# Genetic Algorithm to Improve SVM Based Network Intrusion Detection System

**Conference Paper** · April 2005

DOI: 10.1109/AINA.2005.191 · Source: IEEE Xplore



**3 authors**, including:

Dan Dongseong Kim
The University of Queensland
**223** PUBLICATIONS **4,412** CITATIONS
SEE PROFILE

Ha-Nam Nguyen
**4** PUBLICATIONS **195** CITATIONS
SEE PROFILE

# Genetic Algorithm to Improve SVM Based Network Intrusion Detection System

Dong Seong Kim, Ha-Nam Nguyen, Jong Sou Park
*Dept. of Computer Engineering, Hankuk Aviation University, Seoul, KOREA*
*{dskim, nghanam, jspark}@hau.ac.kr*

## Abstract

*In this paper, we propose Genetic Algorithm (GA) to improve Support Vector Machines (SVM) based Intrusion Detection System (IDS). SVM is relatively a novel classification technique and has been shown higher performance than traditional learning methods in many applications. So several security researchers have proposed SVM based IDS. We use fusions of GA and SVM to enhance the overall performance of SVM based IDS. Through fusions of GA and SVM, the "optimal detection model" for SVM classifier can be determined. As the result of this fusion, SVM based IDS not only select "optimal parameters" for SVM but also "optimal feature set" among the whole feature set. We demonstrate the feasibility of our method by performing several experiments on KDD 1999 intrusion detection system competition dataset.*

## 1. Introduction

This paper proposes a fusion method to improve Support Vector Machines (SVM) based Intrusion Detection System (IDS) by optimizing "*detection model*" in network intrusion detection system through the operation of Genetic Algorithm (GA). SVM is relatively a novel classification technique and has been shown higher performance than traditional learning methods in many applications such as bioinformatics and pattern recognitions [17]. In network security field, therefore, several researchers have adopted SVM to IDS. Fugate *et al.* [4] have exploited SVM for computer intrusion detection. Mukkamala *et al.* [12] have also adopted SVM to network-based IDS and compared its performance to neural network-based IDS; the results show us that SVM gives better performance than neural network in terms of processing capacity and accuracy. Kim and Park [9, 15] have proposed host and network-based IDS using two classes SVM and data mining techniques. Moreover, Hu *et al*. [5] have suggested host-based anomaly detection method using robust SVM. Meanwhile, in views of modeling IDS, these approaches are analogous to each other. Although SVM based IDSs improved the performance of IDS in terms of detection rates and speed of processing, there are still rooms for improvement. Recently, as networks become faster, there is an emerging need for security analysis techniques that will be able to keep up with the increased network throughput [10]. Moreover, when the numbers of features of audit data become very large, the detection rates of IDS can be degraded, since it should process the large number of features of vast amount of audit data. So Mukkamala *et al*. [13] have tried to find out important features among the whole features of audit data to minimize processing burden and maximize detection rates.

This paper proposes an enhanced anomaly detection model based on fusions of GA and SVM. The fusion of GA and SVM is not only able to select *"optimal feature set"* but also is able to figure out *"optimal parameters"* for SVM classifier. Since GA is one of the most powerful tools for searching in large search spaces and it imposes few mathematical constraints in the shape of the function to optimize [11]. Actually combination of GA and SVM is very novel methods, so that several researches have been working on combination of GA and SVM to enhance the performance of classification for SVM [1, 3]. By using GA for SVM, proposed system is capable of figuring out "*optimal detection model"* for network intrusion detection system, for that reason, our method is able to minimize the overheads and maximize performance of SVM classifier in network intrusion detection system. The fusion of GA and SVM provides higher detection rates than the system only applies SVM to IDS [4, 5, 9, 12, 15]. Therefore, our system is able to enhance the overall performance of SVM-based IDS. We perform several experiments on proposed system using KDD 1999 CUP dataset used in intrusion detection competition [7, 8].

