제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

침입탐지 시스템에서 Alert 의 패턴 학습을 이용한 False Positive 감소에 대한 연구

심철준*, 곽주현*, 원일용*, 이창훈*
*건국대학교 컴퓨터공학과
e-mail : spsj1004@cse.konkuk.ac.kr

Research on False Positive Alert reduction using pattern matching technique

Chul-Jun Sim*, Ju-Hyun Kwak*, Il-Yong Won*, Chang-Hun Lee*
*Dept. of Computer Engineering, KonKuk University

요 약

False Positive Alert 은 IDS 가 공격이 아닌 것을 공격으로 잘못 판단하는 것이다. 이러한 False Positive 는 시스템에 직접적인 피해를 주지는 않으나, 시스템 관리자가 적절한 대처를 하기 어렵게 하므로 IDS 의 새로운 문제점으로 대두되고 있다. 본 논문에서는 이러한 False Positive 를 줄이기 위해 IDS 에서 나오는 Alert 중 False Positive 를 필터링 하는 방법에 대해 제시한다. 공격에 대한 Alert 과 False Positive Alert 의 시간 패턴을 각각 분석, 학습 함으로써 그 후의 Alert 의 False Positive 여부를 판별한다.

1. 서론

NIDS 는 네트워크 상에 지나가는 패킷을 분석 함으로서 실시간에 침입을 탐지 하기 위한 시스템이다.

NIDS 는 패킷을 가지고 Rule 과 비교를 해서 공격을 탐지하는 것이다. 그러므로 NIDS 에서는 Rule 이 중요하다. 만일 Rule 이 정확하지 않으면 침입이 아닌 패킷들을 공격으로 오인 할 수가 있다.

과거에 비해 스위칭 기술의 발달과 Bandwidth 의 향상 등 네트워크 기술의 향상이 두드러지면서 네트워크 속도가 증가하고 그에 비례해서 네트워크 상의 트래픽이 증가 함으로서 이벤트의 양이 많아 지므로 False Positive 의 양도 많아 지고 있다.

관리자는 NIDS 에서 나오는 모든 Alert 을 가지고 공격에 대한 대응을 한다. 이러한 경우 정상적인 행위를 공격으로 오인 해 정상적인 사용자에게 피해를 줄 수 있다. 또한 관리자가 수동으로 Alert 을 처리한다고 해도 실수를 범할 수 있고, 많은 양의 Alert 을 관리자 혼자 처리하기에는 역부족이다.

그러므로 False Positive 는 IDS 의 실용성을 낮추는 치명적인 문제로 대두되고 있다. 이러한 False Positive 를 낮추기 위한 연구들은 현재 초기단계로 여러 방법론들이 적용되고 있다.

본 연구에서는 각 공격 타입 별로 시간 패턴을 통계화 시켜 이를 학습알고리즘 중 하나인 IBL 을 적용시켜 실제 공격에 대한 Alert 와 False Positive Alert 를 구분할 수 있는 방법을 제시한다.

2. 관련연구

2.1 기존의 False Positive 해결 방법

[1] 네트워크 환경에 따른 NIDS 의 환경 설정

시스템 관리자가 네트워크와 시스템의 상황에 맞게 침입탐지 시스템의 환경을 설정함으로써 False Positive 을 줄이는 것이다.

NIDS 의 환경 또는 침입에 대한 Rule 을 default 로의 설정은 네트워크 환경에 맞지 않다. 일반 관리자가 침입탐지 시스템의 환경 또는 Rule 을 default 로 설정하면 정상적인 패킷을 해킹으로 오인 할 수가 있다. 그러므로 관리자는 네트워크 상황에 맞게 몇 번이고

시행의 오류를 겪으면서 침입탐지 시스템의 환경을 맞추어야 한다. 이러한 과정은 회사의 보호 정책에 고려하여 수행되어야 하기 때문에 관리자 능력에 성능 의존도가 크다.

[2] DB 사용 설정

각 보호대상 시스템의 취약점을 DB 화 시켜 둔 것을 자산 DB 라고 한다.

