情報保護學會論文誌
第13券 第2號, 2003. 4

데이터마이닝 기법을 이용한 비정상행위 탐지 방법 연구

박 광 진*, 유 황 빈**

Anomaly Detection Scheme Using Data Mining Methods

Park Kwang-jin*, Ryou Hwang-bin**

요 약

네트워크 환경에서의 다양한 침입은 심각한 위험을 초래 할 수 있기 때문에 침입을 효과적으로 탐지하기 위해 데이터마이닝 기법을 발전시켜 왔다. 비정상행위 탐지 기술은 순수 데이터로 학습한 후, 비정상행위를 탐지하기 때문에 정교한 정상행위 패턴 생성이 필수적이다. 순수한 학습 데이터의 생성은 시간과 비용이 많이 드는 단점이 있다. 따라서 네트워크 상의 데이터에 대한 특징을 파악하는 것이 중요하다.

본 논문에서는 데이터마이닝의 연관규칙 및 클러스터링기법을 비정상행위 탐지에 적용하였고, 패킷내의 판정 요소에 정보이론 척도를 적용하여 불필요한 데이터를 필터링하는 방법을 제시하였다. 또한 가변길이 트랜잭션을 네트워크 상의 분석 단위를 정의하는 기준으로 제시하여 행위 패턴 생성에 보다 묘사성이 높음을 보였다.

ABSTRACT

Intrusions pose a serious security risk in a network environment. For detecting the intrusion effectively, many researches have developed data mining framework for constructing intrusion detection modules. Traditional anomaly detection techniques focus on detecting anomalies in new data after training on normal data. To detect anomalous behavior, precise normal pattern is necessary. This training data is typically expensive to produce. For this, the understanding of the characteristics of data on network is inevitable.

In this paper, we propose to use clustering and association rules as the basis for guiding anomaly detection. For applying entropy to filter noisy data, we present a technique for detecting anomalies without training on normal data. We present dynamic transaction for generating more effectively detection patterns.

Keyword : *Intrusions detection, data mining, anomaly detection, misuse detection, training data, detection patterns*

Ⅰ. 서 론

최근 컴퓨터 기술의 급속한 발전으로 인해 기존의 텍스트 위주의 사용자 환경에서 벗어나 이미지, 그래픽, 오디오 및 비디오 데이터 등을 제공하는 멀티미디어 사용자 환경으로 변환하고 있다.

침입탐지는 네트워크 보호나 방어에 중요한 요소이다. 침입탐지시스템은 네트워크의 행위 데이터(BSM이나 tcpdump 등) 정보를 분석하여 침입 여부를 판정한다. 침입탐지의 주요 방식은 오용탐지와 비정상행위 탐지가 있다. 오용탐지는 공격 행위 패턴인 시그네쳐를 이용하여 탐지하는 것으로, 알려지지 않은 새로운 공격에는 취약하다.

비정상행위 탐지는 네트워크의 데이터로부터 정상적인 행위 모델을 만들어, 이 정상 모델로부터 얼마나 벗어나 있는가를 찾아내어 탐지하는 것으로 새롭

* 광운대학교 컴퓨터과학과(kjpark@kisa.or.kr)
** 광운대학교 컴퓨터과학과(ryou@kwangwoon.ac.kr)

고 알려지지 않은 공격에 유용하다.

비정상행위 탐지의 기본 전제는 네트워크 데이터가 뚜렷한 특징과 규칙성이 있다는 것이다. 즉 지속적인 정상행위와 듬성듬성한 비정상행위가 일어난다는 것이다. 비정상행위 탐지 모델을 구축하는 과정은 우선 데이터의 특징을 파악하고, 다음에 특성에 가장 잘 어울리는 모델을 선택한다. 그러나 비정상행위 탐지는 전문가의 지식에 의존하는 등 탐지 방법이 매우 제한적 이였기 때문에 네트워크의 복잡도를 따라가기가 힘들다. 따라서 네트워크의 데이터에 깊숙이 파묻힌 규칙성을 찾아 모델을 구축하는 것이 중요하다[1].

