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

인스턴스 기반의 학습을 이용한 비정상 행위 탐지

홍성길*, 원일용*, 송두헌**, 이창훈*
*건국대학교 컴퓨터공학과,
**용인송담대학 컴퓨터소프트웨어학과
e-mail : k1tank@nownuri.net
lionguy, clcc, chlee@konkuk.ac.kr
dsong@ysc.ac.kr

Abnormaly Intrusion Detection Using Instance Based Learning

Seong-Kil Hong*, Do-Jin Kim*, Il-Yong Won*, Doo-Heon Song**, Chang-Hun Lee*
*Dept. of Computer Science, Kon-Kuk University
**Dept of Computer Software, Yong-In Songdam College

요 약

비정상 행위의 탐지를 위한 침입탐지 시스템의 성능을 좌우하는 가장 큰 요인들은 패킷의 손실없는 수집과 해당 도메인에 알맞은 분류 기법이라 할 수 있다. 본 논문에서는 기존의 탐지엔진에 적용된 알고리즘의 부류에서 벗어나 Instance 기반의 알고리즘인 IBL(Instance Based Learning)을 선택하여 학습시간의 단축과 패턴생성에 따른 분류근거의 명확성을 고려였다. 또한, 기존 IBL 에 포함되어 있는 Symbolic value 의 거리계산 방식에서 네트워크의 로우 데이터인 패킷을 처리하는데 따르는 문제를 해결하기 위해 VDM(Value Difference Matrix)을 사용함으로써 탐지률을 향상시킬 수 있었다.

Symbolic value 간의 거리계산에 따른 성능향상의 정도를 알아보기 위해 VDM 적용 유무에 따른 실험결과와 탐지엔진에 적용되었던 알고리즘들인 COWEB 과 C4.5 를 이용한 결과를 비교분석 하였다.

1. 서론

네트워크 패킷을 통해 들어오는 크래커의 침입 여부를 탐지하기 위해서는 감사 자료의 구성을 위한 선택과 어떠한 방법으로 패킷을 분석하여 침입의 여부 또는 공격유형의 형태를 인식 할 수 있는 지는 침입탐지 시스템을 개발하는데 있어 매우 중요하다.

다양한 형태의 많은 공격 감사자료를 기록하거나 알려지지 않은 공격을 탐지하기란 매우 어렵다. 따라서 아직까지 완벽한 보안 시스템은 존재하지 않는다.

본 논문의 침입탐지 시스템(IDS: Intrusion Detectioin System)은 네트워크에서 비정상 행위를 탐지하기 위한 탐지 모델을 생성하기 위한 알고리즘들이 그 특성에 따라 탐지 모델이 다르게 나타나며, 특히 네트워크 환경에 종속되는 단점을 보완하고자 네트워크 기반 침입탐지 시스템을 구현하는 데 있어 기존의 시스템에서 적용되어 왔던 성향에서 벗어나 인스턴스를 기반으로 하는 알고리즘인 IBL 을 적용하였으며, 기존 알고리즘에 VDM(Value Difference Matrix)을 적용함으로써 성능을 향상 시켰다.

위에서 서술한 침입탐지 시스템은 실시간으로 네트워크상의 패킷들을 원시 데이터로 사용하며, 원시데이터를 IBL 에 적용하기 위한 이벤트의 형태로 변환하게 된다. 원시 데이터로부터 가공된 Event 는 IBL(Instance Based Learning)의 학습알고리즘에 의해서 미리 생성된 패턴을 바탕으로 비정상 행위를 판단한다.

네트워크의 환경에 따른 탐지 시스템의 신뢰도를

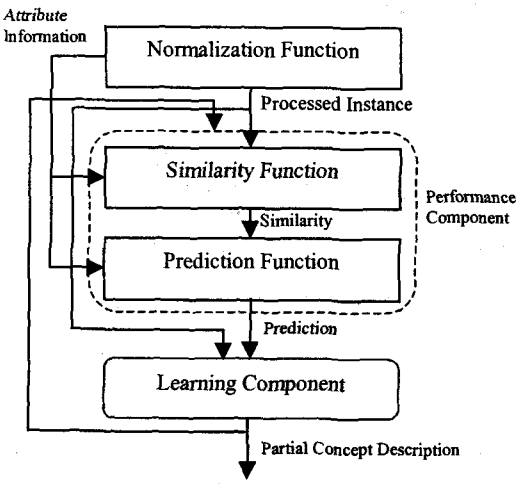
고려하기 위하여 Cobweb 과 C4.5 로 구성된 평가 알고리즘들로 탐지 알고리즘의 탐지율을 비교하였다