IDS 에서 나온 Alert 가 발생을 하면 그 Alert 는 자산 DB 로 이동 하여 취약성을 검색 한 다음 동일한 취약점들이 있으면 그 Alert 을 출력 한다. 그러나 취약점을 검색 했을 때 취약점에 그 정보가 없으면 공격임에도 불구하고 공격의 Alert 가 발생을 하지 안는다.

자산 DB 는 False Positive 만이 아니라 전반적인 Alert 의 수를 감소시키는 방법으로 실제 관리에 있어서는 꼭 필요한 방법이다.

그러나 그 시스템이 취약한 부분이 아닐 경우 실제 공격의도가 있었다고 해도 이를 무시하게 된다는 단점이 있다.

위의 연구들은 환경설정을 통해 줄이는 방법이 대부분이기 때문에 네트워크 상황이 자주 바뀌면 관리자가 관리하기 어렵다는 문제점을 가지고 있다. 특히 자산 DB 를 사용할 경우 각각 시스템의 취약성들에 대한 지속적인 update 가 부담이 되며 또한 그 시스템에 적용이 되지 않는 공격이라 할지라도 공격의 의도가 있었기 때문에 그 공격에 대한 Alert 들을 무시하는 것은 논란의 요소가 있다. 아무리 공격이 자산 DB 에 속해 있지 않는 공격을 한 것이라도, 나중에는 취약성이 있는 공격을 유발 할 수 있기 때문에 미리 예방을 해야 한다.

2.2 IBL

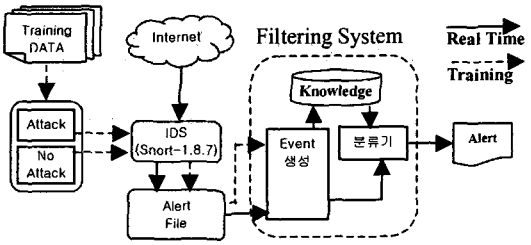
IBL 은 C4.5 나 CN2 와 같은 학습 알고리즘의 하나이다. 기존의 결정트리 (C4.5) 나 Rule 생성(CN2) 방식들이 동일한 클래스의 Data 들의 공통점을 찾아 이를 트리나 Rule 의 형태로 분류하는 방식인데 비해 IBL 은 Data 각각의 예를 모두 보존하거나 (IBL1) 또는 유사한 Data 들의 대표 Data 를 보존하여 이것과 Test Data 를 비교하는 방식이다.

IBL 은 복잡한 Rule 이나 법칙이 특정 클래스의 분류 기준이 되는 문제에 대해서는 약할 수도 있으나 하나의 클래스가 여러 가지 형태로 분류되는 형태의 Data 에 대해서는 효율적인 모습을 보인다.

3. 연구내용

3.1 시스템의 구성

본 논문에서 제안하고자 하는 침입탐지 시스템에서의 False Positive 는 아래 [그림 1]과 같은 구조로 구성되어 있다.



[그림 1] Filtering 시스템의 전체 구조

이 시스템에서의 전체적인 구조에 설명은 우선 Training Data 의 Normal 과 Abnormal 에 대해 분류를 한 다음 IDS 에 돌려 Alert 를 발생 한다. 그 다음 발생한 Alert 를 Event 생성기에 의해 학습을 시켜 Knowledge 을 만든 다음 실시간 네트워크에서 들어오는 Alert 를 Event 화 하여 Knowledge 와 실시간 네트워크에서 들어온 Alert 의 Event 를 가지고 서로 비교 하면서 Alert 를 발생하는 것이다.

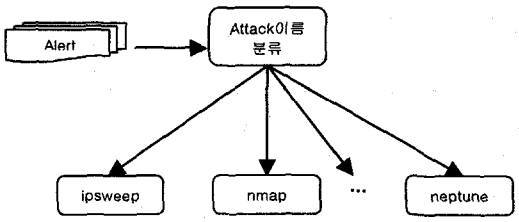
3.2 Event 생성

3.2.1 Alert 그룹 생성

DOS 를 포함한 많은 종류의 Attack 은 공격 시 IDS 에서 대량의 Alert 를 발생시킨다. 이러한 지속적인 Alert 의 발생은 공격의 지속 시간 등에 영향을 받기 때문에 이 공격의 심각성을 표현하는 지표가 될 수도 있으나 관리자의 상황파악을 더 어렵게 할 수도 있다. 그러므로 본 시스템에서는 이러한 Alert 들을 다음과 같은 기준으로 그룹화 시킨다.