일반적으로 비정상행위 탐지 모델은 순수 데이터(비정상이 포함되지 않은 정상 데이터)의 학습에 의해 구축된다. 이 모델의 성능은 모델링 방법과 가능한 학습 데이터의 질과 양에 크게 좌우된다[2]. 이 방법은 순수 데이터로 비정상행위 탐지 모델을 학습하는 것으로 몇 가지 결점이 있다. 첫째, 순수 데이터를 얻기가 쉽지 않다. 둘째, 불필요한(noisy or anomalous data)데이터에 의한 학습은 심각한 결과를 초래한다. 즉 학습 데이터에 침입이 숨겨져 있다면 공격을 정상으로 판정하는 모델이 만들어지게 된다. 셋째, 실시간으로는 학습 대상 데이터가 순수 데이터라고 보장하기 어렵다.

데이터내에서 불필요한 데이터를 찾는 것은 첫째, 불필요한 데이터를 찾아내서 제거하면, 남아있는 순수 데이터를 비정상행위 탐지 모델에 적용할 수 있다. 둘째, 비정상행위는 이벤트 상에서 거의 잘 나타나지 않으므로 그 자체가 관심 대상이다[3].

본 논문에서는 판정 단위 요소로 고정길이 트랜잭션과 가변길이 트랜잭션을 비교·분석하여 가변길이 트랜잭션이 판정 요소를 분석하는 묘사성이 높음을 제시하였다. 또한 네트워크의 패킷 데이터내에서 불완전한 데이터를 제거하여 정상행위 패턴 생성이 가능토록 하였고, 데이터마이닝의 연관규칙 및 클러스터링을 이용하여 비정상행위의 탐지 영역을 분석하여 시스템 설계 방향을 제시하였다.

Ⅱ. 관련연구

침입탐지시스템은 데이터 수집 및 가공, 분석 및 침입탐지, 보고 및 대응 등 일련의 과정으로 이루어진다. 이 중 수집된 데이터를 가공하여 분석하기 위해서는 분석 단위의 결정이 중요하다. 대부분 정적 시간 단위의 패킷량에 의한 고정길이 트랜잭션을 분석 단위로 이용하였고, 가장 적절한 시간 간격을 정하기 위한 연구도 이루어져 왔다[1,4].

비정상행위 탐지 모델은 사용자의 행위 패턴 변화를 통해서 침입을 탐지하는 방법이다. 과거의 경험적인 자료에서 통계적인 값으로 변환하여 비정상행위를 판단하기 때문에 자료의 양이 많을수록 정확하게 침입을 탐지할 수 있다[5].

데이터마이닝 기법은 프로그램과 사용자 행위를 설명하는데 필요한 특징 패턴을 추출한다. 데이터마이닝은 결과에 대한 유용성과 불확실성을 정량화 할 수 있어야 하며, 수행 결과에서 수많은 패턴 및 새로운 정보를 얻게 된다[6].

일반적인 비정상행위 탐지 기술은 순수 데이터로 학습한 후, 새로운 데이터에서 비정상행위를 찾는데 초점을 맞추었다[7]. 최근에는 순수 데이터에 의한 학습 없이, 비정상행위를 찾는 방법으로 데이터의 확률분포 측정[8], 벡터 스페이스 이용 방법[9]등이 제안되었다.

Ⅲ. 침입탐지와 데이터마이닝

3.1. 침입탐지시스템의 정의

침입이란 컴퓨터가 사용하는 자원의 무결성, 비밀성, 가용성을 저해하는 일련의 행위 집합을 말한다. 침입탐지시스템은 시스템의 비정상적인 사용, 오용, 남용 등을 알려주는 시스템이다. 이러한 시스템은 감사 기록, 네트워크의 트래픽 기록 등으로부터 사용자 행위에 대한 정보를 분석함으로써 수행된다. 그리고 침입탐지시스템의 목표는 침입자에 대한 불법적인 사용을 명시하는 것이고, 다른 하나는 합법적인 사용자에 의한 오용이나 남용을 알아내는 것이다[10].

