情 報 保 護 學 會 誌
第19卷 第2號, 2009. 4

통계적 척도 선택 방법에 따른 네트워크 침입 분류의 성능 비교

문 길 종*, 김 용 민**, 노 봉 남***

요 약

네트워크 기술의 발달에 따른 서비스의 증가는 네트워크 트래픽과 함께 취약점도 증대하여 이를 악용하는 행위도 늘어나고 있다. 따라서 네트워크 침입탐지 시스템은 증가하는 트래픽의 양을 처리할 수 있어야 하며, 악의적인 행동을 효과적으로 탐지 할 수 있어야 한다. 증가하는 트래픽을 효과적으로 처리하고 탐지의 정확성을 높이기 위해 처리 데이터를 감소시키는 기술이 요구된다. 이러한 방법들은 크게 데이터 필터링, 척도 선택, 데이터 클러스터링의 영역으로 구분되며, 본 논문에서는 척도 선택의 방법으로 데이터 처리의 감소 및 효과적 침입탐지를 수행할 수 있음을 보이고자 한다. 실험 데이터는 KDDCUP 99 데이터 셋을 이용하였으며, 통계적 척도선택의 방법으로 분류율, 오탐율, 거리값, 규칙, 선택된 척도 등을 제시함으로써 침입 탐지 시 데이터 처리량이 감소하였고, 분류율은 증가, 오탐율은 감소하여 침입 탐지 정확성이 높아짐을 알 수 있었다. 또한 본 논문에서 제시한 방법이 다른 관련연구에서 제시한 선택 척도보다 높은 정확성을 보임으로써 보다 유용함을 증명할 수 있었다.

Ⅰ. 서 론

네트워크 기술이 발달함으로써 빠른 서비스, 정보 획득, 거리 감소 등의 혜택이 증가하고 있다. 하지만, 이와 같은 네트워크 발달의 증가로 네트워크와 시스템의 취약점을 이용하여 시스템을 파괴하는 악의적인 행위들도 증가하고 있다. 현재 서비스 거부 공격, 웜, 봇, 스캐닝 공격 등과 같은 악의적인 행위로 인해 발생하는 네트워크와 시스템의 장애에 대한 문제는 정보보호 분야에서 중요한 문제로 대두되고 있다. 따라서 이러한 행위들을 탐지하고 방어할 수 있는 침입탐지 시스템이 요구되고 있지만, 침입탐지 시스템은 계속해서 증가하는 새로운 서비스들로 인한 네트워크 트래픽을 효과적으로 처리하지 못하고 있다. 그러므로 침입탐지 시스템에서는 불필요한 트래픽을 효과적으로 줄일 수 있는 기술이 필요하다.

데이터를 감소시킬 수 있는 방법은 크게 데이터 필터링(data filtering), 척도 선택(feature selection), 데이터 클러스터링(data clustering)으로 나눌 수 있다. 본 논문에서는 이들 방법 중 척도 선택을 중심으로 연구 및 비교, 분석한다. 척도 선택의 방법은 데이터의 양을 줄일 수 있을 뿐 아니라 침입탐지에 유용한 척도를 선택하여 이용함으로써 침입탐지의 정확성을 높일 수 있다.

본 논문에서는 신뢰성 있는 실험 및 검증을 위해 DARPA 98 데이터 셋을 바탕으로 네트워크 연결 데이터를 추출하여 구성한 KDD (Knowledge Discovery in Database) CUP 99 데이터 셋을 실험 데이터로 사용한다. 침입탐지를 위한 유용한 척도 선택을 위해서 유클리디언 거리(Euclidean distance), 상대 복잡도(relative entropy), 카이제곱 테스트(chi-square test), J-S 거리(Jensen Shannon divergence) 등의 방법을 제시하고, 이를 이용하여 척도 선택을 위한 실험을 한다. 그리고 탐지 성능을 평가하기 위해서 결정트리 알고리즘 중 하나인 C4.5를 이용하여 침입탐지 규칙을 생성하고 실험 결과를 비교한다.

본 논문의 2장에서는 네트워크 공격 분류, 침입 탐지 시스템의 분류, 데이터 감소 기술, 거리 측정 알고리즘,

* (주)정보보호기술 (alcorjjong@gmail.com)
** 전남대학교 문화콘텐츠학부 (ymkim@chonnam.ac.kr)
*** 전남대학교 전자컴퓨터정보학부 (bbong@chonnam.ac.kr)

침입 탐지를 위한 척도 선택에 관련된 연구에 대해서 살펴본다. 3장에서는 생성된 확률 분포의 특징에 대해서 살펴보고, 거리 측정 알고리즘을 적용하는 예제를 보인다. 4장에서는 거리 측정 알고리즘을 적용해서 얻은 실험 결과를 비교 및 분석하고 선택된 척도에 의해 생성된 규칙과 탐지 실험결과를 보인다. 마지막으로 5장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 관련 연구

2.1 네트워크 공격 분류

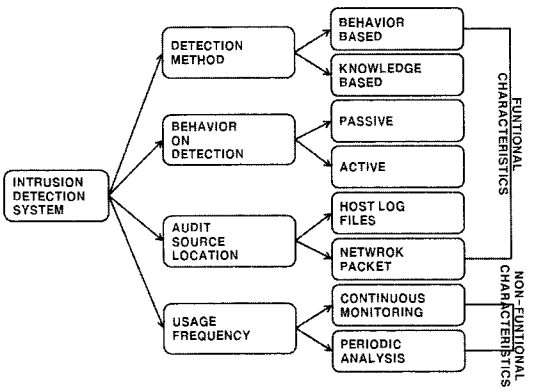
DARPA 데이터 셋은 공격을 효과적으로 분류한 대표적인 데이터 셋으로써, 공격을 서비스 거부 공격(Denial of Service), 프로브(Probes), R2L(Remote to Local), U2R(User to Root)로 나누고 있다[1]. 하지만 이 중에서 U2R의 경우는 주로 네트워크가 아닌 시스템의 로그에서 특징을 찾을 수 있어 네트워크 공격에서는 제외된다[2,3]. 현재, 네트워크 공격은 웜과 봇, 봇넷, DDoS(Distributed Denial of Service) 등의 공격으로 나뉘며, 이것은 기존의 서비스 거부 공격 및 프로브 공격들의 일부 특징을 가지고 있는 공격들이다.

