특집논문-08-13-1-11

네트워크 데이터 모델링을 위한 효과적인 성분 선택

김 호 인[a], 조 재 익[a], 이 인 용[a], 문 종 섭[a]‡

Effective Feature Selection Model for Network Data Modeling

Hoin Kim[a], Jaeik Cho[a], Inyong Lee[a], and Jongsub Moon[a]‡

요 약

네트워크 데이터 모델링은 침입 탐지 시스템의 성능 평가, 네트워크 모니터링, 네트워크 데이터 분석 기법 연구에 있어서 반드시 필요한 연구이다. 네트워크 데이터의 모델링에는 반드시 네트워크의 실제 데이터를 분석하고, 분석된 데이터를 이용하여 효과적으로 데이터를 구성 하여야만, 실제 네트워크 데이터의 충분한 정보를 모델링 된 데이터에 반영할 수 있다. 본 연구에서는 대규모의 네트워크 데이터에서 실제 네트워크에서 사용가능한 모든 성분에 대해 수량화 하였으며, 수량화 된 데이터를 통계적 분석 방법을 통하여 모델링 데이터에서 가장 효과적인 분류 기준으로 작용할 수 있는 성분을 분석하였다.

Abstract

Network data modeling is a essential research for the evaluation for intrusion detection systems performance, network modeling and methods for analyzing network data. In network data modeling, real data from the network must be analyzed and the modeled data must be efficiently composed to reflect a sufficient amount of the original data. In this paper the useful elements of real network data were quantified from packets captured from a huge network. Futhermore, a statistical analysis method was used to find the most effective element for efficiently classifying the modeled data.

Keyword : Network Data Set, Network Feature Selection, Intrusion Detection, Network Data Analysis

Ⅰ. 서 론

오늘날 네트워크의 발달과 더불어 여러 가지 침입 탐지 기술 및 침입 분석 기술이 요구되고 있으며, 또한 현재의 침입 탐지와 관련된 여러 가지 방어 시스템의 검증을 필요로 하고 있다. 침입과 관련된 여러 가지 방어 시스템 및 방어 알고리즘과 관련된 연구에 있어서 연구 및 시스템을 검증할 수 있는 네트워크 데이터 셋이 반드시 필요하다. 이 네트워크 데이터 셋은 논리적으로 충분히 이해할 수 있는 네트워크 데이터의 정보를 충실히 반영하여야 하며, 반영되는 정보이 량이 데이터의 량에 비해 커야 한다. 즉, 소규모의 데이터 셋을 이용하여도 충분히 네트워크 방어 장비의 검증을 할 수 있는 데이터 셋이여야 한다.

이러한 데이터 셋 구성은 네트워크의 실제 데이터를 모델링하는 기술로부터 출발한다. 또한 소규모의 데이터를 이용하여 최대한 많은 정보 (논리적 설명이 가능한 오차 범위 이내의 동일성)를 반영하는 것, 즉, 네트워크 데이터 모델링에 있어서 효과적인 성분 선정은 반드시 선행 되어야 될 연구이다. 네트워크 데이터, 즉 패킷 데이터는 여러 가지 정보를 갖고 있으며 패킷 헤더의 주된 내용으로는 시간 정

a) 고려대학교 정보경영공학전문대학원
Center for Information Security Technologies(CIST), Korea University
‡ 교신저자 : 문종섭(jsmoon@korea.ac.kr)
※ 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원 사업의 연구결과로 수행되었음 (IITA-2007-(C1090-0701-0025))

보, 발신지 주소, 수신지 주소, 여러 가지 패킷의 상세 정보 등으로 구성이 된다[1][2]. 패킷의 헤더는 헤더만으로 많은 정보가 확인이 되며, 헤더 자체의 정보를 가지고 실제 네트워크의 통신이 가능한 것이다[3]. 이러한 패킷의 여러 가지 정보 중 과연 어떤 패킷의 정보가 패킷 데이터의 정보에 많은 비중을 갖고 있는지 확인 해 보고 확인 된 결과의 성분을 모델링 연구의 기반 연구화 할 수 있다. 또한 침입 탐지 시스템이 오용 탐지 혹은 시그너쳐 기반 탐지를 떠나 어떤 최소 성분을 이용하여 탐지 할 수 있는지에 대한 기준을 확인할 수 있을 것이다.