3.2 침입탐지의 범위와 분류

침입탐지 방법에는 호스트에서 감사 로그를 수집하고 내용을 분석하는 호스트 기반 탐지와 네트워크를 통해서 정보를 중앙 감시 시스템에 모아 분석하는 네트워크 기반 탐지로 나눌 수 있다. 어떠한 내용을 탐지할 것인가에 따른 분류로는 시스템이나 응용 프로그램의 약점을 통해서 시스템에 침입할 수 있는 잘 정의되고 알려진 공격을 탐지하는 오용탐지와 컴

퓨터 자원의 비정상적인 행위나 사용에 근거한 알려지지 않은 공격을 탐지하는 비정상행위 탐지로 나눌 수 있다.

비정상행위 탐지 방법으로는 대부분 임계값을 두고 그 값을 기준으로 판정하는 통계적 방법을 이용하고 있다. 특히 데이터마이닝 기법이 침입탐지에 도입되어, 대규모 감사 데이터베이스에서 그 특징을 추출하여 오용 시그네처나 정상행위 프로파일을 생성하여 침입탐지 판정에 이용되고 있다. 데이터마이닝 기법은 알려진 공격 패턴을 생성하여 판정하는 오용탐지 보다는 정상행위간의 연관성 및 유사성 등 그 특징을 추출하는 묘사성이 뛰어남으로 네트워크 상에서의 비정상행위 탐지에 적합하다. 특히 비정상행위 탐지에 데이터마이닝 기법을 적용하면 다양한 각도에서 네트워크 사용 형태를 연관규칙이나 클러스터링 기법을 통해 정상행위 프로파일 생성이 가능하므로 보다 정확한 판정을 기대할 수 있다.

3.3 데이터마이닝 알고리즘

RIPPER[4]는 자동적이고 적응성 있는 탐지 모델 구축에 데이터마이닝 기법을 적용한 대표적인 툴이다. 중심 아이디어는 각 네트워크 접속이나 호스트 세션을 묘사하는 일련의 특징을 추출하는 감사 프로그램을 만드는 것이다. 그리고 데이터마이닝 프로그램을 침입과 정상행위로 구분하여 정확히 끄집어내는 규칙을 학습하는데 적용하였다.

연관규칙(association rule)은 대용량의 자료가 기록되어 있는 데이터베이스에서 자주 발생하는 항목간의 상호 연관성을 찾아내는 방법이다[11]. X→Y, [C,S], C : Confidence, S : support 형식으로 표현되며 X∩Y = 0, S = support(X∪Y),

$$C = \frac{support(X \bigcup Y)}{support(X)} \quad (1)$$

클러스터링(Clustering)은 대용량의 사건들이 기록되어 있는 데이터베이스에서 유사 작업 군을 탐색하는 기법이다. 기존의 클러스터링 기법에서는 트랜잭션 정보를 이용하지 않고 클러스터를 생성하였다. 하지만 작업 단위가 트랜잭션 단위이기 때문에 침입탐지 환경에서는 트랜잭션 정보가 사용자의 정상행위 패턴 생성에 있어서 매우 중요하다.

본 논문에서는 비정상행위도[12]를 이용하여 네트워크 사용자의 행위 패턴을 생성토록 하였다. 비정상행위도는 네트워크 행위 트랜잭션이 정상행위 패턴과 유사하다면 비정상행위도가 낮게 나타나고 반면, 이상 행위를 하였을 경우에는 비정상행위도가 높게 나타난다. 특히 클러스터링인 경우, 하나의 판정 요소에 대해서 클러스터가 두 개 이상 생성되었을 경우는 두 클러스터 사이의 중심점을 구하여 가까운 위치의 클러스터와 비정상행위도를 구하게 된다.

Ⅳ. 침입탐지시스템의 모델링

4.1 학습데이터의 생성

일반적으로 정상행위 모델을 생성하기 위해서는 정상행위만으로 이루어진 데이터를 가지고 학습해야 한다. 따라서 정상행위 패턴 생성에는 비정상행위 또는 불필요한 데이터가 섞이지 않는 순수 데이터가 필요하나, 일반적인 네트워크 상에서의 패킷 데이터는 순수 데이터로 이루어져 있다고 보기가 힘들다. 따라서 인위적으로 순수 데이터를 생성해서 학습 데이터를 만들어야 한다.