서비스 거부 공격은 비정상적으로 컴퓨터 자원을 소모시키기 위한 시도로써, 일반적으로 악의적인 의도로 인터넷의 웹서버를 무력화 시키는데 사용한다. 서비스 거부 공격은 크게 논리적(logic) 혹은 소프트웨어 공격과 범람(flood)으로 분류할 수 있다[4].

2.2 침입탐지시스템

침입 탐지 시스템은 악의적인 사용자로부터 컴퓨터 시스템으로의 불법적인 접근 및 오용행위를 감지하여 관리자에게 통보하거나 각종 침입행위를 기록하기 위한 시스템으로 시작하였다[5]. 일반적으로 침입 탐지 시스템은 호스트와 네트워크 기반으로 분류할 수 있으며, 네트워크의 발달에 따라 호스트를 통한 침입보다 네트워크를 통한 침입이 급증하고 있어, 네트워크 기반 침입탐지 시스템이 호스트 기반 침입탐지 시스템보다 더 중요시 되고 있다[6].

또한 침입탐지 시스템은 [그림 1]과 같이 분류될 수 있다[7]. 시스템의 정상적인 행동을 기반으로 탐지에 사용했을 때, 행동 기반(behavior-based), 공격에 관한 정보를 사용하여 탐지에 사용할 때, 지식 기반(knowledge-based)이라 한다. 탐지 행동(behavior on detection) 분류는 탐지 시스템이 공격에 대한 대응에 따라 능동적(active)과 수동적(passive)으로 나뉜다.



〔그림 1〕 침입 탐지 시스템의 분류

감사 근원지 위치(audit source location)에 의한 분류는 입력 정보의 종류를 바탕으로 정해진다. 또한 사용빈도(usage frequency)에 따른 침입 탐지 시스템의 분류는 침입 탐지 시스템이 실시간 모니터링 혹은 주기적인 모니터링을 하느냐에 따라서 분류되기도 한다.

2.3 침입탐지 연구

일반적인 침입 탐지 모델은 규칙 기반 패턴 매칭 시스템으로 Denning에 의해서 제안되었고[8], 현재 감사 레코드와 일치하는 프로파일 사이의 유사도를 이용하여 탐지하는 것이다. NIDES (Next-generation Intrusion Detection Expert System)는 통계를 바탕으로 한 가장 대표적인 침입탐지 시스템으로 장기 프로파일과 단기 프로파일을 비교하여 유사성을 측정한 것이다[9]. 하이퍼뷰(Hyperview)는 신경망을 사용하는 대표적인 침입탐지 시스템으로써, 신경망[10]과 전문가 시스템(expert system)의 두 가지 모듈로 구성되어 있다. 또한 Lippmann은 키워드 기반 침입탐지의 방법에 신경망을 적용하였다. HMM(Hidden Markov Model)은 관찰된 심볼들(symbols)의 서열을 모델링하기에 유용한 알고리즘으로, 다른 방법에 비해 시스템 콜 이벤트(system call event)를 모델링하는데 좋은 성능을 보인다. 하지만, 정상적인 행동을 모델링하기 위해서는 많은 시간이 요구된다[11]. 또한 Lazarevic은 HMM을 비정상행위 탐지에

적용하는 연구를 수행했다[12]. MARS (Multivariate Adaptive Regression Splines)는 연속적인 이진 변수에 대해서 정확한 예측 모델을 자동으로 생성하는 방법으로, 최적 변수 변환과 상호작용, 고차원 데이터에 숨겨져 있는 복잡한 데이터 구조를 찾는 것에 탁월하다[13].

Abraham은 MARS를 기초로 한 침입 탐지 시스템을 제안하였고, LGP(Linear Genetic Programming)를 기초로 한 침입 탐지 시스템은 Mukkamala에 의해 제안되었다[14]. 단일 척도 탐지의 약점을 극복하기 위해, 다수의 척도를 사용하는 침입 탐지 방법이 제안되기도 했다. Chebrolu는 베이지안 네트워크와 규칙기반 방법인 CART(Classification and Regression Trees)의 장점을 혼합한 침입 탐지 시스템을 제안하였다[15].

2.4 데이터 감소 기술

침입 탐지 시스템에서 요구되는 감사 데이터의 양은 매우 작은 네트워크 단에서도 데이터 전체를 검사하기에는 불가능하다. 또한 척도들 사이에는 인간이 발견하기에는 복잡한 관계가 수 없이 존재한다. 침입 탐지 시스템은 효과적으로 침입을 탐지하기 위해서 처리해야할 데이터의 양을 감소시킬 필요가 있다. 만약 실시간으로 침입 탐지가 요구될 때, 데이터 처리 문제는 가장 민감한 문제가 될 것이다. 데이터 감소(data reduction)를 위한 방법은 데이터 필터링, 척도 선택, 데이터 군집화 방법이 존재한다.

데이터 필터링(data filtering)의 목적은 침입 탐지 시스템이 처리해야 하는 데이터의 양을 줄이는 것이다. 침입 탐지에 있어서 불필요한 데이터가 다수 존재하므로, 이 데이터가 침입 탐지 시스템에서 처리되기 전에 제거되어야 시스템의 효율성을 높일 수 있다. 이 방법의 특징은 데이터 저장 공간을 줄일 수 있고, 데이터 처리 시간을 줄일 수 있으며, 침입 탐지율을 향상시킬 수 있도록 할 수 있다. 하지만 유용한 데이터를 잘못 삭제할 수도 있으므로, 유의해서 사용해야 한다.