2. 관련연구

2.1 IBL(Instance Based Learning)

제안된 침입 탐지 시스템에서 탐지모델의 생성기법은 하나의 사건을 설명할 수 있는 Instance 를 기반으로 하며, Instance 와 Instance 사이의 거리를 나타내는 유사도를 측정하여 패턴을 분류한다. 각 패턴은 PCD(Partial Concept Description)로 구성되며 하나의 PCD 는 패턴을 이루는 Instance 들 중 가장 높은 유사도와 해당 Instance 의 속성값들로 구성되며, 학습과정에서 생성된 PCD 는 하나의 Instance 를 처리할 때 마다 계속 업데이트함으로써 분리 패턴을 생성하게 된다.

IBL 의 학습과정은 [그림 1]과 같이 3 단계로 분리되며 각 단계는 Instance 를 받아들여 일반화하여 학습을 위한 일반화된 Instance 를 생성하는 Processed Instance 를 생성하는 Pre-Process, 유사도 측정과 분류될 카테고리를 예측하기 위한 Prediction 을 생성하는 Performance Component, 다시 Processed Instance 를 받아들여 개념정보인 PCD 를 생성하고 이를 업데이트하는 Learning Component 로 구성된다.



[그림 1] IBL 의 패턴생성 과정

모든 Instance 의 속성들은 Numeric value 또는 Symbolic value 로 채워진 속성값들로 구성된다. 이러한 Instance 의 속성값들 중 가장 높은 값과 가장 낮은 값을 탐색한 후 선택된 속성들을 제외한 모든 Instance 들은 linear normalizing 과정을 거치게 된다.

$$Normalize\ \_attribute\ (x_z, a) = \frac{x_a - a_{\min}}{a_{\max} - a_{\min}}$$

위의 과정으로 일반화된 속성값들은 Performance Component 에 의해 Instance 사이의 관계를 설명하기 위한 유사도와 입력된 Instance 가 특정 카테고리에 속할 수 있는 예측치를 결정한다. 아래의 두 함수는 Similarity function 에 속하며 속성간의 거리를 측정하고 이것을 이용하여 속성사이의 유사도를 도출한다.

$$Similarity\ (x, y) = \frac{1}{\sqrt{\sum_{i \in p} Attribute\ \_defference\ (x_i, y_i)}}$$

$$Attribute\_difference(x_i, y_i) = (x_i - y_i)^2 \quad i\ is\ numeric\text{-}valued$$
$$= x_i \neq y_i \quad VDM$$

Similarity Function 에 의해서 계산된 유사도를 바탕으로 Instance 들을 묶고, 이렇게 묶인 k 개의 가장 유사한 Instance 들 중 가장 유사도가 높은 Instance 의 속성값들을 Learning Component 에게 넘겨 주는데 이때 이것을 Prediction 이라 한다. 이렇게 생성된 Prediction 을 기반으로 분류된 학습데이터에 대한 정보를 업데이트한다.

2.2 평가(Evaluation) 알고리즘

탐지 알고리즘을 구성하기 위해 계층형 개념 군집화(Hierarchical Conceptual Clustering)학습 알고리즘으로 알려져 있는 COBWEB 과 결정 트리 생성기법인 ID3 의 단점을 보완하여 개발된 C4.5 를 적용하였다.

2.2.1 COBWEB

COBWEB 은 인간이 사물을 분류하는 과정인 점진적 개념 형성(Incremental Concept Formation)을 모델로 하여 개발되었다. 이것은 사물을 하나씩 관찰하여 개념을 형성하면서 하향 분류하는 방법이다[5]. COBWEB 은 레코드로 구성된 데이터들을 입력으로 받아, 트리의 형태로 클러스터링을 한다. 트리의 각 노드는 하나의 개념이 되고, 각 노드에는 속성값이 요약되어진 개념정보를 저장하고 있다. 개념 정보는 속성값의 도메인별 확률값이나 평균과 표준편차이다. 속성이 명목형(Nominal) 도메인인 경우에는 확률값이고, 연속형인 경우에는 평균과 표준편차가 된다. 개념정보는 새로운 레코드의 부류를 결정하는데 사용되는 정보가 된다.

