International Journal of Fuzzy Logic and Intelligent Systems, vol. 8, no. 4, December 2008 pp. 316-321

The Design and Implementation of Anomaly Traffic Analysis System using Data Mining

Se-Yul Lee*, Sang-Yeop Cho and Yong-Soo Kim***†**

*** Department of Computer Science, Chungwoon University,**
San29 Namjang-Ri, Hongseong-Eup, Hongseong-Gun, Chungnam, 350-701, Korea
E-mail: pirate@chungwoon.ac.kr
**** Department of Internet, Chungwoon University,**
San29 Namjang-Ri, Hongseong-Eup, Hongseong-Gun, Chungnam, 350-701, Korea
E-mail : sycho@chungwoon.ac.kr
***** Department of Computer Engineering, Daejeon University,**
96-3 Yongun-Dong, Dong-Gu, Daejeon, 300-716, Korea
E-mail : kystj@dju.kr

Abstract

Advanced computer network technology enables computers to be connected in an open network environment. Despite the growing numbers of security threats to networks, most intrusion detection identifies security attacks mainly by detecting misuse using a set of rules based on past hacking patterns. This pattern matching has a high rate of false positives and can not detect new hacking patterns, which makes it vulnerable to previously unidentified attack patterns and variations in attack and increases false negatives. Intrusion detection and analysis technologies are thus required. This paper investigates the asymmetric costs of false errors to enhance the performances the detection systems. The proposed method utilizes the network model to consider the cost ratio of false errors. By comparing false positive errors with false negative errors, this scheme achieved better performance on the view point of both security and system performance objectives. The results of our empirical experiment show that the network model provides high accuracy in detection. In addition, the simulation results show that effectiveness of anomaly traffic detection is enhanced by considering the costs of false errors.

Key Words : Detection Systems, False Errors, Anomaly Traffic, Patterns Analysis, Data Mining

1. Introduction

Nowadays, networked computer systems play an increasingly important role in our society. They have become the targets of a wide array of malicious attacks that invariably turn into actual intrusions. This is the reason why computer security has become an essential concern for network administrators. Too often, intrusions wreak havoc inside Local Area Networks and the time and cost to repair the damage can grow to extreme proportions. Instead of using passive measures to fix and patch security holes once they have been exploited, it is more effective to adopt a proactive approach to intrusions. In addition to the well-established intrusion prevention techniques such as data encryption and message integrity, user authentication and user authorization, as well as the avoidance of security flaws inherent to many off-the-shelf applications, intrusion detection techniques can be viewed as an additional safeguard for networked computers. One of key research areas is analysis system which many companies have adopted to protect their information assets for several years. In order to address the security problems, many automated detection and analysis systems have been developed.

However, between 2006 and 2008, more than 120 new attack techniques, which exploited web & SQL Servers of Microsoft Ltd. that is one of the most widely used web servers, were created and published. Recently, several detection and analysis systems have been proposed based on various technologies. A "false positive error" is an error that detection system sensor misinterprets one or more normal packets or activities as attack. The detection system operators spend very much time to distinguish events. On the other hand, a "false negative error" is an error resulting from that attacker is misclassified as a normal user. It is quite difficult to distinguish intruders from normal users. It is also hard to predict all possible false negative errors and false positive errors due to the enormous varieties and complexities of today's networks. The detection system operators rely on their experience to identify and resolve unexpected false error issues. This paper proposes a method to analyze and reduce the total costs based on the asymmetric costs of errors in the detection system. This method adopts the model which has shown successful results for detecting and identifying

Manuscript received Nov. 2, 2008; revised Dec. 12, 2008.
† Corresponding author

unauthorized or abnormal activities from the networks [1].

The proposed method is to minimize the loss for an organization under an open network environment. This study employs the network model for detection. Furthermore, this study analyzes the cost effectiveness of the false error levels and presents experimental results for the validation of our detection model.

The section 2 presents the introduction of detection systems and the studies of data mining approaches for detection systems. This Anomaly Traffic Anlaysis System(ATAS) model is addressed in detail in Section 3. In Section 4, the asymmetric costs of false negative errors and false positive errors are validated by experimental results. Finally, this paper is concluded with the summary, contributions, and limitations.