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05)
1550-445X/05 $20.00 © 2005 **IEEE**

## 2. Proposed Method

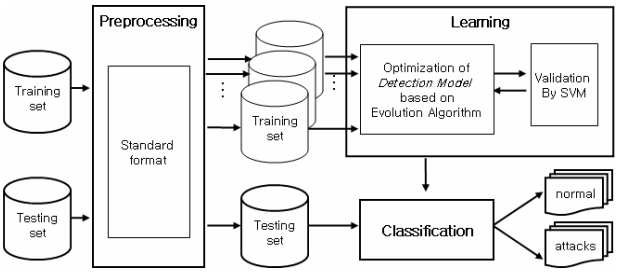


**Figure 1. Overall Structure of Proposed Method**

The overall structure and main components of proposed method are depicted in Figure 1. GA builds new chromosomes and searches the *optimal detection model* based on the fitness values obtained from SVM classifier. A chromosome is decoded into a set of features and parameters for a kernel function, which are used by SVM classifier. The SVM classifier is used to evaluate the performance of a *detection model* represented by a chromosome. In SVM classifier, *n*-way cross-validation is used to prevent overfitting problems, and the detection rates from *n* tests are averaged to obtain a fitness value.

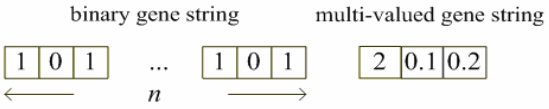


**Figure 2. Structure of a chromosome used in operation of GA**

According to *no free lunch theorem* [2] on machine learning, there is no superior kernel function in general, and the performance of a kernel function rather depends on applications. Also, the parameters in a kernel function play the important role of representing the structure of a sample space. The optimal set of the parameters maximizing the detection performance can be selected by a machine learning method. In a learning phase, the structure of a sample space is learned by a kernel function, and the knowledge of a sample space is contained in the set of parameters. Furthermore, the optimal set of features also should be chosen in the learning phase. In our method, GA technique is exploited to obtain the optimal set of features as well as the optimal parameters for a kernel function. Simulating a genetic procedure, GA creates improved detection models containing a set of features and parameters by the iterative process of reproduction, evaluation, and selection process (see Figure 2). At the end of learning stage, the *optimal detection model* consists of a set of features and parameters for a kernel function. The optimal detection model is obtained after learning phase to be used to classify new pattern samples in classification phase [14].

## 3. Experiments

### 3.1. Description for Dataset

We used the KDD Cup 1999 data [7], and Stolfo *et al.* [8] defined higher-level features that help in distinguishing normal connections from attacks. Total numbers of derived features are 41. We randomly extracted instances for learning set, validation set and testing set from kddcup.data.gz (training dataset) and corrected.gz (testing set) respectively. We only used Denial of Service (DOS) type of attacks out of whole kinds of attacks [14].

### 3.2. Experimental Settings

In learning and validation phase, we randomly split the set of labeled training sample (kddcup.data.gz) into two parts: one set is learning set, which is used as the traditional learning set for adjusting model parameter in the SVM. The other set, validation set, is used to estimate the generalization error during learning step. Our ultimate goal is to decrease generalization error and improve detection rates. We used simple generalization method: 10-fold cross validation with 2500 samples to reduce overfitting problem [2]. After completing cross validation (in other word, learning and validation), we perform classification to decision model constructed as the result of cross validation by using testing set (corrected.gz).

### 3.3. Experimental Results

We applied GA to figure out optimal decision model of SVM. The result of validation test is depicted in Figure 3. Because of characteristics of GA, the detection rates monotonously increase when all of three kernel functions are used in learning phase. While GA was executed for 20 generations, the learning process using SVM with neural kernel function achieved the highest detection rates among the kernel function compared.