척도 선택(feature selection)은 중복되거나 가치가 없는 척도들을 원시 데이터에서 제거함으로써, 학습 모델의 성능을 개선시킬 수 있다. 또한 중요한 척도가 무엇인지, 척도들 간에 서로 어떤 관련이 있는지 설명함으로써, 데이터를 이해하기 쉽게 한다. 또한, 데이터 차원성(curse of dimensionality)의 완화, 일반화 정도(capability of generalization) 강화, 학습 처리 능력 향상, 모델 해석력 개선 등도 척도 선택의 장점이다.

침입 탐지 분야에서의 척도는 침입 탐지를 위해 사용되는 정보, 탐지된 침입 시도, 정상적인 행위 등에서 추출할 수 있다. 척도 선택은 오용 행위를 가장 잘 나타내는 척도들을 찾기 위해 사용하거나, 오용 행위간의 구분을 하기 위해 사용할 수 있다.

데이터 군집화(data clustering)는 객체들을 다른 그룹들로 분류하여 묶는 것으로, 보다 정확하게 데이터 셋을 부분집합으로 나누는 것으로써, 각 부분집합에서의 데이터는 다수의 공통된 특성을 공유한다. 데이터 군집화는 정의된 거리 측정(distance measurement) 방법에 따라 특성이 달라지며, 통계적으로 데이터를 분석하기 위해 다양한 분야에서 사용하는 일반적인 기술이다. 군집화는 침입 탐지 데이터에서 감춰진 패턴을 찾고, 탐지에 유용한 척도를 찾는데 사용되며, 실제 데이터 대신에 클러스터의 특징들을 저장함으로써 데이터와 저장 공간을 감소시킬 수 있다.

2.5 척도 선택 연구

본 절에서는 침입 탐지를 위한 척도 선택에 관련된 연구에 대해서 살펴본다. 모든 연구는 KDDCUP 99 데이터 셋을 이용하며, 본 논문에서 41개의 척도는 이름 대신 1부터 41까지 번호로 나타낸다[24].

LIDM(Lightweight intrusion detection model)은 Yang에 의해 제안되었고[16], 비용 면에서 효율적이고 효과적이다. 척도 선택 시에 정보 획득량과 카이 제곱 테스트를 사용하여, 불필요한 척도를 제거함으로써 데이터를 단순화하여 빠르고 정확한 탐지를 할 수 있고, 탐지 방법으로 최대 엔트로피 모델(Maximum Entropy model)을 사용하였다.

Kuchimanchi는 척도 선택 방법으로 NNPCA(neural network principal component analysis)와 NLCA(nonlinear component analysis)라는 두 개의 신경망 모델을 연구했다[17]. PCA(principal component analysis)와 제안된 두 개의 신경망 모델을 비교하여, NLCA 방법이 실험 데이터 크기의 약 30% 만큼을 줄일 수 있었고, NNPCA 방법은 대략 50%를 줄일 수 있다는 결과를 보였다.

Zhang은 데이터 마이닝 방법 중 하나인 랜덤 포레스트(random forest)를 이용해 척도 선택 연구를 수행 하였다[18]. LDA, ICA(Independent Component Analysis), PCA와 같은 척도 감소 방법이 Venkatachalam에 의해 비교

〔표 1〕 각 방법으로 선택된 척도 번호

방법	척도
IG	3, 5, 6, 10, 13, 23, 24, 27, 28, 37, 40, 41
카이제곱	3, 5, 6, 10, 13, 23, 24, 27, 28, 37, 40, 41
표준편차	1, 2, 4, 5, 6, 10, 13, 16, 21, 22 23, 24, 27, 31, 32, 34, 36, 37
랜덤포레스트	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41
LDA	1, 2, 3, 5, 7, 8, 11, 12, 14, 17, 22, 23, 24, 25, 26, 30, 32
ICA	3, 5, 6, 12, 23, 24, 25, 26, 28, 31, 32, 33, 35
PCA	1, 4, 5, 6, 7, 8, 9, 11, 12, 38, 39, 40, 41
MI	1, 2, 3, 4, 5, 6, 10, 12, 16, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 37, 40
CBFS	3, 4, 6, 10, 12, 25, 29, 37
MBM	1, 2, 3, 5, 7, 8, 11, 22, 23, 24, 25, 26, 30, 32
CART	3, 5, 6, 12, 23, 24, 25, 32, 33, 34, 36
러프셋	3, 4, 5, 24, 33, 41

실험되었다[19]. 그 연구는 Gmix, RBF, Binary tree, LAMSTAR, SOM, ART와 같은 신경망 알고리즘을 사용하는 오용 침입 탐지 시스템에 초점이 맞춰져 있다. 이 실험에서 PCA는 LDA, ICA보다 탐지율, 오탐율, 비용면에서 좋은 성능을 보였다.

Chebrolu는 마코브 블랭킷 모델(markov blanket model)과 결정 트리 분석을 이용하여 척도 선택의 유용성을 실험하여 보였으며[20], 효율적이고 효과적인 침입 탐지 시스템을 제안하기 위해서 베이지안 네트워크, CART, 그리고 베이지안 네트워크와 CART가 결합한 침입 탐지 모델에 대해 연구하였다[21].

Chou는 KDD CUP 99와 UCI 데이터 셋을 사용하여 상호 정보량(mutual information)과 쌍별 상호 정보량(pairwise MI)과 같은 척도 선택 방법을 연구를 수행하였다[22]. 이 방법들에 의해 선택된 척도들은 C4.5와 네이브 베이즈(Naive bayes)에 의한 실험 결과로 증명하였고, 성능 평가를 위해 CBFS(Correlation Based Feature Selection)와 FCBF (Fast Correlation-Based Filter)와 같은 척도 선택 알고리즘을 비교 실험하였다. Zainal은 유용한 척도 선택과 데이터의 분류를 위해 러프 셋(rough set) 이론을 연구했다. 유전자 알고리즘에 의해 선택된 26개의 척도 중, 러프 셋 알고리즘을 이용하여 다시 여섯 개의 중요 척도를 선택했다. 또한 제안한 이론과 SVDF(Support Vector Decision Function Ranking), LGP(Linear Genetic Programming), MARS (Multivariate Regression Splines)의 결과를 비교하였다[23].

