Proceedings of KFIS 2002 Spring Conference, 2002. 5. 25

지능형 침입 탐지 시스템에 관한 연구

On Design of the Intelligent Intrusion Detection System

이민규, 한명묵
경원대학교 전자계산학과 대학원

Min-Kyu Lee, Myung-Mook Han
Dept. of Computer Science
Kyungwon University
E-mail : minkyu75@web.kyungwon.ac.kr

요 약

본 논문에서는 정보보호에서 지능형 침입탐지시스템(Intrusion Detection System :IDS)의 한 모델을 제안한다. 이 모델은 데이터 마이닝 분야와 정보보호 분야의 결합된 방법을 이용한다. 즉, 계산환경을 격상하거나 새로운 공격 방법들 때문에 내장된 IDS를 보완 할 필요가 종종 있다. 현재 사용하고 있는 많은 IDS들은 전문적인 지식을 손으로 작성했기 때문에 IDS들의 변환은 가격이 매우 비싸며, 속도가 느리다는 단점이 있다. 이에 본 모델은 침입탐지 모델을 적응 적으로 구축하는데 데이터 마이닝 구조를 활용한다.

데이터 마이닝(Data Mining : DM)의 기술인 연관 규칙, 순차 패턴, 분류, 군집화, 유전자 알고리즘 기법(GA)인 Selection, Crossover, Mutation, Evaluation, Fitness Function의 기능을 접목하여 단점을 보안하고 처리 성능을 최대로 하는 즉, 보다 안전한 지능형 침입 탐지 시스템(IDS) 모델을 제안한다.

1. 서론

네트워크를 기반으로 한 컴퓨터 시스템이 현대 사회에 있어서 더욱 더 불가결한 역할을 하는 것에 따라, 네트워크 기반 컴퓨터 시스템은 적과 범죄에 의한 침입 목표가 되었다. 그러므로 기밀에 관련되는 정보가 저장된 것과 온라인 조작되고 있는 것에 따라 네트워크 시스템의 안전은 더욱 더 중요하게 되고 있다. 침입탐지시스템은 이와 같이 우리 시스템을 보호하는 것을 돕기 위한 중요한 기술이 되었다.

침입 탐지 기술은 비정상행위 탐지와 오용 탐지로 분류할 수 있다. 비정상행위 탐지 시스템은 정상적인 사용자 프로파일로부터 크게 벗어나는 활동에 주목한다. 정상적인 시스템 사용에 관한 프로파일에서 벗어나는 행위들을 탐지한다. 오용 탐지는 시스템의 알려진 취약점들을 이용한 공격 행위들에 대한 공격 특징 정보를 통해 침입을 탐지한다.

대부분의 IDS는 전문 지식의 수동의 기호화에 의해 개발되는 수제 서명에 의거한다. 이 시스템은 공격의 알려진 서명에 감시되고 있는 시스템 위에서 활동에 필적한다. 이 접근에 관한 주요한 문제는 시스템이 새로운 공격을 찾기 위해 개괄할 수 없다라는 것이다. 최근, 관심이 되는 것이 IDS의 발견 모델을 형성한다 것에 데이터 마이닝을 기반으로 둔 것이 있었다. 이 방법은 알려진 공격과 보통의 행동의 그들의 모델을 알려지지 않은 공격을 찾기 위해 일반화할 수 있으며, 그들은 또한 도메인 전문가에 의해 감사 자료의 어려운 분석을 필요로 하는 수동으로 코드화 되었던 모델보다 더 빠르고 더 자동화되었던 방법으로 생성될 수 있다. 침입을 찾는 효과적인 몇 개의 데이터 마이닝 기술은 개발되었다

이 논문 목적은 IDS의 설계에서 더 규칙적이고 자동화된 시스템 제안이다. 일반적으로 침입 탐지 모델의 다양한 감사 데이터 소스를 적용될 수

있는 도구의 집합을 제안하는 것이다.

이 논문의 2장에서는 지능형 침입 탐지 모델과 DM과 GA에 대해 설명하고, 3장에서는 감사 데이터 마이닝의 데이터 선별에 기술하고, 4장에서는 특징 구성에 따른 IDS 모델을 제안한다. 마지막으로 5장에서는 결론과 앞으로의 연구 방향을 논한다.