2. Intrusion Detection System(IDS)

An intrusion is an unauthorized access or usage of the resources of a computer system [2]. Intrusion Detection System (IDS) is the software with the functions of detecting, identifying, and responding to unauthorized or abnormal activities on a target system [3, 4]. The goal of IDS is to provide a mechanism for the detection of security violations either in real-time or batch-mode [5, 6]. These violations are initiated either by outsiders attempting to break into a system or by insiders attempting to misuse their privileges [7]. IDS collects information from a variety of systems and network sources, and then analyzes the information for signs of intrusion and misuse [8]. The major functions performed by IDS are monitoring and analyzing user and system activity, assessing the integrity of critical system and data files, recognizing activity patterns reflecting known attacks, responding automatically to detected activity, and reporting the outcome of the detection process.

Intrusion detection is broadly divided into two categories based on the detection method: misuse detection and anomaly detection. Misuse detection works by searching for the traces or patterns of well-known port attacks. Clearly, only known attacks, which leave characteristic traces, can be detected by this method. This model of the normal user or system behavior is commonly known as the user of system profile. A major strength of anomaly detection is its ability to detect previously unknown attacks.

The IDS is categorized according to the kind of audit source location that they analyze. The IDS is classified as either a network-based intrusion detection or a host-based intrusion detection approach for recognizing the deflecting attacks. When IDS looks for these patterns in the network traffic, it is classified as network based intrusion detection. When IDS looks for attack signatures in the log files, it is classified as host based intrusion detection. In either case, these products look for attack signatures and specific patterns that usually indicate malicious or suspicious intent. Host based IDS analyzes host bound audit sources such as operating system audit trails, system logs, and application logs. Network based IDS analyzes network packets that are captured on a network.

The current IDS has contributed to identifying attacks using historical patterns. But they have difficulty in identifying attacks using a new pattern or with no pattern [9]. Previous studies have utilized a rule based approach such as USTAT, NADIR, and W&S [10-12]. They lack flexibility in the rule to audit record representation. Slight variations in an attack sequence can affect the activity rule comparison to a degree that intrusion is not detected by the intrusion detection mechanism. While increasing the level of abstraction of the rule base does provide a partial solution, it also reduces the granularity of the intrusion detection device. These limitations in rule based systems can be summarized as follows: the lack of flexibility and maintainability in the acquisition process of rules, the lack of predictive capability, the lack of automatic learning capability, a high rate of false alarms or missing alarms, and difficulty in applying organizational security policies.

Many recent approaches of IDS have utilized data mining techniques, for example, the Computer Misuse Detection System (CMDS), the Intrusion Detection Expert System (IDES), and the Multics Intrusion Detection and Alerting System (MIDAS) using neural networks. These approaches build detection models by applying data mining techniques to large data sets of an audit trail collected by a system [13]. Data mining based IDS collects data from sensors which monitor several aspects of a system. Sensors may monitor network activity, system calls used by user processes, and file system accesses. They extract predictive features from the raw data stream being monitored to produce formatted data that can be used for detection. Data gathered by sensors are evaluated by a detector using a detection technique.

3. Anomaly Traffic Analysis System(ATAS)

3.1 Architecture of ATAS

The ATAS model consists of network based detection model and monitoring tool(Fig. 1) [14]. The model adopts the problem solving methodology that uses previous problem solving situations to solve new problems. The model does preprocessing by packet analysis module and packet capture module. The packet capture module captures and controls packet. The packet capture module does real-time capturing and packet filtering by using the monitoring tool of Detector4win Ver. 1.2 [15]. In the packet filtering process, packets are stored according to the features that distinguish normal packets from abnormal packets. The packet analysis module stores data and analyzes half-open

state. After storing packets, the packets, which are extracted by audit record rules in the packet analysis module, are sent to the detection module [17]. Of figure

The input and the output of detection module, namely STEP 1, are traffic and alert, respectively. The traffic is an audit packet and the alert is generated when an intrusion is detected. The detection module consists of session classifier, pattern extractor, and pattern comparator. The session classifier takes packet of the traffic and checks whether or not the source is the same as the destination. There is a buffer for the specific session to be stored. And, if the next packet is arrived, it is stored in the corresponding buffer. If all packets of the corresponding buffer are collected, all packets of the corresponding buffer become output on session. The output session becomes an input to the pattern extractor or pattern comparator according to action mode. The action mode consists of learning mode and pre-detection mode. The output session from the session classifier is sent to the pattern extractor in the learning mode and to the pattern comparator in the pre-detection mode. Fig. 2 is the block diagram of the STEP 1.

