithms. GA will be used as optimizer of intrusion recognition system. Moreover, in part of classifier, researcher designs SVM that can classify attacks for IDS. Thus in this part researcher designed a system in terms of optimizer and classifier that the classifier is SVM algorithms, which use with GA for intrusion recognition system. In addition after dataset is chosen it should be divided into two parts as a training dataset and testing dataset.

Next section is implementation of the system. Where researcher implements the chosen data set, an algorithm for optimizer and an algorithm as classifier that will work together and identify whether is an attack or normal. For considering and comparing these algorithms together, first of all, researcher implements SVM and get the result then should hybrid two algorithms (SVM and GA) to get the proper results.

In the last section is result and finding. After the implementation of SVM and hybrid model researcher will analyze the data and make result comparison. Moreover, for each implementing of the algorithm will find Accuracy and False Alarms. After analyzing and make result comparison it shows that in hybrid model percentage of accuracy have been increased and also false alarms are in good percentage.

Simulation research tool is being used by the majority of researcher community that estimates how event might occur in the real world (Mehdi Barati ,2012). In this paper MATLAB were used for the implementation.

PERFORMANCE EVALUATION

A. Accuracy

The accuracy is the proportion of the total number of predictions that were correct. In this study the total accuracy is measure of the number of events being predicted as attacks correctly. Figure 5 and Table 1 shows the total accuracy of whole features for all scenarios in terms of SVM and hybrid model. From the table ,the highest accuracy is belong to hybrid model with rates of 98.3333 and the minimum rates is for SVM with rates of 94.8000 which is better than SVM .When GA optimized SVM the accuracy will increase with higher accuracy.

Table 1.Total Accuracy for SVM and hybrid model

Algorithm	Accuracy
SVM	94.8000
GA+SVM	98.3333

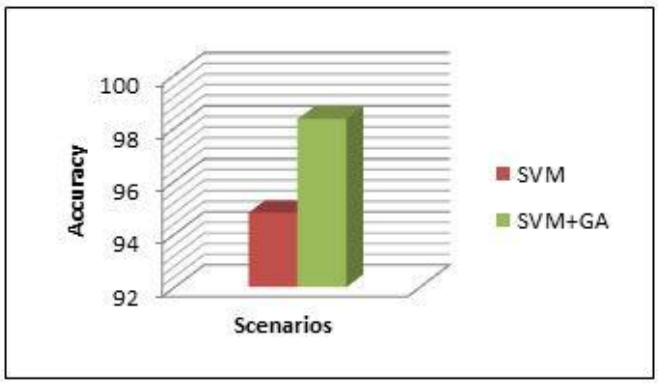

Figure 5. Total Accuracy for SVM and hybrid model

B. False/True Alarms

The following Table 2 and Figure 6 illustrate the false/true alarms of both scenarios in terms of SVM and hybrid model. Based on result in table Figure 6 the maximum true positive is belonging to hybrid model with 99.4987. It means that the rates of attack which is correctly predicted as attack is more than 99 percent, and the minimum true positive is for SVM with rates of 98.7500.Moreover you can see false positive part that the maximum rate recorded for SVM with rates of 2.1429 and minimum rates is for hybrid model with rates of 1.7806 .

In the other hand we have true negative that will indicates the numbers of normal events which successfully labeled as normal, for this part the maximum recorded is for hybrid model with rates of 98.2194 and minimum rates is belong to SVM with rates of 97.8571. In addition last result shows the number of attacks which incorrectly predicted as normal (False Negative) also maximum rate is for SVM with rates of 1.2500 and minimum recorded rates is belong to hybrid model with rates of 0.5013.

Table 2. False/True Alarms for SVM and hybrid model

Alarms / Algorithm	True Positive	False Positive	True Negative	False Negative
SVM	98.7500	2.1429	97.8571	1.2500
SVM+GA	99.4987	1.7806	98.2194	0.5013

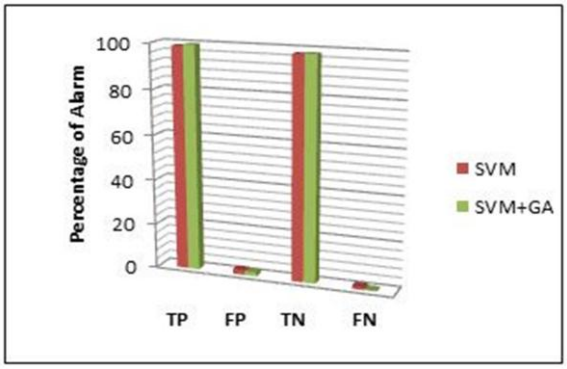

Figure 6. False/True Alarms for SVM and hybrid model

C. The mean squared error (MSE)

The MSE is an estimator among many different ways to evaluate the main differences between values implied by an estimator and also the true values from the quantity being

estimated (E.L. Lehmann and G. Casella, 1998). The following figure of 6 and Table 3 shows the Mean Squared Error for both scenarios in terms of SVM and hybrid model. The maximum MSE is recorded for SVM with rates of 0.0520and the minimum error is belonging to hybrid model with rates of 0.0167.

