2002 년 한국정보처리학회 춘계학술발표논문집 제 9 권 제 1 호

베이지안 네트워크를 이용한 분산 IDS 설계

김도진*,이정현*,황숙희*,황준원*,이창훈*
*건국대학교 컴퓨터공학과
e-mail : lionguy,corona,hwang,jwhwang,chlee@konkuk.ac.kr

A distributed IDS design on global network

Do-Jin Kim*, Jung-Hyun Lee,Suk-hee Hwang, Chang-hun Lee*
*Dept. of Computer Science, Kon-kuk University

요 약

광역 네트워크상에서 침입탐지는 백본망에서의 기가비트를 처리할 수 능력이 시스템에 필요로 하고 있다. 하지만 많은 loss 와 시스템 미치는 부하로 시스템 자체에 큰 영향을 미친다. 따라서 본 논문에서는 이러한 단점을 보완하기 위하여 백본망에 있는 각 local network 에 분산 에이전트를 설치하고, 여기에서 발생한 데이터를 다중회귀분석의 회귀계수를 메인 시스템에서 보내 처리함으로써 전체 및 각 Local 네트워크에 대한 밸런스를 조절하고, 감시하는 기능을 제공하는 시스템의 설계방법을 제시한다.

1. 서론

인터넷 사용자가 급격히 증가함에 따라 정보통신 산업의 발전이 되었지만, 이에 따른 역기능 또한 크게 증가하고 있는 추세이며, 이를 차단하고, 탐지하는 기술이 해킹기술을 앞지르지 못하고 있으며 수동적 입장에서 해킹 사례를 분석하거나, 접근 차체를 차단하는 방법을 택하고 있지만 새로운 해킹 시도에 노출되고, 피해가 계속되고 있다. 따라서 우리가 제안하는 Anomaly IDS 는 능동적 입장에서 해킹 기법에 대해 대처하고, 새로운 형태의 해킹기술을 탐지함으로써, 보호하려는 시스템에 대한 능동적 보안수단을 제공한다. 본 논문에서는 기존에 제시되어 왔던 대형 네트워크상에서의 Anomaly IDS 에서의 문제점을 보완하는 분산 Agent 를 통해서 트래픽을 가공한 데이터를 통계적으로 학습하여 패턴을 생성하고 탐지하는 시스템을 설계하였다.

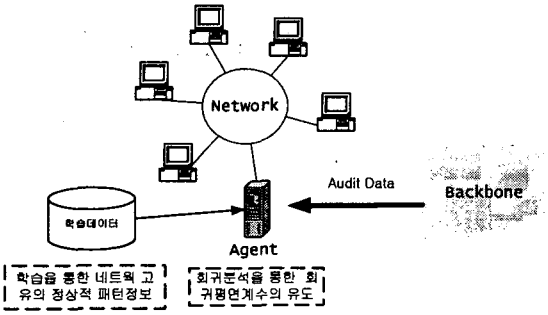
본 논문에서 제시한 것은 Global Network 에서의 비정상행위를 탐지하기 위한 방법으로 한정되며, 광역 네트워크로 이루어진 네트워크 내에 있는 시스템으로 들어가는 비정상적 패킷을 탐지하는 시스템이다. 분산 에이전트가 네트워크에서 발생하는 트래픽을 처리하는 방법을 2 장에서 알아보고, 3 장에서 제안한 베이안 추론식을 이용하여 Agent 가 보낸 자료를 분석하는 방법에 대해서 알아본다. 4 장에서는 제안한 시스템의 설계에 대해서 알아보고, 마지막으로 5 장에서 결론을 맺는다

2. 분산 에이전트

각 Local network 상에는 Agent 를 통해서 네트워크로 들어오는 데이터의 1 차 가공.처리한다. 이러한 가공처리는 네트워크 프로토콜의 Header 정보를 통계처리하기 위한 데이터의 수집을 하는 역할을 수행한다.

기존의 광역 네트워크에서의 침입탐지를 목적으로 하는 경우 Backbone 에 트래픽을 해결할 수 있는 어떤 완충장치 없이 설계한다면 IDS 가 트래픽을 처리하기 위해 많은 리소스를 필요로 하게 되므로 오버헤드의 부담은 큰 문제가 될 것이다. 분산 에이전트는 IDS 가 단독으로 존재할 경우보다 IDS 에 가해지는 오버헤드를 최소한으로 줄이고, 광대역의 네트워크를 감시.보고 할 수 있다는 장점을 가지게 된다.

아래 <그림 1>에서 Agent 의 역할을 보이고 있다.