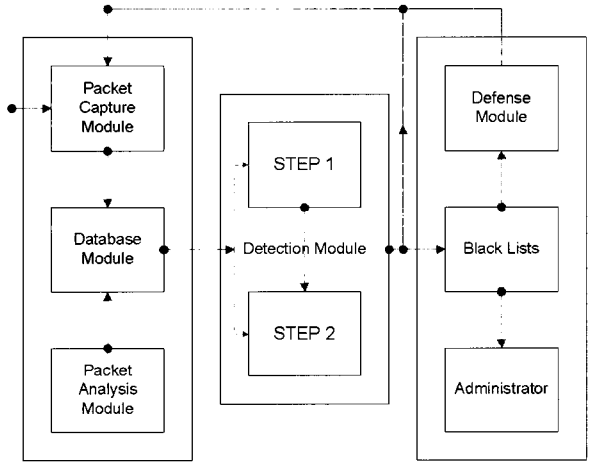

Fig. 1. Architecture of ATAS [14]

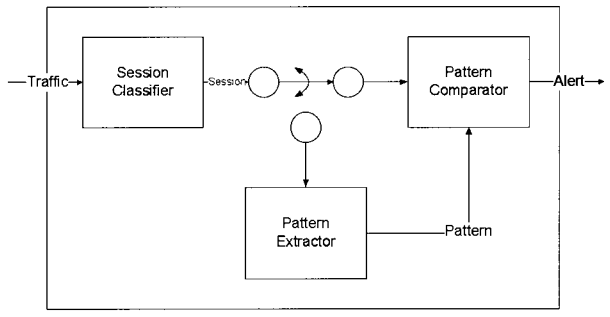

Fig. 2. A Block Diagram of STEP 1[17]

The pattern extractor collects the sessions, which have the same destination, and extracts common pattern. Each extractor has two features. The first feature is a head part that appears in common sessions, which have the same destination, when sessions are arranged by size packets using the time sequence. The second feature is the minimum length of the sessions that have the same destination. The length of session is the number of packets of session.

The pattern comparator compares packets with patterns of the rule based system. If the probe packets and patterns of the rule do not correspond, the pattern comparator considers the probe packets as the abnormal session and generates an alert signal. Thus, the pattern comparator receives a session and patterns of the rule as an inputs. From the input session the data size and the length of session are extracted. If there is a mismatch in one of two features, the pattern comparator considers a session as the abnormal session. What we must consider for the pattern extraction is whether we extract the pattern continuously or we extract the pattern periodically. We generally call the former the real-time pattern extraction and the latter the off-line pattern extraction. The real-time pattern extraction is better than off-line pattern extraction in the viewpoint of updating the recently changed pattern. But, it is difficult to update the pattern when probes occur. For the pattern, if possible, normal traffic becomes patterns of the rule. Otherwise, an abnormal traffic sometimes becomes pattern of the rule. And the abnormal intrusion traffic is considered as the normal traffic. It is called false negative error. The model uses detection module, namely STEP 2, to compensate the false negative error by using fuzzy cognitive maps. The detection module of model is intelligent and uses causal knowledge reasoning utilizing variable events that hold mutual dependences. When CPU usage capacity increases because syn packet increases, the weight of a node W_{ik}, has the value of range from 0 to 1 in fuzzy cognitive maps. The weight is the effect value of path analysis calculated using quantitative Micro Software's Eview Ver. 3.1[15]. The total weighted value of a node depends on path between nodes and iteration number. It is expressed as the following equation.

$$N_k(t_{n+1}) = \sum_{i=1}^{n} W_{ik}(t_n) N_i(t_n) \qquad (1)$$

$N_k(t_n)$: the value of the node k at the iteration number t_n
t_n : iteration number
$W_{ik}(t_n)$: weight between the node i and the node k at the iteration number t_n

On the above equation, the sign of weight between the node i and the node k depends on the effect from the source node to the destination node.

3.2 Analysis for costs of errors

The analysis of costs of errors is presented in the Fig. 3. The purpose of Fig. 3 & equation 2 are to analyze the relationship between the total costs and detection system errors and to find the optimal threshold of network model that minimizes `the total

costs for intrusion detection. The solution provides the weights of errors while the weights can be adjusted to enhance the effectiveness of intrusion detection by controlling the threshold value of the activation function. The activation function produces the level of excitation by comparing the sum of these weighted inputs with the threshold value. This value is entered into the activation function, for example the sigmoid function, to derive the output from the node.