2. 지능형 침입 탐지 모델

침입 탐지를 위한 기본 전제는 감사 메커니즘이 시스템 이벤트의 레코드가 가능할 때와 합법적인 행동과 침입들의 구별되는 증거가 감사 데이터에서 명백하게 이루어 질 때이다. 감사 데이터의 양과 감사 데이터의 분야의 시스템 특징의 수에서 감사 데이터는 완전한 크기를 가지기 때문에 능률과 지능 데이터 분석 도구들은 시스템 활동의 행동의 발견을 요구한다.

일반적으로 데이터 마이닝은 데이터의 방대한 양에서 기술된 모델들을 추출하는 과정을 한다. 최근 데이터 마이닝에서 재빠른 발전은 통계, 패턴 인식, 학습 기계, 데이터 베이스 분야 등 여러 분야에서 알고리즘들을 이용되게 만들어져 있다. 아래의 알고리즘 유형들은 마이닝 감사 데이터에서 사용되는 일부분이다.

2.1 Classifications

분류는 이미 각 클래스로 구분되어 주어진 데이터 집합만으로 미래의 다른 데이터를 구분할 수 있게 각 클래스에 대한 의미 있는 모델을 만들어 내는 방법이다.

체계적 분류를 위해 결정 트리 분류를 사용한다. 결정 트리 분류기를 사용하는 이유는 첫째로 신경망이나 베이스 분류기에 비해 결정 트리 분류기는 관계형 데이터베이스의 질의 언어인 SQL 문으로 바꾸기 쉽고 신경망을 훈련시키는 것은 일반적으로 시간이 많이 걸리기 때문에 결정 트리 분류기에 초점을 두고 있다.

2.2 Link analysis

데이터 베이스 레코드에서 분야 관계를 결정한다. 감사 데이터에서 시스템 특징의 상호관계들은 정상적으로 사용되는 프로파일들의 구조화하는 기반처럼 조사 될 수 있다. 예를 들어, 사용자의 command 역사 데이터의 부분에서의 command 와 argument사이의 상호관계를 가진다.

2.3 Sequence analysis

순차적 패턴의 모델이다. 이 알고리즘은 감사 이벤트가 자주 함께 일어나는 시간기반의 순차적으로 일어나는 패턴을 발견할 수 있다. 이 이벤트 패턴은 침입 탐지 모델들의 일시적 통계 측정의 통합을 위한 지침을 제공한다.

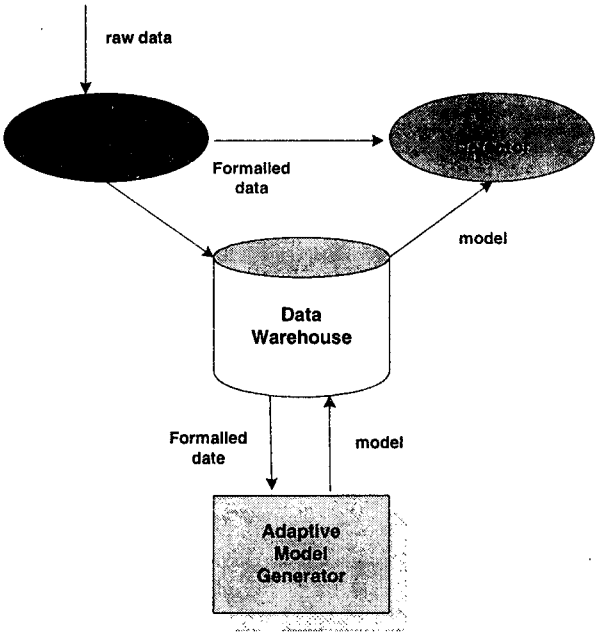
예를 들면, 감사 데이터로부터의 패턴은 이전에 측정되어 포함된 패턴들의 네트워크-기반 denial-of-service(DOS) 공격들은 포함한다.

2.4 유전자 알고리즘을 활용한 분류

GA는 효율적이고 검색 방법이 독립적이고, 분류기 규칙을 학습하기 위해 사용된다. 또한 GA는 특징 구성, 매개 변수를 조절하고, 특징 선택, 개념 학습에 적용된다.

데이터 마이닝 시스템의 대부분은 전통적인 기계학습알고리즘의 변형을 사용한다. 기계 학습에서, 복잡한 시스템의 학습과 시스템의 적절한 출력을 만드는 두 가지 목적을 가지고 있다. 기계학습은 GA 기계학습이거나 GBML(genetic based machine learning)이라 불리는 GA를 기반으로 하고 있다.