<그림 1> 분산 Agent 의 구조와 역할

2.1 필드의 생성 및 선택

비정상 행위에 대한 판단을 하기위해서는 네트워크의 패킷들 중에서 판단을 내릴 수 있는 특정 항목을 선택 하여야 한다. 이러한 과정을 특징 추출 또는 선택이라 하며, 제시하고자 하는 시스템 특징 선택 방법은 필요한 packet 에서 추출한 이벤트 필드 값과 여러 확률에 의해서 생성된 이벤트 필드를 통계적 방법에 의해서 처리하는데 그 목적이 있고, Anomaly IDS 에서 통계적 처리방법은 네트워크에 있어서의 데이터 분포를 선택된 데이터를 이용해 표현할 있으며, 처리 방법이 간단하므로 같은 시간에 더 많은 종류의 요소를 비교분석이 가능하다라는 장점이 있다. 그러나 정규분포형태의 데이터를 가정으로 하고, 각각의 필드가 독립적이므로 여러 필드가 복합적으로 고려되어야 하는 패턴의 감지가 불가능하다라는 단점을 가지고 있다.

위와 같은 이유로 각 필드의 선택에 있어 카이제곱 검정을 통해서 정상데이터와 모의 해킹을 통한 데이터 수집, 빈도를 구하고, 이론적인 빈도와 실제빈도의 차이를 이용 독립성 혹은 관련성 있는지 판단, 이론빈도와 실제 빈도간의 차이가 크면 X^2 값이 증가하게 되고, X^2 의 값이 큰 값을 선택하였을 때 필드는 <표 1>과 같다.

필드	접근유의 확률	채택 여부
1.전체 패킷량	0.422	O
2.tcp	0.375	O
3.udp	0.041	O
4.icmp	0.001	O
5. 연결 회수(syn - syn/ack - ack)	0.375	O
6. 비정상 종료 패킷수	0.017	O
7. 서로 다른 서비스의 수/전체 연결 회수	0.002	O
8. 세션 연결 회수(동일 서비스)	0.024	O

<표 1> X^2 검정을 통한 필드 선택

UDP, ICMP, 비정상종료, 연결 수, 서비스 필드는 접근 유의율이 5% 미만 이지만 유사도가 너무 작아 독립적이다.

2.2 다중회귀분석

회귀분석은 둘 또는 그 이상의 변수간의 상관성 또는 관련성을 검토하며, 하나 이상의 독립변수로써 종속변수를 추정하면 추정의 정확도를 높일 수 있는 통계적 절차를 다중 회기 분석이라 한다.

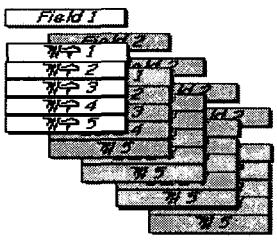
다중회기 분석은 분산 IDS 시스템 중 Agent 에 적용하였다. 이러한 다중회귀분석의 적용은 종속변수를 추정하는데 가장 유용한 정보를 이용할 수 있다는데 있다. 그리고, 하나 이상의 독립변수로써 종속변수를 추정하면 추정의 정확도를 높일 수 있다.

선택된 필드에서 가장 유용한 정보를 가지고 있는 필드들로 다른 필드를 예측하는 통계적 절차라고 할 수 있다.

필드수가 k 개, 즉 X_1, X_2, .. ,X_k 라고 가정하면 추정하고자 하는 회귀식은

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + .. + b_k X_k$$

다음 같은 다중회귀 식에서의 회귀평면계수를 만들 수 있다. 여기에 본문의 관심은 회귀평면계수를 베이지안 네트워크를 추론 서버로 보낸다는 것이다.



<그림 2> 다중 회귀평면계수 Table

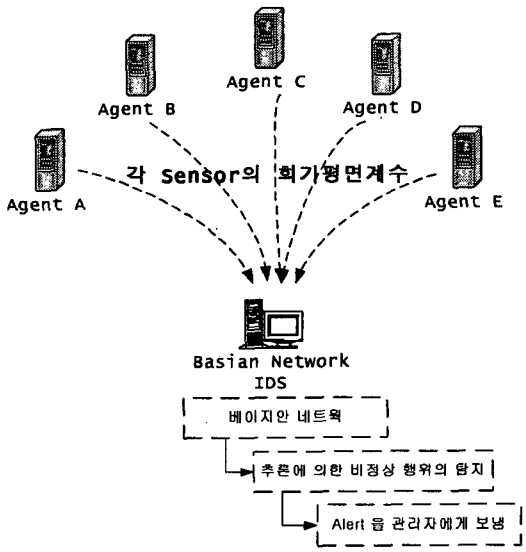
<그림 2>는 5 개의 local 네트워크에서 다중회귀분석을 통해서 얻은 회귀평면계수 이다. 이 회귀평면계수는 베이지안 네트워크를 구성하기 위한 데이터 셋으로 필드의 가중치를 의미한다.