The cost of attacks or errors has been considered in designing IDS [16]. The cost of a false negative error is much higher than that of a false positive error because an organization may suffer from various security incidents compromising confidentiality, integrity, and availability, when IDS does not detect real attacks. This paper introduces the concept of the asymmetric costs of errors to calculate overall misclassification costs. The performance of detection system is optimized when the total costs are minimized.

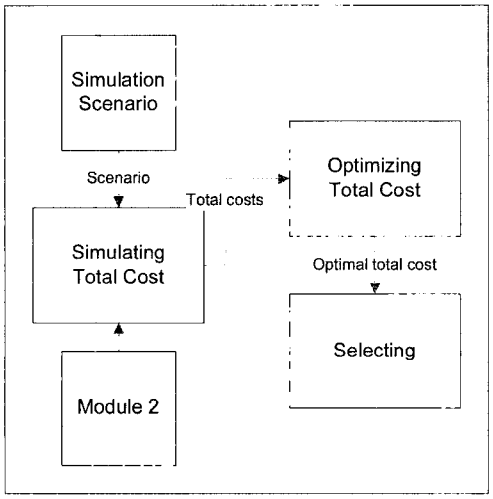

Fig. 3. A Block Diagram of Error's Cost

A false negative error, which is the cost of not detecting an attack, is occurred when the detection system does not function properly and mistakenly ignores an attack. This means that the attack will succeed and the target resource will be damaged. Thus, a false negative error should take a higher weight than a false positive error. The false negative errors are therefore described as the damage cost of the attack. The cost function for detection system can be defined as follows:

$$A_{total}(x) = \omega_1 A_1 + \omega_2 A_2 + ... + \omega_n A_n$$

$$= \sum_{i=1}^{n} \omega_i A_i \qquad (2)$$

$A_{total}(x)$: Total cost
ω_i : Weight for each cost A_i
A_i : Cost for each error i

To measure each cost, we used the errors that are the misclassified by our detection methods. The cost ratio of a false positive error and a false negative error varies depending on the characteristics of the organization. Thus, we found out the minimal total costs by the simulation of adjusting the weights one hundred times. The threshold values can be searched to minimize the total costs for a specific cost ratio of false negative errors to false positive errors.

4. Simulation and Performance Evaluation

For the performance evaluation of the proposed model, we used the 'KDD'99(Knowledge Discovery Contest Data) data set' by MIT Lincoln Lab, which consists of 5,000,000 labeled data (training data, which consist of syn and normal data) and 2,000,000 non-labeled data (test data). We utilize a network model to apply the proposed method for the above data. Three-layer feed-forward network is used to detect an intrusion. Logistic activation function is utilized in the output layer. The number of hidden nodes is selected through experiment with n/2, n, and 2n of nodes (n is the sum of input nodes) by fixing the input and output nodes. A series of experiments is conducted to analyze the effects of changing the value of the threshold values of false negative errors and false positive errors (Fig. 4).

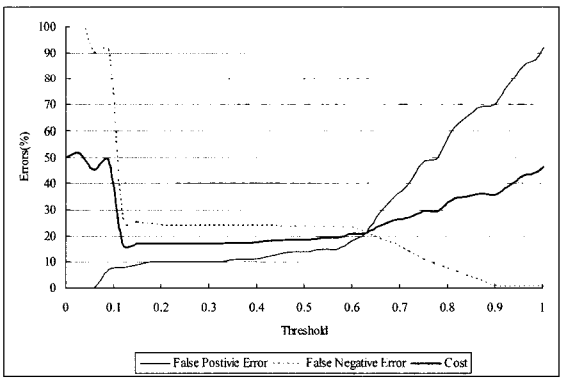

Fig. 4. The performance of ATAS with cost of errors

As the threshold value increases, false positive errors increase while false negative errors decrease. After the ratio of false negative errors over false positive errors is given, the threshold value that minimizes the total cost can be determined. Let us suppose that the cost of false negative error is equal to that of false positive error. We can find that the optimal point of the threshold is 0.12 from Fig. 4. When the output is larger than the threshold value, the output is interpreted as an attack, and normal vice versa.