- 공격유형이 동일한 Alert
- 일정시간 내에 지속되는 source ip

본 실험에 분류 기준시간은 15 분으로서 이 시간 내에 같은 ip 로부터 발생한 동일한 Alert 는 같은 계열의 공격행동으로 간주하여 그룹화 시켰다.
이러한 그룹기준을 통해 실험에 사용된 DARPA Data (4.1) 에 발생한 모든 공격 행위 별로 그룹화 시킬 수 있었다.



[그림 2] Alert 그룹 생성 과정

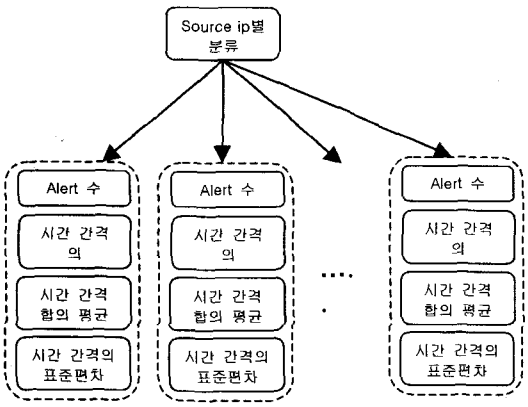
3.2.2 통계 처리

각각의 공격행위를 상징하는 Alert 그룹에 해당되는

Alert 들은 각각 Alert 가 발생한 시간, source ip, destination ip 등을 포함하고 있다. 이러한 Alert 의 각각의 그룹들을 학습 시키기 위해 다음과 같은 형태로 각 그룹의 특징을 추출하였다.

attr1 : Alert 유형(메시지의 타입)
attr2 : Alert 의 수
attr3 : Alert 의 지속시간
attr4 : Alert 들 사이의 평균 시간간격
attr5 : attr4 의 편차

실제로 동일 공격 툴이나 script 등을 사용한 공격에 대해서 발생하는 Alert 들의 시간간격이나 숫자 등 매우 많은 유사성을 지니는 것으로 발견되었다.



[그림 3] 통계처리 과정

3.3 Filtering system

Filtering 시스템은 학습 Data 를 Abnormal 과 Normal 로 분류하고 (3.2.1)와 같이 Alert 그룹들로 분류한 후 Alert 그룹들마다 통계적인 특징들을 추출한다. (3.2.2) 이러한 Alert 그룹들은 attr1 ~ attr5 까지의 attribute 를 갖는 instance 들로 입력되어 IBL 에 Normal 과 Abnormal 의 특징이 학습된다.

IBL 은 이러한 instance 들을 전체(IBL1) 또는 일부의 대표 instance (IBL4)를 저장해 둔 후 새로운 Alert 가 들어온 경우 기존의 instance 와 비교하여 가장 유사한 instance 의 Normal / Abnormal 을 구해 새로운 Alert 의 Abnormal 여부를 결정한다.
이때 Abnormal 로 판정된 Data 들에 대해서만 사용자에게 Alert 을 발생시키게 되며 그렇지 않은 Data 는 False Positive 로 발생시키고 제거한다.

4. 실험 방법 및 결과

IDS 에 관련된 네트워크상의 실험 시에 실제 Attack 여부를 정확하게 검증할 방법이 어렵다. 그래서 이러한 문제를 해결하기 위해 비교적 객관성을 인정 받은 DARPA Data 의 Tcp dump 파일을 사용하여 off-line 상에서 실험을 수행하였다.
이러한 Tcp dump 파일은 DARPA 에서 제공하는 Attack list 를 통해서 Attack 패킷과 Normal 패킷으로 분류되어 역시 널리 사용되는 공개 IDS 인 snort 를 통해 Alert 를 발생 시킴으로써 실제 공격에 대한 Alert 와 False Positive Alert 들을 추출하였다.
그리고 이를 통해 발생한 False Positive Alert 와 Attack 에 대한 Alert 중 약 80%를 Training Data 로 나머지 20%를 Test Data 으로 분류한 후에 IBL 을 통해 False negative Alert 의 감소율 과 False Positive Alert 감소율 을 구하였다.