[표 1]은 본 절에서 설명한 탐지척도 선택 알고리즘에 의해 선택된 척도를 보인 것이다.

Ⅲ. 확률 분포를 이용한 척도 선택

본 논문의 실험에서 사용할 KDDCUP 99 데이터 셋에 대해 살펴보고, 척도 선택 연구를 위한 전체적인 과정을 살펴본다.

3.1 KDDCUP 99 데이터 분석

침입 탐지에 대해 연구, 조사, 평가를 위해 MIT 링컨 연구실에서는 DARPA 1998 침입 탐지 평가 데이터 셋을 수집하였고[24], KDDCUP 99 침입 탐지 대회에서 이 데이터를 가공한 데이터를 제공하였다.

KDD CUP 99 데이터 셋[25]은 DoS, R2L, U2R, Probes의 네 가지 공격 유형과 정상으로 분류한다. *kddcup. data_10_percent.data*는 *kddcup.data*에 포함된 데이터 중, 공격 데이터 일부가 다른 공격들에 비해 너무 많은 데이터를 가지고 있기 때문에, 학습에 영향을 줄 수 있어 데이터 개수를 조정한 학습 데이터이다. 그리고 실험 데이터로 사용한 것은 *corrected.data*로써, 학습에 포함되어 있던 22가지 공격 외에 named, xsnoop, snmpgetattack 등의 새로운 공격을 포함하고 있다. 본 연구에서는 학습에 사용되지 않은 공격에 대한 탐지에 대해서도 분류실험을 하기 위해 각 공격을 각 공격유형으로 분류하여 실험한다.

본 논문에서는 KDD CUP 99 데이터 중에서 네트워크와 관련이 적고, 다른 공격 유형에 비해 현저하게 적은 데이터 개수를 가진 R2L을 제외하였다. [표 2]는 학습과 탐지실험에서 사용되는 각 공격 데이터 개수를 보여준다.

KDD CUP 99 데이터 셋은 네트워크의 연결정보로 구성되어 있고, TCP 연결의 기본적인(basic)의 척도, TCP 연결 안에서의 콘텐츠 정보, 트래픽(traffic)에 관한 척도 세 가지 척도 유형으로 구분할 수 있다. 척도들

〔표 2〕 학습 및 실험 데이터의 개수

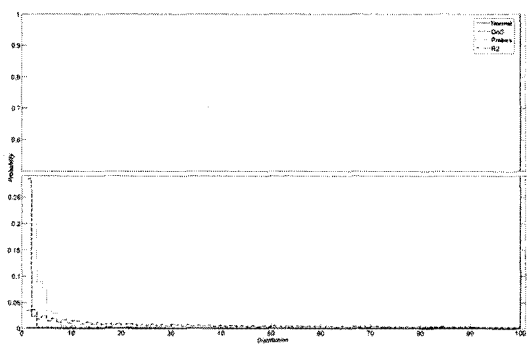
데이터 / 공격유형	훈련 데이터	실험 데이터
Normal	97,277	60,593
DoS	391,458	229,853
R2L	1,126	16,347
Probes	4,107	4,166

은 이산적인(discrete) 척도와 연속적인(continuous) 척도라는 두 가지 속성으로 나뉜다. 네트워크 연결 상의 *duration, ugent* 등의 정보는 연속적인 속성을 가지고 있으며, *protocol_type, service* 등의 척도는 이산적인 속성을 갖는다. [그림 2]는 각 척도들과 클래스의 데이터 구성을 보여준다.

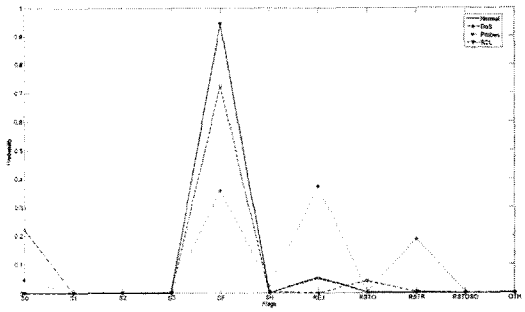
본 연구에서 각 척도간 거리를 측정하기 위해서 먼저 확률 분포(probability distribution)를 구해야한다. 확률 분포를 구하기 위해, 비모수적 방법(non-parametric method)에서 가장 일반으로 사용되는 히스토그램(histogram) 방법[26]을 이용하여, 연속형 값의 확률 분포(probability distribution)를 생성한다. 연속형 값을 갖지 않는 이산형 척도에 대해서는 척도가 갖는 값을 기준으로 확률 분포를 생성한다. 히스토그램 방법을 이용할 때, 전체 값의 최소값과 최대값을 동일한 크기의 구간으로 나누고, 그 구간에 해당하는 확률 분포를 구한다[27].

0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,9,9,1.00,0.00,0.11,0.00,0.00,0.00,0.00,0.00,0.00,Normal.
0,tcp,http,SF,239,486,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,19,19,1.00,0.00,0.05,0.00,0.00,0.00,0.00,0.00,0.00,Normal.
0,tcp,telnet,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,6,5,0.83,1.00,0.00,0.00,0.83,0.33,0.00,5,6,1.00,0.00,0.20,0.33,1.00,0.83,0.00,0.00,DoS.

〔그림 2〕 실험 데이터의 구성



〔그림 3〕 척도 *dst_host_count*의 확률 분포



〔그림 4〕 flag의 확률 분포

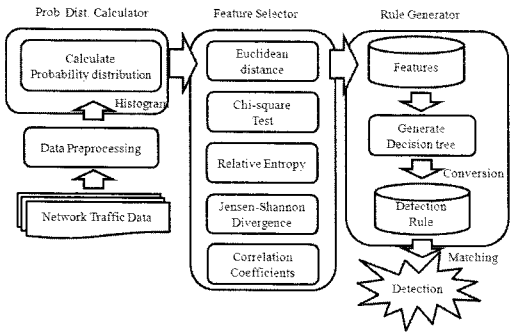
[그림 3]은 연속형 척도의 유형을 보여주는 대표적인 예로써, 32번째 척도인 *dst_host_count*의 확률 분포를 나타낸 것으로써 확률의 변화가 없는 0.3에서 0.5 사이의 확률 분포 구간은 생략하여 표현했다. 척도가 갖는 최소값 0과 최대값인 255를 100등분 하여 구간 값을 2.55로 나누어 확률 분포를 구한 것이다. 이 척도는 목적지 호스트에 접속하는 횟수를 나타낸 것으로 DoS 공격 유형의 99.6% 이상이 255번의 접속 횟수를 나타냄으로써 가장 큰 특징을 가지는 것을 알 수 있었다.