또한 Eskin[8]의 확률 분포를 이용한 비정상행위 탐지 방법은 비정상행위와 정상행위는 뚜렷이 구별되어야 하고, 비정상행위 수는 정상행위 수에 비해 매우 적어야 한다는 전체 조건이 성립해야 한다. 이 경우 정상행위와 비슷한 수의 공격이 일어나는 Syn-flood DOS 등의 공격 유형에는 취약하다.

본 논문에서는 네트워크 상의 일반 패킷 데이터에서 불필요한 데이터를 추출하고 난 후에 학습하는 방법을 제시하고자 한다. 우선 탐지 모델을 만들기 위해서는 데이터 내에서의 비정상행위를 찾아야 한다. 왜냐 하면 모델을 왜곡시킬지도 모르는 비정상행위를 제거함으로서 보다 나은 모델을 만들 수 있기 때문이다. 이를 위해 엔트로피 값을 데이터 내의 비정상행위를 찾기 위한 수단으로 이용한다.

4.1.1 엔트로피의 특징

전형적인 엔트로피의 해석은 분류된 데이터를 엔코드 하는데 요구되는 비트수라 할 수 있다. 엔트로피 값은 클래스 분포가 몰려 있거나 데이터가 순수하면 적어진다. 예를 들어 모든 데이터 아이템이 하나의 클래스에 속하면 엔트로피는 0이고, 수신 측은 오직 하나의 출력만 있게 된다. 엔트로피 값이 크다는 것은 클래스의 분포가 많고 데이터가 덜 순수하

다는 것이다.

비정상행위탐지에서의 엔트로피는 수집 데이터의 규칙성(regularity)을 측정하는데 사용할 수 있다. 엔트로피가 적으면 서로 다른 레코드 수가 적다는 것으로, 중복성 또는 규칙성이 높기 때문에 미래의 이벤트를 예측하는데 도움을 준다. 왜냐하면 많은 이벤트는 계속 반복되기 때문이다.

그러므로 비정상행위탐지 모델은 보다 적은 엔트로피를 가진 데이터를 이용하여 구축하면 더 나은 성능을 가질 수 있다. 즉, 수집 데이터가 하나의 이벤트를 가지면 이외의 모든 것은 공격으로 간주되고, 수집 데이터가 많은 이벤트를 가지게 되면 더욱더 복잡한 모델이 필요하게 된다[13].

4.1.2 불필요한 데이터의 필터링

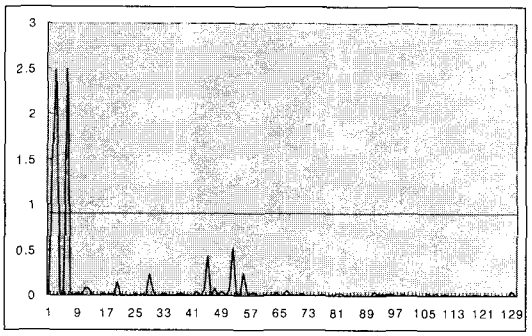
비정상행위는 데이터 내의 규칙성을 계산함으로서 찾게 된다. 정상적인 데이터에는 본질적인 규칙성이 존재하고, 경험적으로 규칙성이 좋으면 좋을수록 성능도 좋아진다.

본 논문에서는 정상적인 TCP 프로토콜 상에서 서비스 종류가 같고 연결 설정이 완료된 경우, source/destination에 대한 규칙성의 척도를 적용해 보고자 한다. 일반적으로 하나의 이벤트내에서는 시작 source address는 다음 destination address와 같은 경우가 많다. 각 이벤트에서 t1=t2이라 하면, t1=x, t2=y로 하여 조건부 엔트로피를 계산하여 일정 임계값 이상인 것은 불필요한 데이터로 간주하게 된다. 계산식은 $H(X/Y)=-\sum P(x,y)\log P(x/y)$이다[2].