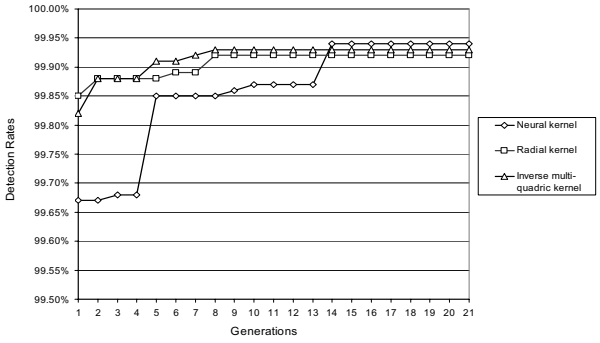


**Figure 3. Detection Rates vs. Generations.**

The result of classification of testing data is summarized and shown in Figure 4. Most of the classification problems using SVM, radial kernel is selected and has showed a good performance, however, in our experiments indicate that neural kernel function for SVM were shown the best results. According to *no free lunch theorem* [2] on machine learning, there is no superior kernel function in general, and the performance of a kernel function rather depends on applications.

In summary, proposed system shows the best result when it couples GA and neural kernel function for SVM. And it also shows that this system can cope with novel attacks very well since it shows detection rates more than 99%. Our system is not only able to select "*optimal feature set*" for audit data but also figure out *"optimal parameters"* for a kernel in SVM classifier. Therefore, our system minimized the number of features that SVM should process and maximize the detection rates for network intrusion detection system suitable to the network environments. And our system provides higher detection rates than the system that only adopts SVM for IDS. The detection rate of proposed system sometimes shows the better performance than the KDD '99 contest winner [16]. These results demonstrate the feasibility of our proposed method. Moreover, when the numbers of features of audit data become very large and networks become faster there is an emerging need for security analysis techniques that can keep up with the increased network throughput. In this viewpoint, it is also remarkably important to optimize detection model in network intrusion detection system.

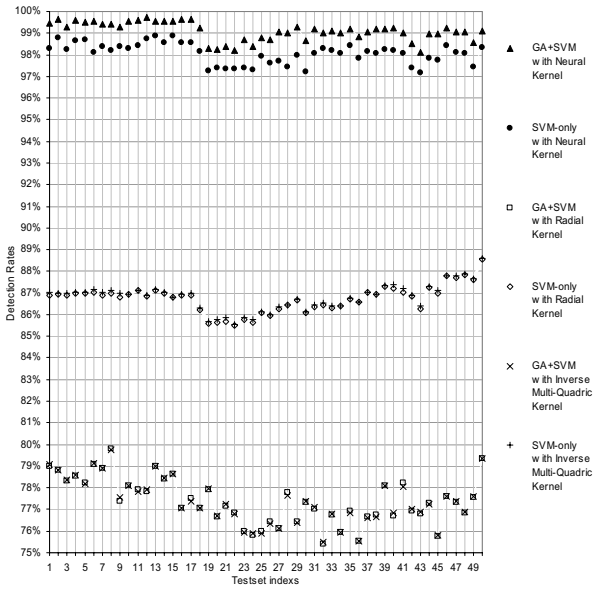


**Figure 4. Detection Rates in Classification Phase.**

## 4. Conclusions

This paper proposed GA to improve SVM based detection model in network intrusion detection system. IDS should cope with misuses as well as novel attacks, and moreover, IDS should give minimum overheads to computer system and IDS itself for processing audit data. SVM have been shown better performance than traditional classification methods [4, 9, 12]. SVM based IDS have been proposed [4, 5, 9, 12, 15] by several researchers. Although SVM based IDS can improve performance of IDS in term of detection rates and learning speed compared to conventional algorithm such as neural network, there are still rooms for improvement. When the numbers of features of audit data become very large, the overall performance of IDS often degraded severely in terms of processing time and detection rates. To cope with these problems, we adopted GA technique which provides fast and excellent optimization to enable IDS to figure out the *optimal detection model* for SVM. We demonstrated the feasibility of proposed system by carrying out several experiments on KDD Cup 1999 intrusion detection dataset. In conclusion, our method is not only able to figure out optimal detection model but also minimize the number of features that SVM classifier should process and consequently maximize the detection rates of IDS.