[그림 4]는 *flag*를 나타낸 것으로써 일반적인 이산형 척도의 확률 분포를 나타내며, *S0, S1, S2, S3, SF, SH, REJ, RSTO, RSTR, RSTOSO, OTH*의 값을 가진다. Probes의 경우 *S0, SF, REJ, RSTR*에서 고른 분포를 보임을 알 수 있고, R2L의 경우는 *SF*에서 가장 높은 분포를 보임을 알 수 있다.

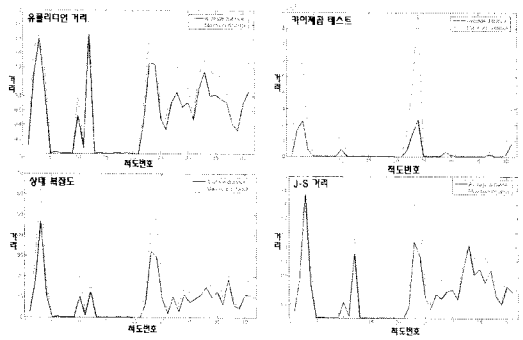
3.2 척도 선택

본 절에서는 유클리디언 거리, 카이 제곱 테스트, 상대 엔트로피, J-S 거리 방법을 이용하여 거리를 측정한다. 거리를 바탕으로 척도를 선택하고, 선택된 척도를 결정트리 알고리즘에 적용한 후, 탐지 규칙을 생성하고 탐지 실험하여 성능을 비교 및 분석한다.

척도 선택 실험은 [그림 5]와 같이 크게 세 부분으로 나뉜다. 첫 번째 부분은 히스토그램 방법을 사용하여 각 공격분류의 각 척도에 해당하는 확률 분포를 구하는 부분이고, 두 번째는 거리 알고리즘을 적용하여 각 척도마다 거리 값을 얻는 과정이다. 세 번째 부분은 선택된 척도를 바탕으로 결정트리 알고리즘 중 하나인 C4.5를 이용해서 탐지 규칙을 생성하고, 생성된 탐지 규칙을 이용하여 탐지하는 과정이다.



〔그림 5〕 척도 선택 실험을 위한 구성도



〔그림 6〕 각 방법의 공격 유형에 대한 거리값

[그림 6]은 네 가지의 방법으로 각 공격 유형의 척도에 대한 거리를 구한 것으로, 각 공격유형 별로 거리를 구한 다음, 네 가지 공격 유형의 평균값을 거리로 사용하게 된다. 각 방법은 값의 범위가 다르므로, 단계적 회귀 방법을 사용하여, 가장 작은 거리 값을 갖는 척도부터 하나씩 제거함으로써, 최종적으로 척도가 한 개 남을 때까지 규칙 생성과 분류를 반복해서 실험한다. 가장 좋은 분류율과 오탐율을 갖는 지점이 유용한 척도를 선택하는 최적의 거리임을 가정한다.

3.3 탐지 규칙 생성

본 실험에서는 침입 탐지의 정확성과 선택된 척도의 유용성을 증명하기 위해서 결정 트리 알고리즘을 사용한다. 결정트리 알고리즘은 데이터 마이닝에서 분류에 주로 사용되는 기법으로 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴을 속성의 조합으로 나타내고, 분류 모형을 트리 형태로 생성하는 것이다. 이렇게 만들어진 분류모형은 새로운 레코드를 분류하고, 해당 분류의 값을 예측하는데 사용된다. 지금까지 정확하고 빠르게 결정트리를 구성하기 위해 다양한 알고리즘이 연구되었고, 보다 개선된 알고리즘들이 계속 발표되었다. 결정 트리는 가지 분리(split) 방법과 가지치기(prune) 방법에 따라 CHAID, CART, QUEST, C4.5 등 여러 가지 알고리즘이 사용된다.

본 논문에서는 탐지 규칙 생성을 위해 C4.5를 사용한다. C4.5는 Ross Quinlan[28]이 개발하였으며, ID3 알고리즘을 확장한 것으로 생성된 결정 트리는 분류를 하기 위해 주로 사용된다. C4.5 결정트리 알고리즘의 가지 형성을 통해 분류 모델을 생성 해주는 알고리즘을 사용하였다[29,30].

[그림 7]은 C4.5 결정트리 알고리즘에 의해 생성된 결정트리를 보여준다. 각 침입 탐지 규칙은 생성된 결정트리의 루트로부터 각 가지 까지의 모든 중간 조건들의 조합이며, 이를 정형화하여 [그림 8]과 같이 탐지 규칙으로 변환한다.

```
Decision Tree:
count > 64 :
| dst_host_diff_srv_rate <= 0.15 :
| | srv_diff_host_rate <= 0 :
| | | protocol_type = icmp:[0] DoS (280685.0)
| | | protocol_type = tcp:
| | | | logged_in = 1:[0] Normal (35.0)
| | | | logged_in = 0:
| | | | | dst_host_diff_srv_rate <= 0.09 :[0] DoS (104689.0)
| | | | | dst_host_diff_srv_rate > 0.09 :
| | | | | | count <= 99 :[0] Probes (6.0/1.0)
| | | | | | count > 99 :[0] DoS (88.0)
| | | protocol_type = udp:
| | | | dst_host_same_srv_rate <= 0.59 :
| | | | dst_host_diff_srv_rate > 0.04 :[0] Probes (10.0)
.....
```