본 논문은 패킷의 여러가지 정보 중 많은 중요도를 가지는 정보의 선정을 위한 효과적인 방법을 제안한다. 제안하는 방법은 주성분 분석을 통해 실제 네트워크 패킷 헤더에서 사용 가능한 모든 속성을 분석해 본다. 주성분 분석을 이용한 실험에서는 MIT/LL에서 공개한 네트워크 데이터 및 고려대학교 일부 네트워크에서 수집한 패킷 데이터를 이용하여 분석 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 MIT/LL 데이터셋, 이를 모델링한 KDD Cup 99'와 이를 분석하기 위해 사용한 주성분 분석에 대하여 설명한다. 3장에서는 실험을 위하여 사용한 실험 데이터의 전처리 과정과 실험결과에 대하여 설명하고 이를 분석한다. 마지막으로 4장에서는 실험 결과를 바탕으로 결론을 맺고 향후 연구 방향에 대해 설명한다.

II. 네트워크 데이터

네트워크 데이터는 네트워크에서 이동되는 모든 종류의 데이터 흐름이라고 할 수 있다. 이를 모델링 하여 데이터셋을 구성하는 것은 네트워크의 침입 및 효율성을 검토 하기 위하여 반드시 확보해야 되는 기술이고, 이 기술을 확보하기 위해서는 효과적인 네트워크 데이터의 성분을 확인하여야 한다. 본 장에서는 현재 일반적으로 많이 사용되고 있고 가장 대용량 네트워크 데이터를 분석한 MIT 의 링컨 연구실 데이터와 해당 데이터를 모델링한 데이터인 KDD CUP 99' 데이터에 대해서 설명한다.

1. MIT 링컨 연구실 데이터

본 논문에서 비교 분석의 대상으로 하고 있는 데이터는 미국 MIT의 링컨 연구실에서 구성한 데이터 셋이다. 해당 데이터 셋은 현재까지 많은 단점들에 대해 연구 결과가 발표 되고 있으나 직접적으로 MIT 의 데이터와 MIT 가 해당 데이터를 이용해 모델링 된 데이터인 KDD 데이터의 비교 분석 연구는 진행되어 있지 않다. MIT 에서는 표본 데이터를 다음 그림과 같은 환경에서 수집 하였다.

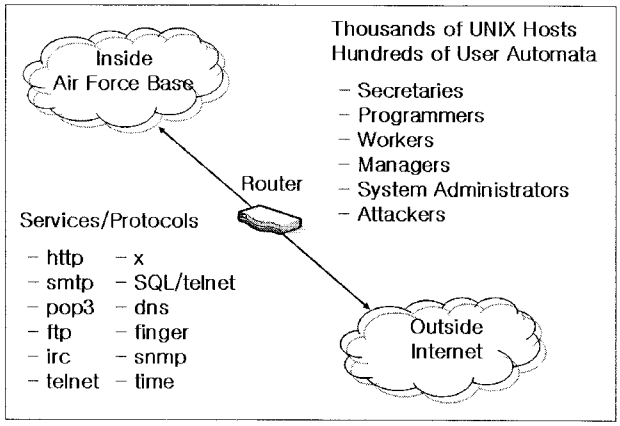


그림 1. MIT 의 모집단 데이터 수집 환경
Fig. 1. The environment for collecting the population data of MIT

MIT에서 수집한 수집 환경은 미국에 소속되어 있는 일부 공군 망에서 패킷 데이터를 수집 하였다. 네트워크 데이터 수집은 공군 망 내부에서 외부로 연결되는 하나만 존재하는 라우터에서 외부와 내부, 모든 데이터가 수집 되었으며, 수집된 데이터는 개인정보 보호의 이유에서 패킷 데이터그램을 삭제한 후 사용되었다. 최대한 많은 수의 단말 노드들이 있는 환경을 구성하기 위하여 가상 환경을 추가 하였으며, 이때 가상 환경에서 일반적인 행위, 공격적인 행위에 대해 실시하도록 하였다[3][4][6].