본 논문에서는 약 10만개의 패킷을 Tcpdump한 자료를 이용하였고, 분석 단위 트랜잭션 중에서 임계치를 초과하는 하나의 트랜잭션을 선택하여, 그 트랜잭션내의 각 이벤트에 대한 엔트로피 값을 측정하였다. [그림 2]에서 임계치 이상인 경우는 불필요한 데이터로 간주하게 된다.

```
Dynamic Filtering : Interval = 1:255743
Max = 10101, Avg = 1911, Bound = 9696
Dynamic Filtering : Bound packet count = 9696
Dynamic Filtering : START
Session ID = 34, Total = 10101
 2 : 210.92.0.222   ⇓◊ 203.240.169.50 : entropy=1.545392
 3 : 203.240.169.50 ⇓◊ 80.117.71.100 : entropy=2.486882
 6 : 211.196.154.175⇓◊ 203.240.169.88 : entropy=2.491832
Dynamic Filtering : END
```

〔그림 1〕 동적 필터링의 결과



〔그림 2〕 동적 필터링의 결과 그래프

Interval : 패킷을 자른 단위의 시간
Max : 가장큰 묶음의 패킷 개수
Avg : 묶음들의 평균 패킷 개수
Bound : 상위 95% 이상 되는 패킷을 자르는 기준 (즉, bound이상이 되는 개수의 묶음은 필터링 대상이 된다)
Session ID : 필터링 대상이 되는 패킷 묶음의 번호
Total : 필터링 대상이 되는 묶음 내의 패킷 개수

4.2 가변길이 트랜잭션의 적용

정상행위 패턴을 생성하기 위해서는 분석 단위에 따른 패킷 데이터의 수집이 필요하다. 본 논문에서는 가변길이 트랜잭션(동적시간 윈도우)을 네트워크 상의 의미적인 트랜잭션 단위를 정의하는 기준으로 제시하고자 한다. 즉 네트워크 상에서 n개의 패킷이 전송되어 올 때 패킷간의 시간 간격이 주어진 임계치를 넘지 않는 패킷은 하나의 트랜잭션으로 묶는다.

4.2.1 입력 정보 및 네트워크의 이상 징후

네트워크 상의 정보는 두 가지 형태로, 패킷해더 정보와 그 내용(payload)이다. 해더 정보는 빌딩에서 왔다갔다하는 사람이 누구인지를 나타내고, 내용은 무엇을 나르는지를 나타낸다. 내용은 암호화되거나, 정보 내용의 공통성을 찾기가 어렵기 때문에 입력 정보로 사용하기 위해서는 복잡한 특징 추출이 필요하다.

침입탐지시스템의 입력 정보로 패킷을 이용한 경우의 이점은 서비스나 사용자의 이벤트에 대해 로그는 사후에 생성되는데 반해, 패킷은 송신되는 이벤트를 직접 탐지하기 때문에 실 시간성이 높다. 네트워크 상에서 Tcpdump 등을 이용하여 해더에서 추출할 수 있는 데이터 형태로는 패킷타입, 수신지/도착

지 주소, 연결을 시도하는 포트번호, 서비스 포트번호 등이다.

네트워크 상에서의 공격에는 공격시 이상 징후가 나타나는 것이 있는 반면, 정당한 패킷을 사용하는 공격은 공격이라고 식별하기가 어렵다. 그러나 피해자가 이러한 패킷을 받았을 때 우리가 인식할 수 있는 패킷 시퀀스를 발견할 수도 있다. 프로토콜 규약이나 보안 취약점을 이용하여, 거짓 IP 주소로부터 온 위조된 패킷, 중복되거나 소스/목적지 주소가 같은 패킷, 일련의 시퀀스 없이 보낸 TCP 패킷이나 유효하지 않는 TCP 시퀀스 번호를 갖는 패킷 등은 IP 프로토콜의 불안정을 초래한다.

공격시 공격자나 피해자에 의해 만들어진 패킷은 비정상을 탐지할 수 있는 몇 가지 이상 징후들이 나타난다. 즉 IP 프로토콜이 불안하고, IP 서비스가 불안정하고 대량의 트래픽을 야기한다. 이 경우 트래픽 양의 곡선이 일정한 형태를 유지하지 않고, 특히 트래픽이 정점에 다다를 때 인바운드 트래픽과 아웃바운드 트래픽 차이가 매우 크게 나타난다[14].