2.2.1 C4.5

C4.5 는 ID3 의 확장 알고리즘이다. ID3 가 가지고 있던 문제점들의 해결 방법으로 Local minimum 을 해결하기 위해 Working Set 과 Training Set 을 분리했으며, Pruning 기법을 이용하여 Leaf 의 Class 가 확률로 표시되고, 속성값이 불완전한 경우에도 판단이 가능하다. 수치데이터의 경우 Thresh hold 에 가까운 데이터의 문제점을 보완하였다. 결정 트리 생성과정은 속성값에 따른 경우의 수를 계산하여 트리를 생성한 후 각 Case 에 따른 무질서도를 계산하여 무질서도가 낮을수록 우수한 카테고리로 분류하게 된다. 생성된 트리는 복잡도를 줄이기 위해 Pruning 기법을 사용하며, Noise 와 정보의 손실 가능성을 고려하여 Leaf 가 분류될 Class 를 확률로 표현한다.

2.4 VDM(Value Difference Matrix)

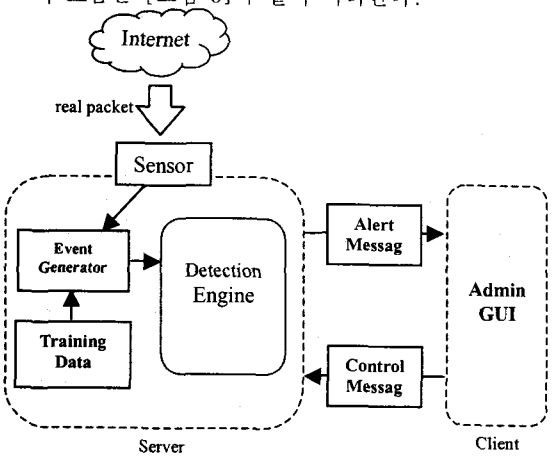
두 인스턴스 사이의 거리를 계산하기 위한 방법으로 VDM 을 사용하였다. 두 인스턴스 사이의 거리를 목적 속성값으로 계산하는 것이 아니라 이웃하는 속성들 사이의 통계값으로 목적 속성간의 유사도를 구한다. 이러한 방법은 이웃하는 속성들 사이의 관계가 목적 속성 사이의 관계에 영향을 줄 수 있다는 생각을 반영한 거리 계산 방법이다.

$$\delta(v1,v2)=\sum_{i=1}^{n}\left|\frac{c_{1_i}}{c_1}-\frac{c_{2_i}}{c_2}\right|^k$$

3. 침입 탐지 시스템의 설계 및 구현

3.1 침입 탐지 시스템의 구조

본 논문에서 제안하고자 하는 침입탐지 시스템은 클라이언트, 서버로 분리되어 특정 네트워크로 유입되는 패킷을 분석하여 탐지 모델에 의한 비정상 행위를 인식한다. 시스템의 전체 구성도와 데이터의 흐름은 [그림 3]과 같이 나타난다.



[그림 3] 비정상 행위 탐지 시스템의 구조

전단부에 위치한 Sensor 는 네트워크로부터 유입되는 모든 패킷을 추출한다. Sensor 는 원시 자료인 Event Generator 에게 전달하게 되는데 이때 Event Generator 는 각 알고리즘들이 처리할 수 있는 의미있는 데이터 형태로 변환해 주는 역할을 하는데, 이때 생성되는 것이 Event 이다.

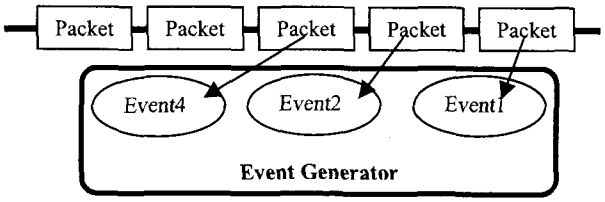
서버는 학습 모드를 통해 원시 데이터인 네트워크 패킷을 추출하여 가공한 Off-line Dataset 을 전달한다. 처리과정을 통해 Instance 를 생성하고, 생성된 Instance 는 IBL 알고리즘에 의한 학습으로 탐지 모델을 구성한다.