패킷에서 개인 정보 등의 이유로 인하여 실제 데이터그램을 삭제하고 적절한 공격 데이터를 추가 삽입하기 위해서 (그림 2) 와 같은 폐쇄 망을 구성하여 패킷을 생성 하였다. 이때 생성된 패킷 데이터에서 2초 단위를 한 시퀀스 단위로 하여 1개의 데이터 셋을 구성하였다. 데이터 셋에 포

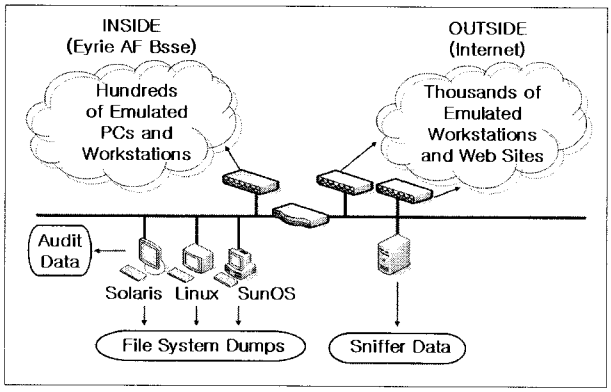


그림 2. MIT의 패킷 재생성 환경
Fig. 2. The environment for regenerating packets of MIT

표 1. KDD CUP 99' 데이터 셋의 TCP 데이터 구조
Table 1. The TCP data structure of the KDD CUP 99' data set

Feature	Description	Type
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol e.g. tcp, udp	discrete
service	network service on the destination e.g., http, telnet	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_btyes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete
land	1 if connection is from/to the same host/port; 0 otherwise	discrete
wrong_fragment	number of "wrong" fragment	continuous
urgent	number of urgent packet	continuous

함되어 있는 정보는 <표 1>과 같다.

각 패킷에 대하여 데이터 셋을 구성한 것이 아니라 2초 단위의 시간 간격으로 패킷을 시퀀스 단위로 구분 하였다. 2초 단위에는 여러 가지의 개수의 패킷이 포함될 수 있으며, 2초 단위에서 가장 많이 발생한 항목의 값을 데이터 셋의 값으로 정의하여 모델링 하였다.

이때 각 데이터의 항목별 분포는 2초 단위의 구분에 따라 최대 분포 값을 사용하였기 때문에 소규모의 발현 데이터는 무시 되었으며, 이로 인하여 비교 분석의 결과에서 모델링 데이터와 표본 데이터의 차이가 발생한다.

2. KDD CUP 99' 네트워크 데이터 셋

KDD CUP 99' 네트워크 데이터 셋은 1998년 MIT 의 링컨 연구실에서 구성된 데이터를 기반으로 제작된 데이터 셋이다. KDD CUP 99' 의 데이터 셋 제작 목적은 다음과 같다.

(1) 침입 탐지 시스템의 개발, 검증 및 침입 탐지 알고리즘의 개발, 검증을 위한 연구 자료 제공
(2) 침입 탐지 시스템의 효율성에 대한 분석 자료
(3) 침입 탐지에 대한 효율성 분석 자료
(4) DARPA 에 대한 연구 지원

KDD CUP 99' 는 지식 자료에 대한 국제 학술 기구로서 1999년에 침입 탐지에 대한 대회를 KDD CUP 99' 데이터 셋을 이용하여 개최하였다[5].

KDD CUP 99' 데이터 셋의 공개된 구성 요소는 다음과 같다.

(1) 외부에서 내부로 오는 네트워크 데이터
(2) 내부에서 외부로 전송되는 네트워크 데이터
(3) Basic Security Module (BSM) 로그 데이터
(4) 윈도우 NT 시스템의 감사 데이터
(5) 전체 호스트의 디렉토리 리스트
(6) 공격 대상 디렉토리의 상세 설명
(7) 내부 파일 시스템의 상세 구조 설명

또한 공개된 데이터의 구성 형식은 다음과 같다.