4.2.2 시간 간격에 따른 패킷량 분석

네트워크를 통해서 이동되는 데이터의 양은 패킷 단위로 전달되고 있으며, 유통되는 패킷의 양은 네트워크의 특성에 따라 다양한 양상을 보이고 있다. 시간에 따른 패킷량을 조사하기 위해 첫째, 고정단위시간에 전송되는 패킷의 양 둘째, 패킷의 전송 사이사이에서의 긴 휴지기를 단위로 구분하는 방법을 비교하였다.

가변길이 분석 단위는 전송되고 있는 패킷 사이의 시간적인 간격을 이용하여 동적으로 패킷의 량을 조사하는 것으로, 모든 패킷의 시간 간격을 순위 화하여 상위 5%의 순위에 들어오는 시간을 선택하였다. 그 결과 약 1.80초 이상의 시간 간격이 있는 경우만을 구분하여 트랜잭션을 산출한 결과, 49개의 분석 단위를 얻었다.

고정길이 분석 단위는 패킷을 구분하는 것은 단지 정적인 시간뿐이지만, 본 실험에서는 동적인 패킷량 조사와 같은 트랜잭션 개수로 맞추기 위해 49개의 분석 단위로 나누었다.

[그림 4]의 정적 패킷량에서는 한 개의 패킷량을 제외하고는 균일함을 나타내지만, [그림 3]의 동적 패킷량에서는 편차가 심한 패킷의 유동 량을 보이고 있다. 이 이유는 정적 패킷량은 이전의 트랜격션이 다음 트랜잭션에 포함되어지기 때문에 각 트랜잭션 간 구분이 모호해 진다. [그림 3]에서는 각 트랜잭션 간 분석 단위가 명확하기 때문에 독립적인 특징을 가지게 된다. 따라서 동적패킷량이 판정 요소를 표현하는 묘사성이 높다고 할 수 있다.

특히 네트워크 상의 행위 트랜잭션은 수시로 일어나기 때문에 의미 있는 단위의 트랜잭션을 만들기 위해서는 동적패킷량을 이용한 가변길이 트랜잭션이 적합하다고 할 수 있다. 또한 고정길이 트랜잭션을 사용하는 경우, 탐지 시간에 있어 고정길이 트랜잭션 크기만큼 탐지 시간이 지연되는 반면 가변길이 트랜잭션을 사용하는 경우 실시간 탐지가 가능하다.

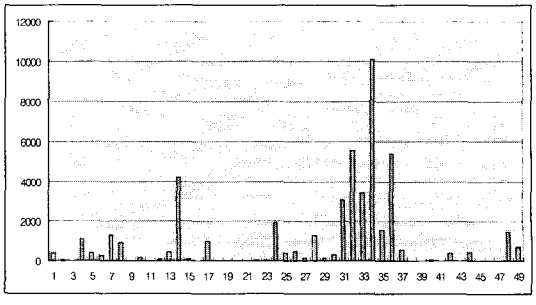
4.3 비정상행위도 산출

본 논문에서는 침입탐지 실험을 위해 가변길이 트랜잭션을 탐지 분석 단위로 이용하였으며, 일반 데이터를 가지고 불필요한 데이터를 필터링하여 정상행위 데이터를 만들었다. 그리고 수차례의 최적화 과정을 거쳐 정상행위를 가장 잘 표현할 수 있는 대표적인 프로파일을 가지도록 하였다.