침입 탐지 모드에서는 네트워크로부터 들어오는 패킷들을 Event Generator 에 의해 Instance 화 하고 이것을 탐지 엔진에 보내게 된다. 탐지 엔진은 학습 과정에서 생성한 침입 탐지 모델을 근거로 Instance 들을 분류하여 비정상 행위를 구별하며 비정상 행위에 따른 Alert Message 를 소켓통신을 이용하여 클라이언트인 Administrator GUI 에 보내게 되고, 관리자는 그에 따른 적절한 Contorl Message 를 Server 에게 보내게 된다.

3.2 Event 생성

추출된 패킷들은 Event generator 에 의해 이벤트로 변환된다. 이벤트로 변환하는 과정에서 패킷을 처리하는 방법에는 특정 시간단위만큼 패킷을 받아서 하나의 이벤트를 결정하는 것과 하나의 패킷을 하나의 이벤트로 처리하는 이 두가지로 나누어 진다.

본 논문에서는 [그림 4]와 같은 하나의 패킷을 이벤트로 처리하는 방법을 사용하고 있다.



[그림 4] 비정상 행위 탐지 시스템의 구조

비정상 행위 탐지 시스템은 수집된 방대한 양의 감사 기록 정보들로부터 의미있는 정보로의 전환 및 축약시키는 단계가 필요하게 되는데 이와 같은 전처리 기능을 실행하는 것이 Event Generator 이다. 센서는 Event Generator 에게 실시간으로 수집한 감사 기록 정보인 네트워크 패킷들을 보내게 되는데 이때 수집되는 데이터들로부터 비정상 행위를 구분하기 위한 판정 요소를 추출한다. 아래 [표 1]은 감사 기록정보로부터 추출하기 위한 판정 요소들이다.

종 류	속 성
공 통	- 탐지 네트워크상의 IP 패킷수 - TCP, UDP, ICMP 패킷수 및 비율
TCP/IP	- TCP 패킷 비율 : Inbound, Outbound - Connection : SYN, SYN/ACK, ACK 비율 - SYN 을 보낸 출발지 IP 의 수 - FIN, RESET 비율 - TCP 헤더의 플래그 비트 값 - TCP payload 의 길이 총합
UDP,ICMP/IP	- UDP 패킷 비율 : Inbound, Outbound - UDP payload 의 길이 총합 - ICMP 패킷 비율 : Inbound, Outbound

[표 1] 이벤트 데이터를 위한 구성요소

본 논문에서는 네트워크상의 의미적인 트랜잭션의 단위를 정의하는 기준을 하나의 패킷으로 정하여 이벤트를 생성한다. 이를 통해 어떤 시점에서 발생된 네트워크 행위를 보다 정확하게 모델링할 수 있다. 위에서 기술한 내용과 같이 판정 요소로써 속성들을 정의하고 그것에 해당하는 속성값을 할당하면 하나의 패킷에 따른 Event 를 생성하게 된다.

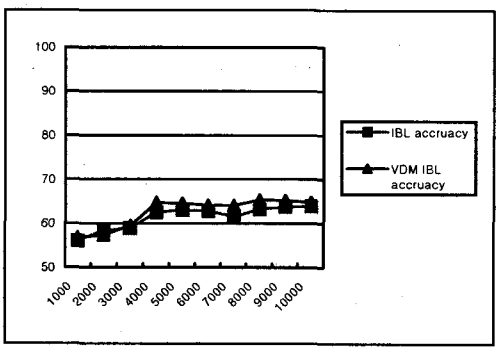
4. 실험환경 및 결과

4.1 실험환경

본 논문에서 제안하는 시스템은 서버와 클라이언트로 구성되어 있다. 서버는 Linux 환경에서 개발되었으며 탐지 네트워크에 직접 배치되어 그 기능을 수행할 수 있다. 클라이언트는 서버와 같은 네트워크 또는 다른 네트워크에 배치되어 소켓통신을 함으로써 침입탐지 결과를 보고 받을 수 있다. 실험을 표준환경의 제공을 위해 DARPA 산하 MIT Lincoln Lab 에서 제공하는 Tcpdump 데이터를 사용하였다. DARPA 데이터는 정상 패킷과 smurf 와 syslog 와 같은 서비스 거부공격으로 이루어진 비정상 패킷들이 포함되어 있다.