〔그림 7〕 C4.5 결정트리에 의해 생성된 탐지 규칙 트리

```
DoS;count > 64;dst_host_diff_srv_rate <= 0.15;srv_diff_host_rate
<= 0;protocol_type = icmp;
DoS;count > 64;dst_host_diff_srv_rate <= 0.15;srv_diff_host_rate
<= 0;protocol_type = tcp;logged_in = 0;dst_host_diff_srv_rate <= 0.09;
Probes;count > 64;dst_host_diff_srv_rate <= 0.15;srv_diff_host_rate
<= 0;protocol_type = tcp;logged_in = 0;dst_host_diff_srv_rate >
0.09;count <= 99;
DoS;count > 64;dst_host_diff_srv_rate <= 0.15;srv_diff_host_rate <=
0;protocol_type = tcp;logged_in = 0;dst_host_diff_srv_rate >
0.09;count > 99;
Probes;count > 64;dst_host_diff_srv_rate <= 0.15;srv_diff_host_rate
<= 0;protocol_type = udp;dst_host_same_srv_rate <=
0.59;dst_host_diff_srv_rate > 0.04
.....
```

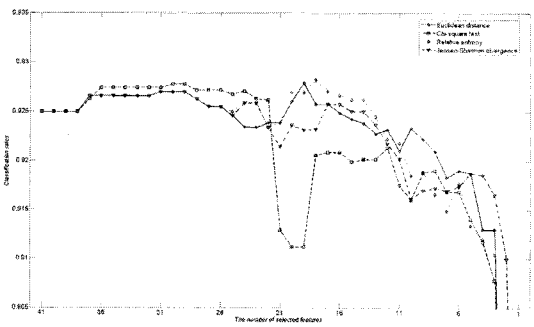
〔그림 8〕 생성된 트리를 변환한 탐지 규칙

Ⅳ. 분류 실험 결과

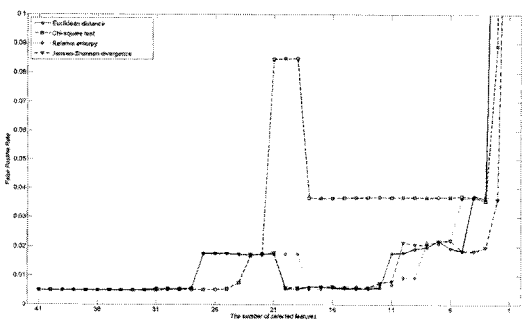
거리 측정 실험에서 선택된 척도를 이용하여 생성한 탐지 규칙을 바탕으로 *corrected.data*에 대해 분류 실험한다.

본 논문의 분류 실험은 C4.5에 의해 생성된 탐지 규칙과 데이터의 비교로 이루어지며, 가장 낮은 거리값을 갖는 척도를 하나씩 제거하면서, 선택된 척도가 한 개 남을 때까지 반복해서 실험한다. 그 중 가장 좋은 분류값을 갖는 선택된 척도의 구성을 최적의 척도로 선택한다. [그림 9]는 네 가지 방법으로 선택된 척도 개수에 따라 분류율이 변화함을 보인다. 전체적으로 좋은 결과를 보인 방법은 상대 엔트로피를 사용한 것으로써, 가장 높은 분류율을 보인 것은 선택된 척도가 18개일 때이다. [그림 10]은 선택된 척도의 개수에 따른 오탐율의 변화를 보인다. 네 가지 방법에서 가장 좋은 분류율을 보인 부분인 18개의 선택된 척도에서의 오탐율은 상대 엔트로피가 가장 좋음을 보였다.

[표 3]은 네 가지 방법을 이용하여 거리를 구하고, 가장 좋은 분류율 결과를 나타낸다. [그림 11]은 각 방법



〔그림 9〕 선택된 척도 개수에 따른 분류율



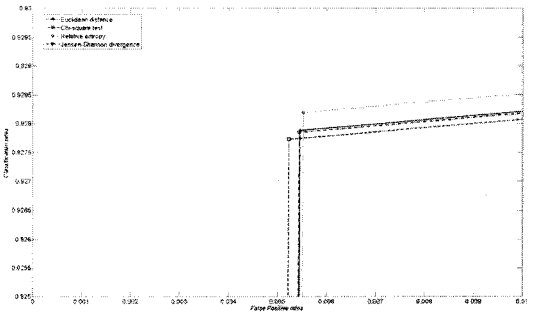
〔그림 10〕 선택된 척도 개수에 따른 오탐율

〔표 3〕 네 가지 방법의 분류율 결과

방법	분류 개수	분류율	오탐 개수	오탐율
유클리디언 거리	288532	0.92788	488	0.005463
카이 제곱 테스트	288484	0.92772	317	0.005253
상대엔트로피	288630	0.92819	335	0.005529
J-S 거리	288244	0.92785	330	0.005446

에서 가장 좋은 결과를 보인 지점의 분류율과 오탐율을 나타내는 ROC(Receiver operating characteristic)로써 상대 엔트로피가 가장 좋은 결과를 보임을 알 수 있다.

[표 4]는 네 가지 방법을 이용한 척도 선택에서 분류율이 가장 좋은 결과를 보일 때, 선택된 척도와 그 순위를 나타낸다. 상대 엔트로피가 네 가지 방법 중, 척도 선택 개수가 가장 적은 18개를 보였고, 유클리디언 거리, J-S 거리, 카이 제곱 테스트 순으로 척도 선택 개수가 적었다.