The performance of networks is calculated by the function of cost, which consists of false positive errors and false negative errors (Table 1). The performance of network model is

measured using the output sample data. The total cost of the network model is 15.32% when the threshold value is 0.5, which is a general value without considering costs of errors. When the optimal point of threshold of 0.12 is applied to the network model from Fig. 4, the cost is 15.95%. The cost decreases and the performance of the intrusion detection model is sensitive to the threshold value. A false negative error is more important in detection system as mentioned in the previous section. We need to concentrate on the decrease of false negative errors by changing the threshold value. The false negative errors have fallen from 9.01 to 7.84% - a decrease of 1.17%. The change in the total cost would be greater as weights are added to the negative false errors.

Table 1. Results of Simulation performance on Test-Bed

Threshold value	Sample	False positive errors (%)	False negative errors (%)	Cost (%)
0.5	Input	24.63	9.23	16.93
	Output	21.63	9.01	15.32
	Total	23.13	9.12	16.13
0.12	Input	25.49	8.45	16.97
	Output	24.06	7.84	15.95
	Total	24.78	8.15	16.46

Fig. 5 shows the movement of the optimal that minimized the total costs depending on the cost ratio. We increase the cost ratio from 1 to 10 by 0.1 and search each minimal total cost by 100 times through the simulation. Simultaneously, the threshold values were optimized when the simulation was performed. As the cost ratio of false negative errors to false positive errors increases, total cost decreases while the amount of the decrease becomes smaller. The performance of intrusion detection varies depending on each cost ratio score. Table 2 depicts the detailed score of the simulation results from Fig. 5.

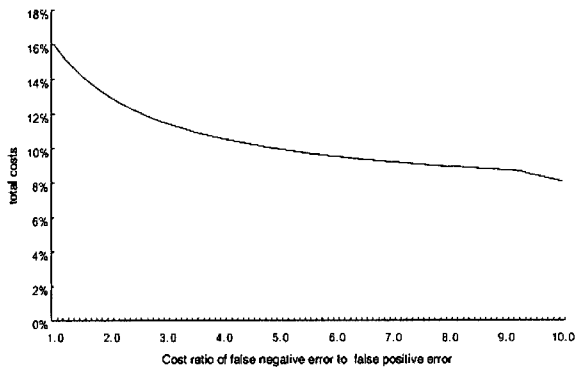

Fig. 5. Results of minimum total costs

When a false negative error takes the weight value five times higher than a false positive error, the total percentage of errors is 9.015%. When the cost ratio is 1, the total percentage of errors is 15.965%. The decreased amount is about 38% compared to the original cost, which has the cost ratio of 1. Thus we come to the conclusion that a success factor for Detection System is the cost ratio and threshold, as well as the classification accuracy. Organizations should provide appropriate policies regarding the cost ratio, and then they can effectively discover the threshold value that minimizes the total cost and determines a cost effective intrusion detection policy based on asymmetric costs of errors in Detection System.

Table 2. Results of ATAS performance for total costs

Weights of false negative errors over false positive errors	False positive errors (%)	False negative errors (%)	Total cost savings (%)
1	24.09	7.84	15.965
2	24.08	6.94	15.510
3	24.95	6.96	15.955
4	24.98	6.94	15.960
5	24.98	6.95	15.965
6	24.98	6.96	15.970
7	24.98	6.94	15.960
8	24.98	6.94	15.960
9	24.95	6.96	15.955
10	24.95	6.94	15.945

5. Conclusion

There have been a variety of researches and systems designed to detect intrusion by using data mining approaches. However, most researches addressed the measure of system performance as providing prediction accuracy without considering the costs of errors in intrusion detection. In this paper, we proposed an ATAS network model based on costs of false positive errors and false negative errors. The first diagram of this research develops a network model for intrusion detection. The second diagram analyzes the system performance based on costs of errors.

The results of the empirical experiment indicate that the network model provides very high performance for the accuracy of intrusion detection. The cost of false negative errors must be much higher than that of the false positive errors to an organization. The total cost of errors is minimized by adjusting the threshold value for the specific cost ratio of false negative errors to false positive errors.

This paper provides insights into the multi-faceted method of

evaluating IDS performance in terms of false negative and false positive errors as well as detection accuracy. Our study will help organizations to employ IDS effectively by reducing the total costs due to the inevitable IDS errors.