3. 베이지안 네트워크 IDS

본 IDS 는 각각의 네트워크를 담당하는 **Agent** 로부터 1 차 가공된 회기평면계수를 받아 **Agent** 간의 인과관계를 밝히고, 비정상 행위가 보고된 네트웍에 대해서는 **Alert** 을 보내게 된다.

이러한 방법은 네트워크간에는 비정상 행위에 대해 서로 영향을 주게 되는데 특히, 근접한 네트워크간에는 이러한 현상이 두드러진다고 보는 것이다.

베이지안 네트워크 IDS 의 비정상행위 추론을 위한 과정을 <그림 3>이 보이고 있다.



<그림 3> Basian Network IDS 의 추론과정

3.1 베이지안 네트워크 추론

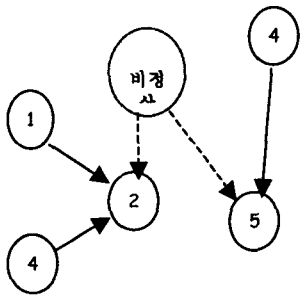
실제 베이지안 네트워크를 구성하기 위해서는 Markov Chain Monte Carlo 방식의 Gibbs Sampling 방법을 이용하면 된다. 하지만 이 알고리즘은 여러 해가 있을 경우 NP-hard 문제가 부딪칠 수 있으며, 부정확할 확률이 높다. 따라서 본 논문에서 여러 과정을 거쳐 좀더 신뢰성 있는 네트워크를 구성, 베이지안 네트워크 추론식의 정확성을 높이는데 노력을 하였다.

통계적인 방법에서 베이지안 통계를 사용하는 이유는 베이지안 통계는 불완전한 데이터 집합을 처리할 수 있으며, 인과 관계에 대한 학습이 가능하고, 도메인 지식과 데이터를 결합하여 사용이 가능하다라는 것인데, 전문가 도메인 지식과 더불어 통계적 데이터 학습을 통해서 새로운 지식을 발견할 수 있다. 또한 해가 너무 많은 NP-hard 문제에 부딪히는 경우가 많은데 베이지안 네트워크를 이용함으로써 데이터의 NP-hard 문제와 overfitting 을 피할 수 있는 효율적이고 원천적인 방법을 제시해 준다.

실제 베이지안 네트워크를 구성하는 요소는 이전에 대부분의 논문이 필드의 값을 이용하여 베이지안을 구성하였지만, 본 논문에서 회귀평면계수를 이용하여 베이지안 네트워크를 구성한 이유는

- 일정 기간의 데이터가 없을 경우 베이지안에서 데이터처리 불능
- 각 필드값에 대한 기울기(가중치)이므로 필드간의 관계표현성을 크게 할 수가 있다.

실험상 n 번 이상의 정상 데이터의 학습을 통해서 n 개의 <그림 2>과 같은 회귀평면계수 테이블을 수집하고, 에이전트 각각에 대한 5 개의 서로 다른 베이지안 네트워크 및 추론식이 만들어 진다.



<그림 4> 계수 1 에 대한 베이지안 네트워크 및 추론식

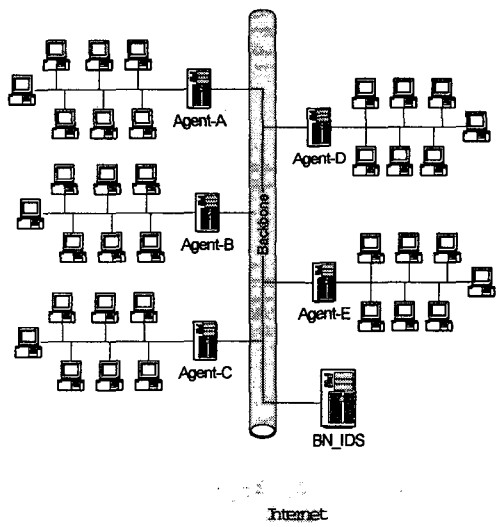
$$P(a|1,2,..4,5)=\frac{p(a)p(f(2)|a,f(1),f(4))p(f(5)|a,f(4))}{\sum_{a} p(d)p(f(2)|d,f(1),f(4))p(f(5)|d,f(4))}$$

아래의 베이지안 네트워크 추론식은 실제 탐지 모드에서 그대로 활용되며 a'인 경험치를 줌으로써 false positive 를 조절할 수 있다.

위의 그림상에 있는 수식은 베이지안 네트워크 추론식이며 f 함수의 의미는 n 개의 회귀평면계수가 정규분포를 갖는다고 가정했을 때 갖는 확률값을 의미한다.

4. 분산 IDS 설계

본 논문에서 제안하는 시스템은 아래 <그림 5>의 시스템 구성도와 같이 보이고자 한다.