Table 3. MSE for both scenarios

Algorithm	MSE
SVM	0.0520
GA+SVM	0.0167

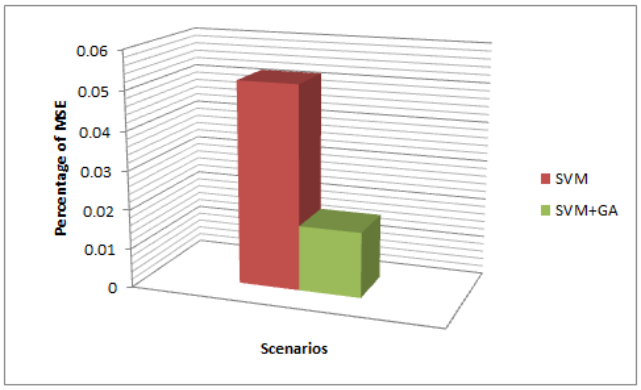

Figure 6. MSE for both scenarios

CONCLUSION

In the summary of this study, the most representative parameters for evaluating performance of IDS are selected which are: Accuracy and false/true alarms. These parameters were compared between primary algorithms (SVM) and hybrid model (SVM+GA). Achieved result is shown in clear figures and tables.

In this study, the researchers achieved a general understanding of algorithms and also review the literature of algorithms in network. The researcher then design and implement the model to perform hybrid model with different metrics. Finally, the results demonstrated by the hybrid model compared with other primary algorithms that have better performance in detecting intrusion.

The result have high Accuracy to detect intrusion while using hybrid model rather than primary algorithms, also result shows good percentage of alarms in terms of: False positive, True positive, False Negative and True Negative when researcher use hybrid model.

REFERENCES

Barati, M., Atefi, K., Khosravi, F., Daftari, Y.A. . (June 2012). *Performance evaluation of energy consumption for AODV and DSR routing protocols in MANET.* Paper presented at the International Conference on Computer & Information Science (ICCIS), Kuala Lumpur.

Carter, Earl. (2001). *Cisco Secure Intrusion Detection System* (Vol. 1). 800 East 96th Street, Indianapolis, Indiana 46240: Pearson Education, Cisco Press.

Casella, E.L. Lehmann and G. (1998). *Theory of Point Estimation* Springer Texts in Statistics Vol. 2nd ed. (pp. 590 p). doi:10.1007/b98854

David Burns, Odunayo Adesina, Keith Barker. (November 4, 2011). *CCNP Security IPS 642-627* (Vol. 1). Indianapolis, IN 46240 USA: Cisco Press.

Dong Seong Kim, Ha-Nam Nguyen, Jong Sou Park. (30 March 2005). *Genetic Algorithm to Improve SVM Based Network Intrusion Detection System*. Paper presented at the 19th International Conference on Advanced Information Networking and Applications AINA 2005.

Earl Carter, Jonathan Hogue. (2006). *Intrusion Prevention Fundamentals* (Vol. 1st). 800 East 96th Street, Indianapolis, Indiana 46240: Pearson Education, Cisco Press.

Guggenberger, Andre. (2008). Another Introduction to Support Vector Machines. Retrieved from http://mindthegap.googlecode.com/files/AnotherIntroductionSVM.pdf

Hugo Gascon, Agustin Orfila, Jorge Blasco. (November 2011). Analysis of update delays in Signature-based Network Intrusion DetectionSystems. *Computers & Security, 30*(8), 613-624.

M. Sadeghi, F. Khosravi, K. Atefi, M. Barati. (2012). Security Analysis of Routing Protocols in Wireless Sensor Networks *International Journal of Computer Science Issues, 9*, 465-472

Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. (2009). *A Detailed Analysis of the KDD CUP 99 Data Set*. Paper presented at the The Second IEEE international conference on Computational intelligence for security and defense applications CISDA'09, IEEE Press Piscataway, NJ, USA ©2009.

Aneetha, S., Indhu, T.S. & Bose, S. (2012). *Hybrid Network Intrusion Detection System Using Expert Rule Based Approach.* Paper presented at the CCSEIT '12 Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology Pages 47-51,ACM New York, NY, USA ©2012.

Salvatore J. Stolfo, Wenke Lee (1998). *Data Mining Approaches for Intrusion Detection.* Paper presented at the Proceeding SSYM'98 Proceedings of the 7th conference on USENIX Security Symposium San Antonio, Texas.

Sandip Sonawane , Shailendra Pardeshi and Ganesh Prasad (2012). A survey on intrusion detection techniques. *World Journal of Science and Technology, 2*(3), 127. doi: 80160124

Sharmila Devi, Ritu Nagpal. (2012). Intrusion Detection System Using Genetic Algorithm-A Review. *International Journal of Computing & Business Research*.

Sinclair, Chris, Lyn Pierce, and Sara Matzner. (1999). *An Application of Machine Learning to Network Intrusion Detection.* Paper presented at the Proceeding ACSAC '99 Proceedings of the 15th Annual Computer Security Applications Conference

Syng-Yup Ohn, Ha-Nam Nguyen, Dong Seong Kim,Jong Sou Park. (2004). *Determining Optimal Decision Model for Support Vector Machine by Genetic Algorithm* (Vol. 3314). Springer Berlin Heidelberg: Springer Berlin Heidelberg.

Whitley, Darrell. (1992). *Foundations of Genetic Algorithms and Classifier*. Morgan Kaufmann Publishers Inc., 297-318.

Wu, Tzeyoung Max. (2009). Intrusion Detection Systems. *Information Assurance Technology Analysis Center (IATAC)*.

Yuebin Bail, Hidetsune Kobayashil (March 27-29, 2003). *Detection Systems: Technology and Development*. Paper presented at the 17th International Conference on Advanced Information Networking and Applications (AINA'03), Xi'an, China.