〔그림 11〕 가장 좋은 결과를 보인 ROC

〔표 4〕 각 방법에서 가장 좋은 결과의 선택된 척도

방법	선택 척도
유클리디언 거리	*12, 3, 23, 24, 33, 2, 32, 41, 28, 4, 35, 34, 36, 27, 30, 37, 40, 29, 10*
카이 제곱 테스트	*24, 3, 2, 23, 41, 10, 21, 4, 29, 40, 37, 30, 25, 27, 31, 1, 11, 39, 34, 38, 35, 32, 33, 6, 36, 19, 26, 17, 28, 14*
상대 엔트로피	*3, 23, 24, 37, 2, 33, 35, 12, 29, 32, 40, 41, 4, 10, 27, 34, 31, 25*
J-S 거리	*3, 23, 33, 12, 24, 35, 32, 37, 34, 2, 36, 4, 40, 30, 29, 41, 27, 38, 25, 28, 31, 10, 39, 22, 26, 1, 11, 6, 19*

〔표 5〕 상대 엔트로피를 이용한 각 공격 유형의 유용한 척도 선택

방법	선택 척도
Normal	*2, 3, 4, 10, 12, 23, 24, 25, 27, 29, 31, 32, 33, 34, 35, 37, 40, 41*
DoS	*2, 3, 4, 10, 12, 23, 24, 25, 27, 29, 31, 32, 33, 34, 35, 37, 40, 41*
Probes	*2, 3, 4, 10, 12, 23, 24, 25, 27, 29, 31, 32, 33, 34, 35 37, 40*
R2L	*2, 3, 4, 10, 12, 23, 25, 32, 33, 34, 35, 37, 40, 41*

〔표 6〕 관련연구 방법과 제안 방법의 결과 비교

방법 / 대상	분류 개수	분류율	오탐 개수	오탐율
정보 획득량	285912	0.919452	642	0.010595
카이 제곱	285912	0.919452	642	0.010595
표준편차	287135	0.923385	335	0.005529
랜덤포레스트	229853	0.739175	60593	1.000000
LDA	286708	0.922012	689	0.011371
ICA	287841	0.925656	576	0.009506
PCA	287303	0.923926	1056	0.017428
상호 정보량	285454	0.91798	1353	0.022329
CBFS	287867	0.925739	368	0.006073
마코브 블랭킷	286199	0.920375	776	0.012807
CART	286778	0.922237	585	0.009655
러프셋	287667	0.925096	313	0.005166
상대엔트로피	288630	0.92819	335	0.005529

상대 엔트로피 방법은 선택된 척도 개수가 가장 적으면서, 분류율과 오탐율에서도 가장 좋은 결과를 보여, 확률 분포 추정에서 이 방법을 이용한다. [표 5]는 상대 엔트로피를 적용하여 각 공격 유형의 유용한 척도 선택 결과를 보여준다.

본 논문에서 방법의 유용성을 증명하기 위해 관련연구에서 제안한 방법에 의해 선택된 척도를 본 논문의 실험에 사용한 C4.5를 사용하여 규칙을 생성하고 침입 분류 실험을 수행하였다. 관련연구에서 제안한 방법들 중, 척도 선택 개수가 적은 방법이 존재하지만, 제안한 방법의 분류율과 오탐율을 비교해 본 결과, 제안된 방법이 가장 좋은 성능을 나타내어, 다른 연구에 비해 보다 정확하고 유용함을 확인할 수 있었다.

V. 결 론

네트워크 트래픽 데이터의 증가로 인해 네트워크 침입 탐지 시스템은 증가하는 방대한 양의 트래픽을 효율적으로 처리하기 위해 데이터 감소 방법이 요구되고 있다. 본 논문에서는 데이터 필터링, 척도 선택, 데이터 군집화와 같은 데이터 감소 방법 중에서, 침입 탐지에 유용한 척도 선택 방법을 제안함으로써 학습 및 탐지 시에 처리해야할 네트워크 트래픽 데이터의 양을 감소시키고, 침입 탐지 및 분류의 정확성을 유도한다.

본 논문에서는 침입 탐지 및 분류의 정확하고 빠른 처리를 위해, 다양한 거리측정 방법을 적용하여 척도를 선택하고 규칙을 생성하여 탐지 실험을 하였다. 본 실험의 데이터로 KDD CUP 99 데이터 셋을 사용하였고, 학습하지 않은 공격에 대해서도 분류를 실험하였으며, 개별 공격의 특징이 아닌 공격 유형의 특징을 찾아내기 위해 KDD CUP 99 데이터 셋의 여러 가지 공격들을 DoS, Probes, R2L의 공격 유형으로 분류하였다. 그리고 각 공격 유형의 41가지 척도에 대한 값을 히스토그램 방법에 적용하여 확률 분포를 계산했으며, 유클리디언 거리, 카이 제곱 테스트, 상대 엔트로피, J-S 거리 방법을 이용하여 정상과 각 공격 유형의 특징을 나타내는 유용한 척도를 선택한 후, 결정 트리 알고리즘 중 하나인 C4.5를 이용하여 탐지 규칙을 생성하고, 분류 실험하였다. 네 가지 거리 측정 방법을 이용하여 거리값이 낮은 척도를 하나씩 제거하여 한 개가 남을 때까지 반복해서 실험하여 선택된 척도, 분류율, 오탐율 등을 비교한 결과, 상대 엔트로피가 모든 면에서 가장 좋은 성능을 보임을 확인할 수 있었다. 본 논문에서 제시한 방법으로 실험한 결과, 침입탐지 시스템에서 처리해야할 데이터 양이 감소했고, 탐지 및 분류 정확성이 증가하여 효율적임을 알 수 있었다.

참고문헌

[1] Richard P. Lippmann and Robert K. Cunningham, "Improving intrusion detection performance using keyword selection and neural networks," *Computer Networks*, 2000.

[2] K. Das, *Attack Development for Intrusion Detection* Evaluation, MIT Master's Thesis, June 2000.

[3] Eric Cole and Hackers Beware, The *Ultimate Guide to Network Security*, New Riders, August 2001.

[4] Advanced Networking Management Lab (ANML), *Distributed Denial of Service Attacks(DDoS) Resources,* http://anml.iu.edu/ddos/index.html, 2008.

[5] Dorothy E. Denning, "An Intrusion-Detection Model," *IEEE Transactions on Software Engineering,* Vol. SE-13, NO. 2, 1987.

[6] G. Lipens and H. Vacaro. "Anomaly Detection: Purpose and Framework," *Proceedings, 12th National Computer Security Conference*, 1989.

[7] Hervé Debar, Marc Dacier, and Andreas Wespi, "Towards a taxonomy of intrusion-detection systems," *Computer Networks* 31(8), 1999.