기계학습방법은 규칙 집합을 찾기 위한 최적화 문제와 기본적으로 다른 면이 있다. 최적화 문제의 목적은 최적화 해를 찾기 위한 것이고, 최후의 한 종류만이 개체에 수렴하면 되지만 기계학습은 더 나은 규칙을 찾기 위한 것이 아니라, 서로 협력하는 규칙 집합을 찾는 것이다. 일반적으로, GBML에서 두 가지 학습이 있다. 전체 규칙 집합을 하나의 개체로 표현하고, 후보 규칙 집합 집합들의 개체 집단을 유지하고, 그리고 규칙 집합들의 새로운 세대를 생성하기 위한 선택과 유전 연산자를 사용하는 것이 자연스러운 방법으로 여겨질 것이다. 즉, 전통적인 GA를 사용하며, 집단 안에서의 각 실재는 학습 문제에 대한 완전한 해를 표현하는 규칙 집합이다.



그 림 1.The Architecture of Data Mining-based IDS

위의 그림1은 데이터 마이닝을 적용한 모델이다. sensor, detectors, warehouse, 그리고 일반적인 구성요소의 모델로 구성되어 있다. 이 구조는 데이터 분석, 분할, 수집뿐만 아니라 유지할 수 있으며 데이터의 보관, 분배도 할 수도 있다.

3. 감사 데이터의 마이닝

데이터 마이닝은 데이터베이스로부터 분석 대상이 되는 데이터를 선별하는 과정(data selection), 선별된 데이터를 적절한 형태로 가공하고 정제(cleaning, transformation)하는 과정, 변환된 데이터에 마이닝 알고리즘을 수행하는 과정(data mining), 알고리즘의 결과를 재가공하고 기존의 지식과 통합하는 과정을 거치게 됩니다.

3.1 분류

분류는 주어진 객체의 성격을 여러 가능한 클래스 중의 하나로 나누는 과정입니다. 주어진 예제 데이터나 학습 데이터를 이용하여 분류규칙을 학습하거나 미리 정해진 클래스에 따라 분류규칙을 만듭니다. 새로 들어오는 객체(데이터)를 이 분류규칙에 따라 나누게 됩니다.

특징 중에 하나인 클래스 라벨(즉 개념)인 레코드 집합을 주어진 분류 알고리즘은 각 개념에 기술하기 위한 대부분 구별할 수 있는 특징 값을 사용하는 모델을 계산할 수 있다. 분류 모델의 정확성은 트레이닝 데이터에서 제공된 특징의 집합에 의존한다.

3.1.1 Meta-Classification

기초 탐지 모델(즉 classifiers)의 수에 의해 예측득의 상관 관계를 논리적으로 배우기 위한 메커니즘으로 Meta-learning을 사용한다. Meta-classifier의 결과는 모든 기반 탐지 모델의 탐지 능력을 결합한다. 이 접근은 fraud detection의 도메인의 관계에서 경험적 평가와 확장된 연구되어질 수 있고, 효과와 접근을 보여질 수 있다.

3.2 연관 규칙

연관 규칙의 목적은 데이터 베이스 테이블로부터 다양한 특징의 상호 협력적인 관계를 정의하는 것이다. 레코드들에 집합이 주어지고 각 레코드는 항목의 집합이다. support(X)는 X집합의 항목을 포함하는 레코드의 퍼센트이다. 연관 규칙은 X→Y. [c. s]로 표현된다. 여기서 X와 Y는 항목 집합이고 $X \cap Y = \varnothing$, s=support($X \cup Y$)는 규칙의 support이고,

$c = \frac{support(X \cup Y)}{support(X)}$ 는 confidence이다.

예를 들면, 쉘 명령 기록 파일로(명령의 연속과 그들의 인수)부터 사용자의 연관 규칙은 식(1)과 같다.

$$trn \rightarrow rec.humor \ ; \ [0.3, \ 0.1] \quad (1)$$

이것은 사용자가 trn을 호출하는 시간의 30%를 가리키거나 rec.humor에서 새로운 소식을 읽는, 명령기록 파일에서 기록된 활동들의 10%가 이 뉴스그룹을 읽는 것을 가리킨다. 여기서 0.3은 confidence, 0.1은 support이다.

연관 규칙 알고리즘을 감사 데이타에 적용하는 동기는 다음과 같다.