(1) 데이터의 속성이 표시된 모든 데이터
(2) 데이터의 속성이 표시된 10% 샘플 데이터
(3) 속성이 표시되지 않은 10% 샘플 데이터
(4) 속성이 표시되지 않은 모든 데이터
(5) 공격의 종류
(6) 표시 방법의 설명

본 논문의 관련 실험에서는 약 1기가바이트의 데이터 속성이 표시된 10% 샘플 데이터를 이용하였다. 데이터 셋의 세부 자료 구조는 다음과 같다.

3. 주성분 분석 이론

주성분 분석(Principal Component Analysis) 는 대응 되는 성분을 이용하여 데이터의 주 성분 및 성분의 성향을 분석할 수 있는 방법이다. 본 논문에서는 네트워크 데이터를 수량화 하였으며 수량화된 데이터를 이용하여 주 성분 분석을 시도 하였다[8].

주 성분 분석에 있어서 대표적인 네트워크 성향, 즉 네트워크 성분 수량화 데이터 중 가장 큰 분산을 갖는 데이터를 확인하고 가장 큰 분산값이 도출된 성분을 기준으로 데이터를 사영한다[7][9]. 이때 사영은 Karhunen-Loeve 사영을 이용하여 사영된 데이터 성분을 기준으로 데이터 성분의 성향이 도출된다. 또한 데이터의 분산값을 이용하여 사영 후 분산 정도를 확인할 수 있다[10][11].

데이터 행렬 x 에 대하여 주성분 y 값을 도출한다면

$$y = E[(y^T x)^2] \tag{1}$$

위의 식은 주성분 y 가 x 중 yi에 대해서 가장 큰 값을 가지는 것을 의미한다. 이것을 이용하여 n 번째 주성분일 경우 n-1 과 n 번째의 수량 차를 이용하여 주성분 값을 확인할 수 있다.

$$\widehat{x_{n-1}} = x - \sum_{i=1}^{n-1} y_i y_i^T x \tag{2}$$

또한 수량 차를 이용하여 n-1 번째 성분의 주 성분을 확인할 경우 데이터 행렬 x 에서 n 번째 주성분 탐색을 위해서는 반드시 n-1 성분을 제외 하여야 한다.

$$y_n = E[(y^T \widehat{x_{n-1}})^2] \tag{3}$$

또한 주성분 분석 결과를 Singular Value Decomposition (SVD) 이용하여 2차원 평면에 차원 축소가 가능하며 이때 성분 성향을 2차원 공간에 사영시켜 확인할 수 있다. 2차원 공간행렬 x ′ 는 다음과 같다.

$$x' = UWD^t \tag{4}$$

U 는 n열, n 행으로 구성된 orthogonal 행렬이며, D 는 n열, n 행으로 구성된 Diagonal 행렬이다. W 는 n열, p 행으로 구성된 행렬이며, 이것이 Singular Value 이다.

Ⅲ. 관련 실험

본 장에서는 본 논문에서 실험에 사용하기 위하여 고려대학교 일부 네트워크의 데이터를 수집에 대해 설명하고 또한 수집된 데이터를 이용한 실험에 대해서도 설명한다. 수집된 네트워크 데이터는 크게 두 종류의 데이터로서 소규모의 네트워크, 대규모의 네트워크 규모로 구분할 수 있다. 본 논문에서는 소규모 네트워크 데이터를 대변하기 위한 데이터로서 12개의 호스트로 구성된 네트워크 데이터, 대규모 네트워크 데이터를 대변하기 위한 데이터로서 1998개의 호스트로 구성된 네트워크 데이터를 수집하였다.

1. 소규모 네트워크 데이터 수집

소규모 네트워크 데이터 수집은 12개의 실제 사용 호스트로 구성된 네트워크 환경이였으며, 고려대학교의 실제

표 2. 소규모 수집 네트워크의 호스트 환경
Table 2. The host environment for capturing a small scale network

Host	Hardware	Software
1	Intel Pentium 4 2.0GHz 512MB Memory Realteck 8139	MS Windows XP SP2 Enterprise Edit.
2		
3		
4		
5		
6		
7		
8		
9		
10		
11	Intel Pentium 4 1.2GHz 1024MB Memory	RedHat Linux 9
12		

대학원의 일부 연구실 구성망이다. 수집은 총 14일동안 전체 패킷 수집을 하였으며, 외부 연결 네트워크에 간단한 스위치 장치를 이용하여 외부 연결 시스템에 원격 수집 하였다. 수집시 손실 데이터 량은 0.00001% 미만이었다. 수집 환경의 호스트는 다음과 같이 구성되었다.