Reference

[1] Lee, W., Stolfo, S. J., "A data mining framework for building intrusion detection models," *IEEE Symposium on Security and Privacy*, pp. 209-220, 1999.

[2] Safavi-Naini, R., Balachadran, B., "Case-based reasoning for intrusion detection," *12th Annual Computer Security Application Conference*, pp. 214-223, 1996.

[3] Denning, D. E., "An intrusion detection model," *IEEE Trans. S. E., SE-13(2)*, pp. 222-232, 1987.

[4] Richards, K., "Network based intrusion detection: a review of technologies," *Computer and Security*, pp. 671-682, 1999.

[5] Debar, H., Dacier, M., "Towards a taxonomy of intrusion detection systems," *Computer Networks*, pp. 805-822, 1989.

[6] Debar, H., Becker, M., "A neural network component for an intrusion detection system," *IEEE Computer Society Symposium Research in Security and Privacy*, pp. 240-250, 1992.

[7] Weber, R., "Information Systems Control and Audit," *IEEE Symposium on Security and Privacy*, pp. 120-128, 1999.

[8] Lippmann, R. P., "Improving intrusion detection performance using keyword selection and neural networks," *Computer Networks,* Vol. 24, pp. 597-603, 2000.

[9] Jasper, R. J., Huang, M. Y., "A large scale distributed intrusion detection framework based on attack strategy analysis," *Computer Networks*, Vol. 31, pp. 2465-2475, 1999.

[10] Ilgun, K., Kemmerer, R. A., "Ustat: a real time intrusion system for UNIX," *Proceedings of the IEEE Symposium on Research in Security and Privacy*, pp. 16-28, 1993.

[11] Hubbards, B., Haley, T., McAuliffe, L., Schaefer, L., Kelem, N., Walcott, D., Feiertag, R., Schaefer, M., "Computer system intrusion detection," *IEEE Computer Society Symposium Research in Security and Privacy*, pp. 120-128, 1990.

[12] Vaccaro, H. S., "Detection of anomalous computer session activity," *Proceedings of the IEEE Symposium on Research in Security and Privacy*, pp. 280-289, 1989.

[13] Helman, P., "Statistical foundations of audit trail analysis for the detection of computer misuse," *IEEE Transactions on software engineering*, Vol. 19, pp. 861-901, 1993.

[14] Se-Yul Lee and Yong-Soo Kim, "Design and analysis of probe detection systems for TCP networks," *International Journal of Advanced Computational Intelligence & Intelligent Informatics*, Vol. 8, pp. 369-372, 2004.

[15] Se-Yul Lee, An Adaptive probe detection model using fuzzy cognitive maps, *Ph. D. Dissertation*, Daejeon University, 2003.

[16] Maxion, R. A., "Masquerade detection truncated command lines," *International Conference on Dependable Systems and Networks*, pp. 219-228, 2002.

[17] Se-Yul Lee, Byoung-Chan Chun, Yong-Soo Kim, "The network model for Detection Systems based on data mining and the false errors," *International Journal of Fuzzy Logic and Intelligent Systems*, Vol. 6, No. 2, pp. 64-68, 2006.

Se-Yul Lee

He received the B. S. degree in the Department of Physics, the M. S. degree in the Department of Information & Communications Engineering, and Ph. D. degree in the Department of Computer Engineering from Daejeon University, in 2003, respectively. He was a Researcher at Insopack Co in 2001. He is currently a professor in the Department of Computer Science, Chungwoon University. His research interests include network security, system security, patterns recognition, and grid middleware.

Sang-Yeop Cho

He received the B. S. degree from Hannam University in 1986. He received M. S. degree and Ph. D. degree in the Department of Computer Engineering from Chungang University in 1988 and 1993, respectively. He is currently an Associate Professor in the Department of Internet, Chungwoon University. His research interests include artificial intelligence, fuzzy logic, and application of Petri net.

Yong-Soo Kim

He received the B. S. degree from Yonsei University, and M. S. degree from Korea Advanced Institute of Science and Technology(KAIST) in 1981 and 1983, respectively. He received the Ph. D. degree in the Department of Electrical Engineering from Texas Tech University in 1993. He was a Researcher at Samsung Electronics Co. from 1983 to 1986. He is currently a Professor in the Department of Computer Engineering, Daejeon University. His research interests include neural networks, fuzzy logic, image processing, pattern recognition, and intrusion detection systems.