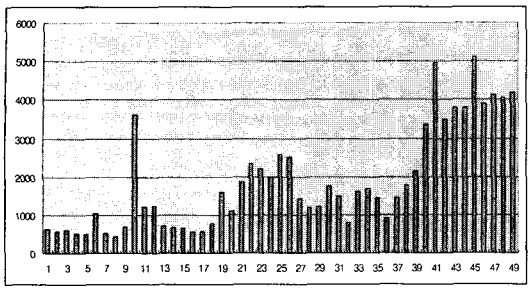
테스트 데이터 : 오용침입과 네트워크 비정상행위데이터

- 오용데이터 : Dos(SYN/UDP/ICMD Flooding), Probing attack
- 비정상행위 데이터

Anomaly1 : 평소보다 많은 다량의 서비스 발생
Anomaly2 : 평소 사용하지 않았던 서비스 발생
Anomaly3 : 낯선 호스트로부터의 로그인 발생
Anomaly4 : 잘 사용하지 않는 포트에 접속
Anomaly5 : 근무 시간외 트래픽 발생



〔그림 3〕 동적인 패킷량



〔그림 4〕 정적인 패킷량

	sysgal	pscan	solaris_land	newpop	anomal1	anomal2	anomal3	anomal4	anomal5	
misuse	■	■	■	■						threshold
clustering	14	38	3	14	80	3	3	44	60	5
association	80	70	69	72	30	80	70	80	30	40

〔그림 5〕 오용행위 및 비정상행위도

본 실험에서는 오프라인테스트를 통해 나온 임계치 에 따라 탐지 여부를 결정하였다. 클러스터링인 경우 평균적으로 정상행위도가 5에 수렴하였고, 연관규칙은 40에 수렴하였다. 이러한 수렴 여부에 따라 각 모델에 있어서 임계치를 5와 40으로 결정하였다. 클러스터링인 경우 임계치가 5이하인 경우는 오용행위 중 solaris_land인 경우이며 비정상행위에 대한 임계치 이하 값을 가지는 경우는 평소 사용하지 않은 서비스에 대한 anomal2와 낯선 호스트로부터의 로그인을 발생하는 anomal3에 대해 낮은 비정상행위도를 보였다. 이러한 결과는 분포와 빈도에 따른 유사성을 가지는 클러스터링의 특성을 잘 나타내 주고 있다. 연관규칙인 경우 오용행위에 대해서는 비정상행위도가 높은 값을 가지며, 비정상행위 경우 평소보다 많은 다량의 서비스를 발생하는 anomal1에 대해서는 클러스터링에 비해 낮은 비정상행위도가 나왔다. 클러스터링에서 임계치 값 이하를 가지는 anomal2와 anomal3에 대해서는 높은 비정상행위도가 나왔으며 근무 시간외 트래픽을 발생하는 비정상행위테스트에서는 클러스터링보다 낮은 비정상행위도를 보였다. 이 실험을 통해서 알 수 있는 것은 클러스터링의 경우 빈도와 분포의 성격을 가지는 행위에 대해서 묘사성이 높고, 연관규칙은 빈도와 분포의 성격보다는 데이터의 의미있는 연결성을 가지는 행위에 대해 묘사성이 뛰어남을 알 수 있었다.

V. 결 론

본 논문에서는 네트워크 상에서 공격을 정확하고 빠르게 탐지하기 위해 비정상행위탐지 방법을 제안하였다. 비정상행위탐지 모델을 정확히 만들기 위해서는 순수한 정상 데이터 만으로 학습하여야 하나, 실제 순수 데이터를 구하기가 어렵다. 이 문제를 해결하기 위해 네트워크 상 송수신되는 패킷의 성질을 엔트로피 값으로 산출하여 불필요한 데이터를 필터링하는데 이용함으로서, 일반 데이터를 가지고도 정상행위 패턴을 생성하는 방법을 제시하였다. 특히 가변길이 트랜잭션을 분석 단위로 이용하여, 보다 정확한 판정 요소를 산출하는 방법을 제시하였다.

본 연구에서는 방대한 데이터 분석을 좀 더 지능적이고 자동적으로 수행하기 위해서 데이터마이닝 기법인 클러스터링과 연관규칙을 활용하였다. 오용행위와 비정상행위에 대한 탐지 영역을 분석하기 위해 대표적인 DoS(Denial of Service)와 Probing, 비정상행위를 발생시켜 실험을 하여, 알고리즘 특성에 따라 서로 다른 탐지 영역을 가지고 있음을 보였다.