총 14일의 데이터 수집 중 2일의 데이터는 일부 사용자의 편협적 대용량 서비스 사용(p2p)으로 네트워크 데이터 중 프로토콜이 평균 편차를 크게 벗어났기 때문에 삭제 하고 12일의 데이터만을 본 연구에서 사용하였다.

2. 대규모 네트워크 데이터 수집

대규모 네트워크 데이터 수집은 고려대학교 일부 네트워크에서 실험하였으며, 총 1998개의 호스트를 대상으로 수집하였다. 수집된 데이터는 네트워크 패킷 헤더부분이 1.7 테라바이트였으며, 이를 이용하여 본 연구에 사용하였다. 대용량 데이터의 전송 수집을 손실을 최소화하기 위해 다중 버퍼 시스템을 이용하여 메인 저장 시스템에 실시간 압축 전송하였다. 개략적인 대규모 네트워크 데이터의 수집 환경은 다음과 같다.

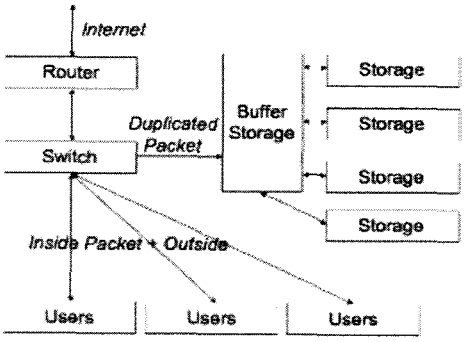


그림 3. 대규모 네트워크 데이터 수집 환경
Fig. 3. The environment for capturing a huge scale network

3. 네트워크 데이터의 수량화 및 주 성분 분석

2장에 설명한 주 성분 분석 이론을 네트워크 데이터 분석에 이용하기 위하여 네트워크 데이터를 다음과 같은 행렬로 구성하였다.

표 3. 네트워크 데이터 행렬
Table 3. The network data matrix

	Featurename	1st Packet	2nd Packet	k'th Packet
p_1	2nd Layer Protocol	p_1, n_1	p_1, n_2	p_1, n_k
p_2	3rd Layer Protocol	p_2, n_1	p_2, n_2	p_2, n_k
p_3	Type of Service	p_3, n_1	p_3, n_2	p_3, n_k
p_4	ICMP Type	p_4, n_1	p_4, n_2	p_4, n_k
p_5	ICMP Type sub code	p_5, n_1	p_5, n_2	p_5, n_k
p_6	IP Header Length	p_6, n_1	p_6, n_2	p_6, n_k
p_7	Header Total Length	p_7, n_1	p_7, n_2	p_7, n_k
p_8	UDP Header Length	p_8, n_1	p_8, n_2	p_8, n_k
p_9	IP Version	p_9, n_1	p_9, n_2	p_9, n_k
p_{10}	IP ID	p_{10}, n_1	p_{10}, n_2	p_{10}, n_k
p_{11}	Fragment Offset	p_{11}, n_1	p_{11}, n_2	p_{11}, n_k
p_{12}	IP Time to Live	p_{12}, n_1	p_{12}, n_2	p_{12}, n_k
p_{13}	IP Checksum	p_{13}, n_1	p_{13}, n_2	p_{13}, n_k
p_{14}	Source Port	p_{14}, n_1	p_{14}, n_2	p_{14}, n_k
p_{15}	Destination Port	p_{15}, n_1	p_{15}, n_2	p_{15}, n_k
p_{16}	TCP Sequence	p_{16}, n_1	p_{16}, n_2	p_{16}, n_k
p_{17}	TCP Acknowledgment	p_{17}, n_1	p_{17}, n_2	p_{17}, n_k
p_{18}	TCP Data Offset	p_{18}, n_1	p_{18}, n_2	p_{18}, n_k
p_{19}	3rd Layer Checksum	p_{19}, n_1	p_{19}, n_2	p_{19}, n_k
p_{20}	UDP ID	p_{20}, n_1	p_{20}, n_2	p_{20}, n_k
p_{21}	ARP Hardware Address	p_{21}, n_1	p_{21}, n_2	p_{21}, n_k
p_{22}	ARP Protocol Address	p_{22}, n_1	p_{22}, n_2	p_{22}, n_k
p_{23}	ARP Command	p_{23}, n_1	p_{23}, n_2	p_{23}, n_k

또한 위의 행렬의 데이터를 수량화 데이터로 구성하기 위해서 일부 데이터를 다음과 같이 치환하였다. 각 성분의 세부 내용은 다음과 같다.