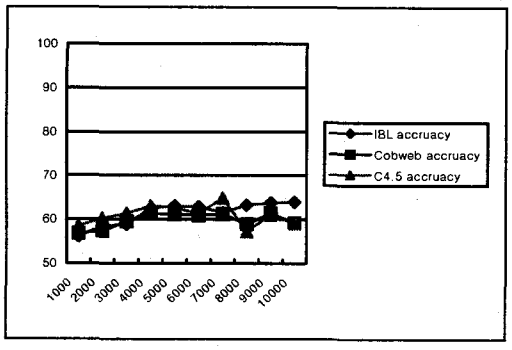
4.2 실험결과

본논문에서 실험한 순수 IBL 과 VDM 을 적용한 IBL 의 실험 결과는 [그림 5]에 비교분석 되어있다.



[그림 5] 기존 IBL 과 VDM 을 적용한 경우의 비교

실험결과로 보아 탐지율이 어느정도 테스트 데이터와 데이터의 양에 영향을 받는다는 것을 알수 있고, Symbolic value 의 유사도 계산문제를 개선하기 위해 VDM 을 적용하였을 경우 탐지율이 더 높게 나온다는 것을 보여주고 있다. 이것은 이벤트를 구성하는 필드의 필드값이 Symbolic value 로 구성되는 것이 많기 때문일 것이라 본다.



[그림 6] 평가모델에 의한 탐지성능 비교

평가모델로 선정한 C4.5 와 Cobweb 의 알고리즘을 이용한 각각의 탐지률 결과는 [그림 6]에서 보는 바와 같으며, IBL 의 성능이 평가 알고리즘보다 우수한 성능을 보이고 있으며, 이벤트의 처리능력에 있어서도 월등한 성능을 보여 주었다.

5. 결론 및 향후과제

본 논문에서는 인스턴스 기반의 학습을 하는 IBL 알고리즘을 적용함으로써 명확한 분류근거의 제시와 패턴을 생성하기 위한 시간과 자원의 낭비가 감소된다는 점이다. 또한 기존 IBL 에 있는 Symbolic value 의 거리계산에 따른 단점을 개선하기 위해 VDM 을 적용함으로써 탐지률이 향상되었으며 Cobweb,C4.5 보다도 탐지률이 우수하다는 것을 알 수 있었다. 그러나 비정상 행위 탐지 시스템은 우수한 성능을 가진 알고리즘을 가졌다 하더라도 시스템이 알고리즘 의존적이기 때문에 알고리즘의 특성에서 벗어나 다양한 도메인에서도 적용이 가능해야 할 것이다.

이러한 문제점을 보완하기 위해서는 오용탐지 방법과 비정상 행위 탐지를 혼합한 침입 탐지 시스템을 고려 해 볼 필요가 있으며, 대역폭이 큰 네트워크에서의 적용을 위한 패킷 수집 및 처리 방법에 대한 연구가 필요할 것이다.

참고문헌

[1] David W. Aha, "A Study of Instance-Based Algorithms for Supervised Learning Tasks", Department of Information and Computer Science University of California, Technical Report, 1990

[2] Cost & Salzberg, "A weighted Nearest Neighbor Algorithm for Learning with Symbolic attribute feature", Journal of Machine Learning, 1993

[3] Kathleen McKusick, Kevin Thompson, "COBWEB/3: A Portable Implementation", Technical Report FIA-90-6-18-2, AI Research Branch, NASA Ames Research Center, 1990

[4] 이정현, "네트워크 기반 비정상 행위에 대한 다계층 침입 탐지 시스템 설계 및 구현", 석사학위논문, 건국대학교 컴퓨터 공학과, 2001

[5] 이효승, "COBWEB 을 사용한 비정상행위도 측정을 지원하는 네트워크 기반 침입탐지 시스템 설계", 석사학위논문, 건국대학교 컴퓨터 공학과, 2002

[6] 김도진, "IBL 을 사용한 네트워크 기반 침입탐지 시스템과 평가 모델의 연구", 석사학위논문, 건국대학교 컴퓨터공학과 2003

[7] 김주영, 강창구, 이극, 이소우 "네트워크 패킷 분석을 통한 침입탐지 기법 개발", 1999

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

[9] J.Frank, "Artificial Intelligence and Intrusion Detection", NCSC, 1994

[10] Murthy and Salzberg, "A System for Induction of Oblique Decision Tree", JAIR 94, 1994