- 감사 데이터는 데이터 베이스 테이블로 구성되어 질 수 있다. 각 행은 감사레코드이고 각 열은 감사레코드들의 field(system feature)이다.
- 프로그램 실행과 사용자의 행동 경험은 시스템 특성 사이의 빈번한 상호협력관계를 나타내는 증거이다.
- 새로운 실행(또는 프로세스)으로부터 총체적인 규칙집합(모든 선행된 실행들의)을 위해 규칙들을 지속적으로 병합할 수 있다.

3.3 Frequent Episodes

많은 공격들의 성질의 이해에서 네트워크 이벤트의 frequent sequential 패턴 들의 연구가 필요하다. 순차적 감사 레코드 패턴을 표현하기 위해 frequent episode를 사용한다. 연관 규칙 알고리즘이 감사 데이터간의 관계를 찾고있는 동안 Frequent Episode는 내부 감사 레코드 패턴을 발견하는데 사용할 수 있다. Frequent Episode는 빈번하게 발생하는 하나의 time-window-특정한 길이-이벤트들의 집합이다. 이벤트들은 window가 이동하는 시간에 특정한 최소 frequency(min_fr)에 발생되어야만 한다. serial 에피소드상의 이벤트들은 적당한 때에 부분 명령에서 발생해야만 한다. 반면 parallel 에피소드는 그런 제약이 없다. X와 Y에 대하여 X+Y는 frequent episode이다.

$$X \rightarrow Y, \ confidence = \frac{frquency(X+Y)}{frequency(X)} \quad (2)$$

, 그리고 support=frequency(X+Y)를 frequent episode rule라 한다. 예를 들면 Web에서 쇼핑몰의 log파일로부터의 frequent serial 에피소드 규칙은

$$home, \ research \rightarrow theory \ ; \ [0.2, \ 0.05], \ [30s] \quad (3)$$

이다. 이것은 홈페이지와 리서치가이드가 명령에 의해 오픈 되어질 때, 30s time window내에 theory그룹의 페이지가 열리는 경우의 20%를, 방문의 sequence가 log파일에서 전제시간(30s)의 5%가 발생함을 보여준다.(즉, 대략 모든 레코드들의 5%정도)

우리는 frequent episode 알고리즘을 감사 흔적들을 분석하는데 적용하려고 한다. 거기에는 프

로그램 실행에서 sequence 정보의 증거가 있다. 사용자 명령들은 비정상행위 탐지를 위한 프로파일을 생각하는데 사용되어진다.

4. 특징구성

침입 탐지에서 데이터 마이닝 접근을 사용하는 가장 큰 도전은 그것이 차례로 다량의 감사 자료에 규칙이 설정하는 프로파일을 계산하는 것을 요구하는 것이다. 그리고 우리가 목표 시스템에서 마이닝 작업에 위압하게 만드는 각 자원의 탐지 모델을 계산이 필요하다. 더욱이, 이 마이닝 학습과정은 탐지 모듈은 정적이 아닌 규칙 집합을 사용하기 때문에 침입 탐지 시스템은 없어서는 안 되고, 연속적인 부분이 되야 한다. 예를 들면, 새로운 버전의 시스템 소프트웨어가 나오면 'normal' 프로파일 규칙을 업데이트해야 한다. 데이터 마이닝은 시간과 저장공간에서 많은 과정이 필요하고, 실시간 탐지는 다양한 경험이 필요하다.

Normal 연결 데이터와 침입 데이터를 Frequent Episode 프로그램에 적용하고, "intrusion only" 패턴들을 발견하기 위해 패턴들의 결과를 비교하게 된다. 패턴 비교 알고리즘은 [3]에서 상세하게 묘사되어 있다. 간단히, 패턴의 수가 매우 많고, 두 개의 데이터 집합으로부터 정확히 매치 되는 패턴이 드물고, 두 개의 episode를 고려한 학습 알고리즘은 아주 다른 특징축의 다른 두 집합의