표 4. 데이터 행렬 치환
Table 4. Permutation of the data matrix

피쳐 종류	세부 내용	데이터 종류	데이터 형식
프로토콜	데이터의 형식에따른 구분	TCP, UDP, ..	문자열
패킷 길이	데이터의 총 길이 표시	0, 1024	숫자
TTL	라우터 이동 횟 수	0, 64, 128	숫자
ID	패킷 데이터의 송신 수신 조합을 위한 라벨링	1, 2, 3, 4, ···	숫자

또한 데이터는 10000개 기준으로 각 속성 종류별 발생 빈도를 이용하여 표준화 하였다. 표준화는 Scalar Multiplication 방법으로서 아래와 같다.

$$Ra' = r(a1, a2) = (ra1, ra2) \quad (5)$$

수량화 된 행렬 데이터를 이용하여 주성분 분석을 하였으며 주성분 분석은 R project 의 표준 PCA 라이브러리를 이용하였다. 주성분 분석 결과 전체 데이터의 95% 정보량을 4개의 성분으로 표현가능하며, Eigen Value 는 다음과 같다.

표 5. Eigen Value
Table 5. Eigen Value

Feature	Value
3rd Layer Protocol	1.173
Header Total Length	1.004
IP Time to Live	0.966
IP ID	0.827
Total	95%

또한 데이터 성분의 차원 사영 결과를 SVD 를 이용하여 2차원 사영 결과는 다음과 같다.

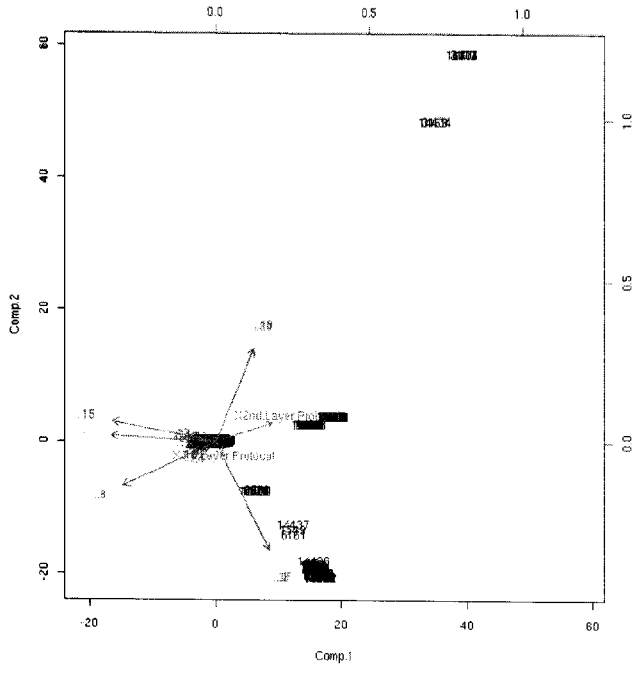


그림 4. 2차원 사영 결과
Fig. 4. Results of a 2 dimensional projection

Ⅳ. 결론 및 향후 연구

주성분 분석을 이용한 네트워크 성분 분석을 통하여 많은 네트워크 데이터 정보 중 4개의 정보만 이용해서 구성하여도 95% 이상의 정보를 포함하고 있다. 또한 현재 대표적으로 사용하고 있는 데이터 셋은 이러한 사전 분석 정보가 없기 때문에 비효율적인 부분이 포함되어 있다. 추후 본 연구를 바탕으로 가장 효과적으로 최소 데이터를 이용하여 최대 정보를 표현할 수 있는 효과적인 네트워크 데이터 셋 구성에 기반 연구로 사용할 수 있을 것이다.