4.1 DARPA Data

DARPA Data 는 미 국방성에서 인터넷 보안 연구를 위해서 인위적으로 실험용 네트워크를 구성하여 일정 기간 동안 발생하는 모든 공격 및 비 공격 패킷 Data 를 DUMP 파일로 저장해 둔 것이다.

이 논문에서 사용한 DARPA Data 는 1998 년도에 수집된 3 주차의 월요일, 목요일, 금요일, 4 주차의 월요일, 목요일, 금요일, 5 주차의 월요일, 화요일, 수요일, 목요일, 금요일, 6 주차의 월요일, 화요일, 수요일, 목요일, 7 주차의 월요일, 화요일, 금요일의 Data 를 사용했고, 그 다음 3,4,5,6,7 주차 ipsweep, neptune 공격을 사용하고, 3,4,5 주차는 Training Data 로 사용을 하고, 6,7 주차는 Test Data 로 사용 했다.

4.2 실험 시스템 및 사용 IDS

4.2.1 실험 시스템

Linux 8.0 을 사용 하고, 그리고 컴파일을 하기 위해서는 egcs 를 사용 하였다.

4.2.2 Snort

Snort 는 공개된 IDS 중 가장 널리 사용되는 IDS 의 하나로서 국내 수많은 상용 IDS 가 이를 기반으로 구현되었으며 현재에도 지속적인 update 가 이루어지는 IDS 이다.

본 실험에는 버전 1.8.7 의 linux 기반 snort 가 사용되었다.

4.3 Data 중 사용된 공격의 유형과 Alert 의 분포

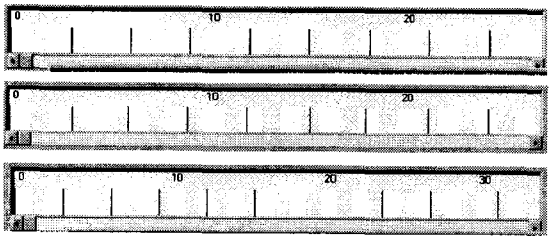
4.3.1 공격의 유형

DARPA Data 의 3,4,5,6,7 주차를 실험해 본 결과 False Positive 는 자주 발생 했지만 실험 할 수 있을 정도의 Attack 는 ipsweep 과 neptune 정도였다.

ipsweep, neptune 공격들은 예상했던 것 이상으로 매우 일정한 Alert 발생 시간패턴을 지녔으며 이는 특정

script 나 tool 을 사용하기 때문에 발생하는 현상으로 예상된다.

아래의 그림은 3 개의 ipsweep 공격으로 초당 들어온 Alert 을 시간화 시킨 것으로 서로 다른 곳에서 온 공격임에도 매우 유사함을 알 수 있다.



[그림 3] ipsweep Alert 의 시간 별 발생 패턴

4.3.2 Alert 의 분포

DARPA Data 를 Snort 로 돌려 나온 ipsweep 과 neptune Attack 의 각각 Alert 들을 Alert 의 시간 간격의 합, 시간 간격합의 평균, 시간 간격의 표준 편차를 계산한 결과는 False Positive 로 나온 ipsweep, neptune Attack 에 대한 Alert 시간 간격 합 과 표준 편차가 다르게 나온다.