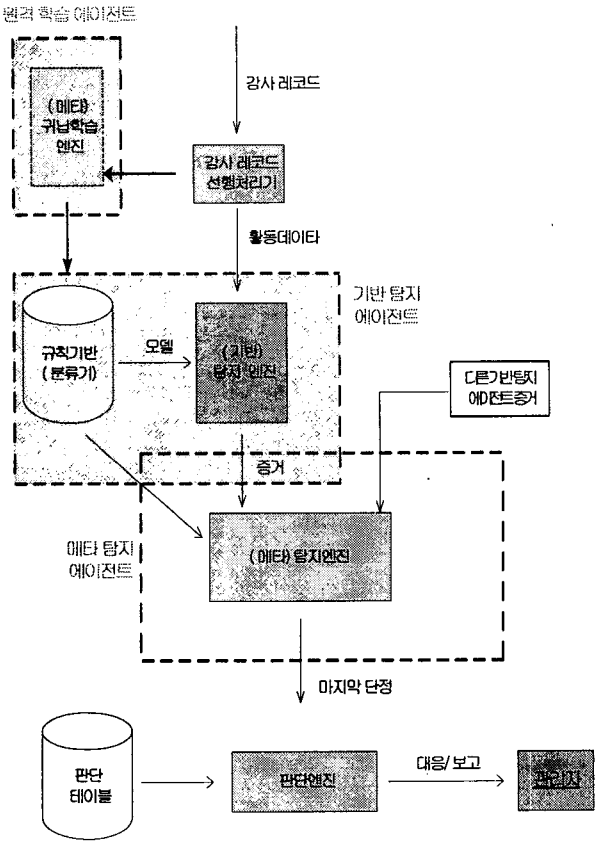


그림2. IDS의구조

관계를 가진 것들은 대부분의 "intrusion only"패턴을 출력하게 된다. 이 침입패턴의 각각은 더 나은 분류 모델의 레코드를 연결하여 특징을 추가하여 지침으로 사용한다.

그림2에서 학습 에이전트와 탐지 에이전트의 두 종류의 지능형 에이전트가 있다. 학습 에이전트는 사용자와 프로그램의 규칙 집합을 유지하고 계산하기 위해 있다. 이것은 기반 탐지 모델과 메타 탐지 모델을 만들어 낸다 학습 에이전트의 일은 감사 데이터의 많은 양으로부터 정확히 계산하는 것이다.

5. 결론 및 향후과제

이 논문에서 침입 탐지의 데이터 마이닝 기술을 사용하는 시스템 구조를 제안했다. 이 구조는 분류, 연관 규칙, frequence episode 프로그램을 사용하여 탐지 모델을 설계하였다. 탐지 모델의 정확성은 올바른 특징 집합과 충분히 트레이닝된 집합에 의존한다. 감사 데이터로부터 구성된 패턴의 계산에 사용하는 연관 규칙과 frequent episode 알고리즘을 제안한다. 또 GA를 사용하여 강건한 개념 학습 시스템을 제안하고, 이러한 방법으로 데이터 마이닝의 핵심적인 기능인 분류를 활용하여 실행할 것이다.

우리는 제안된 모델에 대한 분류기에 대한 실험과 연구가 계속 이루어 질 것이다. 앞으로 분류기의 수행 능력을 향상시키기 위한 연구가 필요하다.

6. 참고문헌

[1] W. lee , S. J. Stolfo. and K. W. Mok "Mining in a data-flow environment : Experience in intrusion detection." submitted for publication, March 1999.

[2] W. lee and S. J. Stolfo. "Data mining approaches for intrusion detection." In Proceedings of the 7th USENI Security Symposium, Dan Antonio, TX, January 1998.

[3] W. Lee, D. J. Stolfo, and K. W. Mok. "Mining in a data-flow enviroment: Experience in intrusion detection." submitted for publication, March 1999.

[4] R. Srikant and R. Agrawal. "Mining generalized association rules." In Proceedings of the 21st VLDB Conference, Zurich, Switzerland, 1995.

[5] R. Srikant. "Fast Algorithms for Mining Association Rules and Sequential Patterns." PhD thesis, University of Wisconsin - Madison, 1996.

[6] H. Mannila, H. Toivonen, and A. I.

Verkamo.
"Discovering frequent episodes in sequences.
In Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining" , Montreal, Canada, August 1995.
[7] P. K. Chan and S. J. Stolfo.
"Toward parallel and distributed learning by meta-learning."
In AAAI Workshop in Knowledge Discovery in Databases, pages 227-240, 1993.
[8] J. Frank.
"Artificial intelligence and intrusion detection: Current and future directions."
In Proceedings of the 17th National Computer Security Conference, October 1994.
[9] R. Agrawal, T. Imielinki, and A. Swami.
"Mining association trules between sets of items in large databases." In Proceedingd of the ACM SIGMOD conference on Management of Data 1993.