참 고 문 헌

[1] Lippmann R.P., Fried D.J., Graf I., Haines J.W, Kendall K.R., McClung D., Weber D., Webster S.E., Wyschogrod D., Cunningham R.K., and Zissman M.A., Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation, in Proceedings of the 2000 DARPA Information Survivability Conference and Exposition(DISCEX), Vol. 2, 12-26, 2000, IEEE Computer Society Press: Los Alamitos, CA.

[2] Lippmann, R.P., Haines J.W., Fried D.J., Korba J., and Das J., The 1999 DARPA Offline Intrusion Detection Evaluation. Computer Networks, 2000. 34(2), 579-595.

[3] J. W. Haines. 1999 DARPA Intrusion Detection Evaluation. Technical Report 1062. MIT Lincoln Laboratory. 2001

[4] Saharon Rosset, Aron Inger. KDD-cup 99. ACM SIGKDD Explorations Newsletter. KDD-99 Conference report. 2000

[5] Norm O'Rourke, Larry Hatcher, Edward J. Stepandski, Using SAS for Univariate & Multivariate Statistics Second Edition, Wiley InterScience, 2005.

[6] Lippmann R.P. and Haines J., Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation, in Recent Advances in Intrusion Detection, Third International Workshop,, RAID 2000 Toulouse, France, October 204, 2000 Proceedings, H. Debar, L.ME, and S.F. Wu, Editors. 2000, Springer Verlag, 162-182.

[7] Sergios Theodoridis, Konstantinos Koutroumbas, Pattern Recognition Third Edition, Academic Press, 2006.

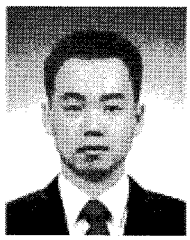
[8] James Lattin, J. Douglas Carroll, Paul E. Green, Analysing Multivariate Data, Thomson Books, 2002.

[9] 한학용, 패턴인식 개론, 한빛 미디어, 2005.

[10] 허명회, 다변량수량화, 자유아카데미, 1999.

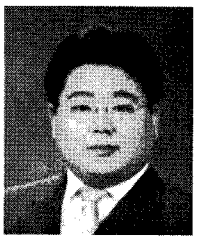
[11] Norm O'Rourke, Larry Hatcher, Edward J. Stepandski, Using SAS for Univariate & Multivariate Statistics Second Edition, Wiley InterScience, 2005.

저 자 소 개



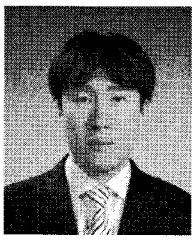
김 호 인

- 2007년 : 동국대학교 멀티미디어공학 학사
- 2007년 ~ 현재 : 고려대학교 정보경영공학대학원 석사과정
- 주관심분야 : 네트워크 모델링, 네트워크 보안, 시스템 보안



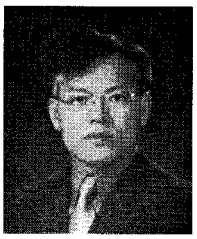
조 재 익

- 2005년 : 동국대학교 컴퓨터학 학사
- 2008년 : 고려대학교 정보보호대학원 석사
- 2008 ~ 현재 : 고려대학교 정보경영공학대학원 박사과정
- 주관심분야 : 패턴인식, 침입 탐지 시스템, 악성 코드의 확산 패턴 분석, 네트워크 모델링, 네트워크 공격 방어 시뮬레이션



이 인 용

- 2007년 : 남서울대학 컴퓨터학 학사
- 2007년 ~ 현재 : 고려대학교 정보경영공학대학원 석사과정
- 주관심분야 : 네트워크 보안, 패턴인식, 시스템 보안



문 종 섭

- 1991년 : Illinois Institute of Technology 전산학 박사
- 1993년 ~ 현재 : 고려대학교 전자 및 정보공학부 교수
- 2001년 ~ 현재 : 고려대학교 정보경영공학대학원 겸임교수
- 주관심분야 : 신경망 이론, 패턴인식, 시스템 보안, 네트워크 보안