[표 1] Ipsweep

Alert 수	시간 간격의 합	시간 간격 합의 평균	시간 간격의 표준편차	Attack 이름	Normal/ Abnormal
133	79201.72	595.5016	4.477456	ipscna	Normal
122	72601.53	600.0126	0.000725	Ipscna	Normal
130	77998.69	604.641	9.230772	Ipscna	Normal
512	27.3359	0.053495	0.20993	ipscna	Abnormal
510	27.00782	0.053061	0.104027	ipscna	Abnormal
510	27.00463	0.053054	0.104013	ipscna	Abnormal
762	3830.329	5.033283	0.027948	ipscna	Abnormal
507	33.24088	0.065693	0.128591	ipscna	Abnormal

[표 2] neptune

Alert 수	시간 간격의 합	시간 간격 합의 평균	시간 간격의 표준편차	Attack 이름	Normal/ Abnormal
10	2633.364	292.596	17.09977	neptune	Normal
200	2676.046	13.44747	25.63748	neptune	Abnormal
241	2691.229	11.21345	21.54569	neptune	Abnormal
200	2727.154	13.70429	26.1307	neptune	Abnormal
200	2690.515	13.52018	25.77922	neptune	Abnormal

4.4 실험 결과

본 실험에서 사용된 ipsweep 과 neptune 의 경우 거의 일정한 Alert 패턴을 발생시킴을 알 수 있었다. 이것은 공격 시 사용하는 tool 이 일정한 시간대별 패킷 발생을 가지기 때문에 발생하는 현상으로 추정된다.

특이한 현상으로는 DARPA Data 에 포함된 ipsweep 의 False Positive 역시 10 분 간격으로 매우 일정한 패턴을 지녔다는 점이었다. 이는 이 Data 에서 발생한 False Positive 가 특정한 네트워크의 기능요소에 의해 주기적으로 발생하게 되는 것이라고 생각된다. 이는 실제 Alert 쪽만을 학습시키는 것이 아니라 False Positive 도 유사한 패턴이 발생할 수 있음으로써 양쪽을 모두 학습시켜야 한다는 근거가 되었다.

[표 3] 실험 결과

	Portsweep	neptune
False Positive 감소	100%	100%
False Negative 증가	0%	0%

위의 성능의 평가는 DARPA Data 가 인위적인 환경으로 구성된 실험용 네트워크 이며 이에 따라 사용된 공격 툴도 일정했으므로 좋은 결과를 얻을 수 있었다.

그러나 실제 네트워크 환경에 적용시키기 위해서는 더 많은 Data 를 수집하여 다양한 형태의 공격유형에 대한 실험이 필요하다고 생각된다.

5. 결론 및 향후 과제

본 연구에서는 각 공격에 의해 발생하는 Alert 의 시간적인 발생 패턴을 분석하고 학습함으로써 False Positive Alert 을 filtering 할 수 있는 방법을 제시하였다.

이를 위해서 DARPA Data 들 중 False Positive 와 Attack 이 가장 많이 등장했던 neptune 이나 ipsweep 을 기반으로 각 공격유형별 Alert 을 분석 함으로서 위의 공격에 대해서는 False Positive 를 100% 감소시킬 수 있었다.

그러나 실제 네트워크 상에서의 다양한 종류와 유형, 도구를 이용한 공격을 고려할 때 동일한 성능을 기대하기는 매우 어렵다. 또한 확보된 학습 Data 의 무결성에 대한 보장이 매우 어렵기 때문에 이런 점도 IDS 관련 연구의 어려움으로 남아있다.

이러한 학습기능을 갖춘 IDS 를 개발하기 위해서는 실시간으로 발생하는 Alert 을 분석하여 이를 판별함과 동시에 다시 학습에 사용할 수 있는 기술을 비롯하여 많은 문제점을 극복해야만 한다.

그럼에도 불구하고 실제 공격 중에 상당수를 차지하는 툴이나 script 에 의한 공격의 Alert 발생 패턴이 매우 일정함을 알아낸 것은 이러한 방법론이 앞으로 다양하게 발전할 수 있음을 제시한다.

참고문헌

[1] Kyu-Eon Lee "네트워크기반 침입탐지시스템의 경고 메시지 축약 모델" 포항공대 HPC Lab

[2] 이은영, "네트워크 기반 침입탐지시스템의 위험도 평가 모델" 한국과학기술원 2003

[3] D.Aha, K.Kibler, Noise-tolerant instance-based learning algorithms. Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (pp.794-799).(1989)

[4] Snort homepage www.snort.org

[5] The Truth about False Positive, Internet Security Systems, White Technical Report, 2001

[6] CVE Vulnerability Search Engine. http://cve.mitre.org