## 5. References


[1] Xue-wen Chen, "Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines", In *2nd IEEE Computer Society Bioinformatics Conference* (CSB 2003), 11-14 August 2003, Stanford, CA, USA, pp. 504-505.

[2] Richard O. Duda,Peter E. Hart and David G. Stork, *Pattern Classification (2nd Edition)*, Wiley Interscience Inc., 2001.

[3] Frohlich, H, *et al*, "Feature selection for support vector machines by means of genetic algorithm", In *15th IEEE International Conf. on Tools with Artificial Intelligence*, 2003, pp. 142-148.

[4] M. Fugate and J.R. Gattiker, "Anomaly Detection Enhanced Classification in Computer Intrusion Detection", In *Pattern Recognition with Support Vector Machines, First International Workshop*, Niagara Falls, Canada, August 10, 2002, Lecture Notes in Computer Science 2388, pp. 186-197.

[5] Wenjie Hu and Yihua Liao and V. Rao Vemuri, "Robust Support Vector Machines for Anomaly Detection in Computer Security", In *Proceedings of 2003 International Conference on Machine Learning and Applications*, Los Angeles, CA, June 23-24, 2003.

[6] Thorsten Joachims, "Making large-scale support vector machine learning practical", In *Advances in kernel methods: support vector learning*, MIT Press, Cambridge, MA, 1999.

[7] KDD-CUP-99 Task Description: http://kdd.ics.uci.edu/databases/kddcup99/task.html

[8] KDD Cup 1999 Data.: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[9] Dong Seong Kim, Jong Sou Park. "Network-based Intrusion Detection with Support Vector Machines", In *Information Networking, Networking Technologies for Enhanced Internet Services International Conference* (ICOIN 2003), Cheju Island, Korea, Feb. 12-14, 2003, Lecture Notes in Computer Science 2662, pp 747-756.

[10] Christopher Kruegel, Fredrik Valeur, Giovanni Vigna and Richard A. Kemmerer, "Stateful Intrusion Detection for High-Speed Networks", In *IEEE Symposium on Security and Privacy*, IEEE Computer Society Press, USA, May 2002.

[11] Melanie Mitchell, *An Introduction to Genetic Algorithms,* MIT Press, 1996.

[12] S. Mukkamala, G. Janoski, A H. Sung. "Intrusion Detection Using Neural Networks and Support Vector Machines", In *Proceedings of IEEE International Joint Conference on Neural Networks*, IEEE Computer Society Press, 2002, pp.1702-1707.

[13] S Mukkamala, A H. Sung. "Feature Selection for Intrusion Detection Using Neural Networks and Support Vector Machines", *Proceedings of the 82nd Annual Meeting of the Transportation Research Board*, National Academics.

[14] Syng-Yup Ohn, Ha-Nam Nguyen, Dong Seong Kim, Jong Sou Park, "Determining Optimal Decision Model for Support Vector Machine by Genetic Algorithm", In *International Symposium on Computational and Information Sciences*, Shanghai, China, December 16-18, 2004, Lecture Notes in Computer Science, pp. 895-902.

[15] Jong Sou Park, Julee Lee, Dong Seong Kim, Sung-Do Chi, "Using Support Vector Ma-chine to Detect the Host-based Intrusion", In *IRC International Conference on Internet Information Retrieval*, 2002, pp. 172-178.

[16] Results of the KDD'99 Classifier Learning Contest, http://www-cse.ucsd.edu/users/elkan/clresults.html

[17] V. Vapnik. *The Nature of Statistical Learning Theory*, Springer, Berlin Heidelberg, New York, 1995.


## Acknowledgements


This work was supported (in part) by the Ministry of Information & Communications, Korea, under the Information Technology Research Center (ITRC) Support Program.