[8] T.F.Lunt, R.Jagannathan, R.Lee et al., "IDES: The enhanced prototype, A real-time intrusion detection system," Technical Report SRI Project 4185-010, *SRI-CSL-88-12*, CSL SRI International, Computer Science Laboratory, October 1988.

[9] Hervé Debar, Monique BECKER and Didier SIBONI, "A Neural Network Component for an Intrusion Detection System," *Proceedings of 1992 IEEE computer society symposium on research in security and privacy.* 1992.

[10] J. Ryan, M. Lin, "Intrusion detection with neural networks," *AAAI workshop*, 1997.

[11] W. Lee, S. J. Stolfo, and K. W. Mok, "Mining in a data-flow environment: Experience in intrusion detection," *In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1999.

[12] Aleksandar Lazarevic, Aysel Ozgur, Levent Ertoz, Jaideep Srivastava, and Vipin Kumar, "A comparative study of anomaly detection schemes in network intrusion detection," *Proceedings of Third SIAM Conference on Data Mining*, 2003.

[13] Ajith Abraham and Dan Steinberg, "MARS: still an alien planet in soft computing?," *Lecture notes in computer science 2074.* 2001.

[14] Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham and Vitorino Ramos, "Intrusion detection systems using adaptive regression splines," *Sixth international conference on enterprise information systems.* ICEIS'04, 2004.

[15] Srilatha Chebrolu, Ajith Abraham and Johnson P. Thomas, "Hybrid Feature Selection for Modeling Intrusion Detection Systems," *Neural Information Processing 11th International Conference*, 2004.

[16] Yang Li, Bin-Xing Fang, You Chen and Li Guol, "A Lightweight Intrusion Detection Model Based on Feature Selection and Maximum Entropy Model," *Communication Technology, International Conference on*, 2006.

[17] Gopi K. Kuchimanchi, Vir V. Phoha, Kiran S. Balagani, and Shekhar R. Gaddam, "Dimension Reduction Using Feature Extraction Methods for Real-time Misuse Detection Systems," *Proceeding of the 2004 IEEE Workshop on Information Assurance and Security*, 2004.

[18] Jiong Zhang and Mohammad Zulkernine, "Network Intrusion Detection using Random Forests," *Proc. of the Third Annual Conference on Privacy, Security and Trust*, 2005.

[19] V. Venkatachalam and S. Selvan, "Performance Comparison of Intrusion Detection System Classifiers using Various Feature Reduction Techniques," *I.J. of SIMULATION*, Vol. 9 No 1, 2008.

[20] Srilatha Chebrolu, Ajith Abraham and Johnson P. Thomas, "Hybrid Feature Selection for Modeling Intrusion Detection Systems," *Neural Information Processing 11th International Conference*, 2004.

[21] Srilatha Chebrolu, Ajith Abraham, and Johnson P. Thomas, "Feature deduction and ensemble design of intrusion detection systems", *Computers & Security*, Volume 24, Issue 4, 2005.

[22] T. S. Chou, K. K. Yen, and J. Luo, "Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms," *International Journal of Computational Intelligence*, Volume 4 Number 3, 2007.

[23] Anazida Zainal, Mohd Aizaini Maarof and Siti Mariyam Shamsuddin, "Feature Selection Using

Rough Set in Intrusion Detection," *TENCON.* 2006 IEEE Region 10 Conference, 2006.

[24] Richard P.Lippmann, David J. Freid et al., "Evaluating Intrusion Detection System: The 1998 DARPA off-line Intrusion Detection Evalution," *Computer Security Applications Conference,* 1991. Proceedings., Seventh Annual, 1999.

[25] KDD CUP 99 DATA Available in http://kdd.ics.uci.edu/databases/kddcup99/kddcup99. html.

[26] Richard O. Duda, Peter E. Hart and David G. Stork, *Pattern Classification, 2nd Edition,* Wiley-Interscience, 2000.

[27] Gil-Jong Mun, Yong-Min Kim, Dongkook Kim, and Bong-Nam Noh, "Network Intrusion Detection Using Statistical Probability Distribution," *The 2006 International Conference on Computational Science and its Applications*, 2006.

[28] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

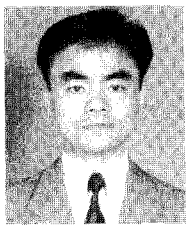
[29] 정일안, *분류 기법을 이용한 네트워크 공격 침입 분류의 자동생성*, 전남대학교 대학원, 2004.

[30] 이성권, *확률분포거리 기반 분류에 의한 악성 봇과 웜 탐지패턴 자동 생성*, 전남대학교 대학원, 2006.

〈著者紹介〉

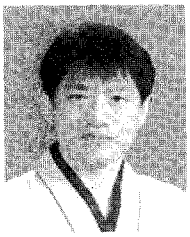
문 길 종 (Gil-Jong Mun)
정회원

2006년 2월: 전남대학교 대학원 정보보호협동과정(이학석사)
2009년 2월: 전남대학교 대학원 정보보호협동과정(이학박사)
2009년 1월~현재: (주)정보보호기술 선임연구원
<관심분야> 네트워크 보안, 침입탐지, 정보보호 등



김 용 민 (Yong-Min Kim)
종신회원

2002년 8월: 전남대학교 대학원 전산통계학과(이학박사)
2004년 3월~2006년 2월: 여수대학교 정보기술학부 전임강사
2006년 3월~현재: 전남대학교 문화콘텐츠학부 조교수
<관심분야> 시스템 및 네트워크 보안, 전자상거래 보안 등



노 봉 남 (Bongnam Noh)
종신회원

1978년 2월: 전남대학교 수학교육과 졸업(학사)
1982년 2월: KAIST 전산학과 졸업(석사)
1994년 2월: 전북대학교 대학원 전산과 졸업(박사)
1983년~현재: 전남대학교 전자컴퓨터공학부 교수
2000년 현재: 전남대학교 시스템보안연구센터 소장
<관심분야> 컴퓨터와 네트워크 보안, 개인정보보호, 사이버사회와 윤리