향후 연구로는 실시간으로 정상행위 패턴을 학습할 수 있도록 가변길이 트랜잭션의 분석 단위 자동설정 방법과 보다 순수한 데이터의 생성을 위해 다양한 판정 요소의 적용에 대한 연구가 필요하다고 하겠다.

참 고 문 헌

[1] Wenke Lee, Salvatore J. Stolfo. "Data Mining Approaches for Intrusion Detection" Proceedings of the 7th USENIX security Symposium, Texas, 1998.

[2] Eleazar Eskin, Wenke Lee, Salvatore J. Stolfo. "Modeling System Calls for Intrusion Detection with Dynamic Window Sizes", 2001.

[3] Leonid Portnoy, Eleazar Eskin and Salvatore J.

Stolfo. "Intrusion detection with unlabeled data using clusterling" To Appear in Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA -2001). Philadelphia, PA: November 5~8, 2001.

[4] Wenke Lee. "A data mining framework for constructing features and models for instruction detection systems", 1999.

[5] Harold S. Javitz and Alfonso Valdes, "The NIDES Statistical Component Descriptio and Justification", Annual report, SRI International, 1994.3.

[6] Wenke Lee, Sal Stolfo, and Kui Mok. "A Data Mining Framework for Building Intrusion Detection Models" In Proceedings of the 1999 IEEE Symposium on Security and Privacy, Oakland, CA, May 1999.

[7] W. Lee, S.J. Stolfo, and K. Mok. "Data mining in work flow environments : Experiences in intrusion detection." In Proceedings of the 1999 Conference on Knowledge Discovery and Data Mining (KDD-99), 1999.

[8] Eleazar Eskin, "Anomaly Detection over Noisy Data using Learned Probability Distributions" ICML00, Palo Alto, CA : July, 2000.[abstract, full paper] Applications. Kluwer 2002[full paper, PDF].

[9] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy and Salvatore Stolfo. "A Geometric Framework for Unsupervised Anomaly Detection : Detecting Intrusions in Unlabeled Data." To Appear in Data Mining for security

[10] Sandeep Kumar, Classification and Detection of Computer Intrusions. Ph. D. Dissertation, August 1995.

[11] R. Agrawal, T. Imielinski, and A. Swami. Swami. "Mining association rules between sets of items in large detabase." In proceedings of the ACM SIGMOD Conference on Management of Data, page 207~216, 1993.

[12] 한국정보보호진흥원, 정보통신 기반구조 보호기술개발, 2001.12.

[13] Wenke Lee and Xiang. "Information- theoretic measures for anomaly detection". In Proceedings of the 2001 IEEE Symposium on Security and privacy, May 2001.

[14] Luca Dert, Stefano, Gata Masellt,"Design and Implemention of an Anomaly Detection System : an Empirical Approach". submitted to NOMS, 2002.

[15] Leonid Portnoy, Eleazar Eskin and Salvatore J. Stolfo. "Intrusion detection with unlabeled data using clustering" To Appear in Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001). Philadelphia, PA : November 5-8, 2001.

[16] Eleazar Eskin, "Detecting Errors within a Corpus using Anomaly Detection." Proceedings of First Conference of the North American Association for Computational Linguistics, 2000. 11~1.

[17] John E. Dickerson, Jukka Juslin, Ourania Koukousoula, Julie A. Dickerson, "Fuzzy Intrusion Detection". 2001.

〈著者紹介〉

박 광 진 (Kwang-jin park)
1982년 2월 : 동국대학교 전자계산학과 졸업
1988년 2월 : 한양대학교 산업대학원 전자계산학 석사
1998년 9월~현재 : 광운대학교 컴퓨터과학과 박사과정
<관심분야> 정보통신정책, 정보보호

유 황 빈 (Ryou Hwang-bin)
1975년 2월 : 인하대학교 전자공학과 졸업
1977년 8월 : 연세대학교 대학원 전자공학 석사
1989년 2월 : 경희대학교 대학원 전자공학 박사
1981년 3월~현재 : 광운대학교 컴퓨터과학과 교수
<관심분야> 멀티미디어, 정보보호, 개인정보정책