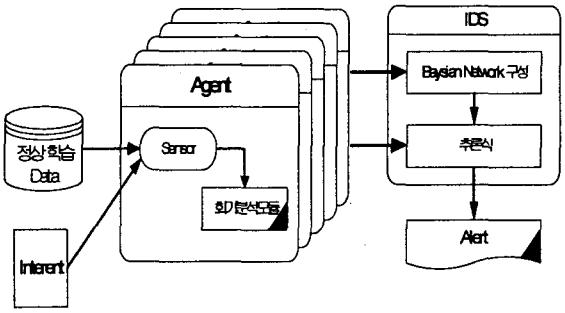


<그림 5> 분산 IDS 시스템의 구성

학습단계에서 각각의 네트워크에 위치하고 있는

Agent 는 정상행위에 대한 학습을 통한 Data 를 가지고 있으며, Sensor 부분은 Audit data(원시 패킷)를 다중회귀분석을 하기 위한 이벤트 테이블을 만들고 다중회귀를 통해서 얻은 회귀평면계수를 Baysian Network IDS 서버로 보낸다. 이 과정을 반복하여 베이지안 네트워크를 만들기 위한 테이블을 만들고, 추론식을 만들어 침입판단 모듈로 추론식을 인코딩시킨다.

탐지모드는 Agent 가 이벤트를 침입판단 모듈로 필드의 확률 값을 보내면 Baysian Network 모듈에서는 각 Agent 간의 인과관계에 대한 학습 및 추론을 수행하게 되며, 비정상에 대한 사후추론치를 알려주고 Alert 모듈을 통해서 경고한다.



<그림 6> 분산 IDS 설계

5. 결론

본 논문에서 제시하는 IDS 시스템은 광역 네트워크에서 분산 Agent 시스템을 가진 IDS 가 네트워크의 밸런스를 에이전트를 통해서 모니터링하고, 각 네트워크의 비정상여부를 탐지한다. 이러한 시스템의 구성은 Network Based 또는 Host Based 의 적용에 불과했던 IDS 의 적용을 트래픽 처리에 따른 오버헤드를 최소화하고 비정상행위의 탐지를 위해 효율적으로 수행하도록 한다.

그러나 정상적인 네트워크 상황에서 Pure 데이터라는 가정 하에 베이지안 네트워크를 학습하였다. 실제 이런 정상적인 데이터를 모으는 과정과 알고리즘이 생략되었다. 이런 Pure 데이터를 수집하고, 비정상적인 데이터가 있을 경우 노이즈를 제거 할 수 있는 알고리즘에 대한 연구가 필요하며, 각 local network 의 모든 필드를 동일하게 선택을 했는데, 이는 네트워크의 특징을 정확히 표현하기 어렵다. 따라서 각 네트워크 마다 서로 다른 필드를 선택함으로써 좀더 각 네트워크에 대한 특징을 효과적으로 추출할 수 있을 것이다.

에이전트의 회귀분석 모듈에서도 반복적 실험을 통한 회기평면계수의 조정으로 더 정확한 분석 시스템의 구성과 베이지안 네트워크의 신뢰성을 높이고, 네트워크간의 인과관계에 대한 연관성 추론확률의 정확도를 높일 수 있을 것이다.

참고문헌

[1] Dorothy E. Denning, "An Intrusion Detection Model", In IEEE Transaction on software engineering, Number 2, February 1987.

[2] Mark Crosbie, Gene Spafford, " Defending a Computer System using Autonomous Agents", Technical Report CSD-TR-95-022, Purdue University, March 11 1994.

[3] Joseph Barrus, "A Distributed Autonomous-Agent Network-Intrusion Detection and Reponse", Proc, 1998 Command and Control Research and Technology Symposium, Monterey CA, June-July 1998.

[4] Jean-Philippe Pouzeol , Mireille Ducasse " Handling Generic Intrusion Signatures" ,2000

[5] Steven T. Eckmann , Giovanni Vigna , Richard A. Kemmerer " STATL: An Attack Language for State-based Intrusion Detection " ,1999

[6] Giovanni vigna , Richard A,Kemmerer " NetSTAT: A Network-based Intrusion Detection Approach" ,1999

[7] A Data Mining Framework for Building Intrusion Detection Models,199x

[8] M. Chung, N. Puketza, R. A. Olsson, and B. Mukherjee, "Simulating concurrent Intrusion for Testing Intrusion Detection Systems: Parallelizing Intrusion", Proc, 18th National Information Systems Security Conference, Baltimore, MD, pp.173~183, October 1995.

[9] D. Anderson et al, "Next Generation Intrusion Detection Expert System (NIDES)", Software Design, Product Specification, and Version Description Document, Project 3131, July 11 1994.