데이터마이닝을 이용한 침입 탐지 시스템의 경보데이터 축약기법

허문행*
*안양대학교 디지털미디어학과
e-mail:moonh@anyang.ac.kr

Aggregation Techniques for Alert Data of Intrusion Detection System using Data Mining

MoonHeang Hu*
*Dept. of Digitalmedia, Anyang University.

요 약

이 논문에서는 데이터마이닝의 클러스터링을 이용한 경보 데이터 축약기법을 제안한다. 제안된 클러스터링 기반 경보데이터 축약기법은 데이터간의 유사성을 이용한 경보 데이터의 그룹화를 통해 생성된 모델을 이용하여 새로운 경보 데이터에 대한 분류를 자동화할 수 있다. 이것은 과거에 탐지된 공격의 형태뿐만 아니라 새로운 혹은 변형된 경보의 분류나 분석에도 이용할 수 있다. 또한 생성된 클러스터의 생성 원인의 분석을 이용한 클러스터 간의 시퀀스의 추출을 통해 사용자가 공격의 순차적인 구조나 그 이면에 감추어진 전략을 이해하는데 도움을 주며, 현재의 경보 이후에 발생 가능한 경보들을 예측할 수 있다..

1. 서론

기존의 침입 탐지 시스템은 알려진 공격 형태를 탐지하는 것은 가능하지만 변형된 형태의 공격이나 새로운 형태의 공격의 탐지는 어렵다는 것이다. 또한 개개의 패킷에 대한 검사만을 통해 해당 연결의 침입 여부를 판단하므로, 공격을 탐지했을 경우에 탐지된 공격이 의도하는 의미나 전략과 같은 고수준의 의미를 포착할 수 없다. 그리고 침입 탐지 시스템이 공격의 탐지 결과로서 실시간으로 생성하는 대량의 경보 데이터를 분석하기조차 용이하지 못하다. 침입 탐지는 시스템이나 네트워크에 대한 불법적인 침입이나 자원에 대한 공격으로부터 시스템의 자원을 보호하기 위한 기술이다. 그러나, 현재의 침입 탐지 시스템은 실제의 공격적인 행위가 발생했을 때 그러한 행위 이면의 논리적인 단계나 공격이 의미하는 전략을 포착하기가 쉽지 않으며, 침입 탐지 시스템이 생성하는 대량의 경보 데이터의 효율적인 분석 또한 용이하지 않다.

이 논문에서는 침입 탐지 시스템의 효율성을 높이기 위해 데이터 마이닝의 클러스터링 기법을 이용하여 경보 데이터를 그룹화하고 그 결과를 이용하여 경보 데이터의 상관 관계를 분석하는 방법을 제안하였다. 즉, 클러스터링 기법을 이용하여 경보 데이터를 사용자가 원하는 개수의 그룹으로 분류할 수 있게 하였으며, 생성된 클러스터 모델을 이용하여 새로운 경보 또한 적당한 클러스터로 분류할 수 있도록 하였다. 또한, 결과 클러스터의 생성 원인이 되는 이전의 경보의 분포를 분석하여 클러스터 간의 시퀀스를 생성하였고, 생성된 각각의 클러스터 시퀀스를 통합하여 클러스터들의 시퀀스를 추출하여 발생한 경보의 향후 가능한 경보 타입을 예측하기 용이한 방법을 제공하였다. 논문의 2 장에서는 관련 연구로써 침입 탐지와 경보 데이터의 상관 관계 분석과 클러스터링 기법에 대해 설명하고 3 장에서는 경보 데이터에 클러스터링 분석을 적용하기 위한 절차에 대해 기술하며 4 장에서는 구현을 위한 시스템의 구성에

대해 기술하고 마지막으로 5 장에서 결론을 맺는다.

2. 관련 연구

침입 탐지는 보호하고자 하는 네트워크나 시스템의 사용을 실시간으로 모니터링하여 시스템의 보안 요소를 침해하는 행위를 탐지하는 기술이다[1]. 이러한 침입 탐지 시스템은 자원, 모델, 기술 세가지 요소로 구성된다. 자원은 침입 탐지 시스템이 목표 시스템에서 보호해야 할 시스템의 자원이다. 모델은 이러한 자원의 행위나 자원에 가해진 행위가 정상적인지, 불법적인지를 결정하는 요소이다. 기술은 미리 설정된 모델과 실제 시스템의 행동을 비교하여 그 행동이 정상적인지, 공격적인지를 확인하는 요소이다. 침입 탐지 기술은 그 방법론에 따라 오용 탐지 기법 (misuse detection)과 이상 탐지 기법 (anomaly detection)으로 분류된다[3]. 그러나, 기존의 침입 탐지 시스템은 가해진 공격에 대하여 저수준의 탐지만 가능하며, 그 공격 이면의 논리적인 절차나 공격의 전략을 포착할 수 없다. 실제 공격의 상황에서 침입 탐지 시스템에 의해 생성되는 많은 양의 경보 데이터를 수동으로 분석하여 공격의 절차나 전략을 추출할 수 있는 경보의 상관 관계를 생성하는 것은 어려운 작업이다. 따라서 경보 데이터 간의 연관성 분석을 통해, 새로운 공격의 탐지나 보다 정확한 탐지를 위한 침입 탐지 모델을 구축하고, 사용자에게는 보다 이해하기 용이한 정보를 제공할 수 있다. 개연적 경보 상관 관계 분석은 경보 데이터의 속성의 유사성을 이용하여 경보 데이터 간의 상관 관계를 분석하는 기법이다[2]. [5]는 경보 데이터의 통합과 상관 관계 분석 기법을 제안하였다. 특히, [5]에서 제안된 상관 관계 분석 방법은 어떤 타입의 경보가 주어진 경보 유형의 다음에 오는지를 기술하기 위한 결과 메커니즘을 이용하였다. 그러나 이 결과 메커니즘은 가능한 모든 경보 데이터들이 서로 관련되기 위한 충분한 정보를 제공하지 않는다는 단점이 있다.

이 논문에서는 데이터 마이닝의 클러스터링 기법을 기반하여 경보 데이터를 분석하고, 이를 이용하여 공격의 시퀀스를 추출하는 방법을 제안한다. 데이터 마이닝은 "다량의 저장된 데이터로부터, 이전에 잘 알려지지 않았지만, 묵시적이고, 잠재적으로 유용한 정보를 추출하는 일련의 작업"이다[Usam96]. 특히, 데이터 마이닝 기법 중에서 클러스터링 분석은 유사도가 높은 데이터들을 같은 그룹으로 분류하여 주어진 데이터의 분포나 패턴을 찾아내는 기법이다[2]. 이와 같은 클러스터링 분석 기법을 경보 데이터의 분석에 적용함으로써, 대량의 경보 데이터를 보다 효율적으로 분석할 수 있으며, 데이터의 그룹화를 통해 고수준의 의미를 추출할 수 있다.

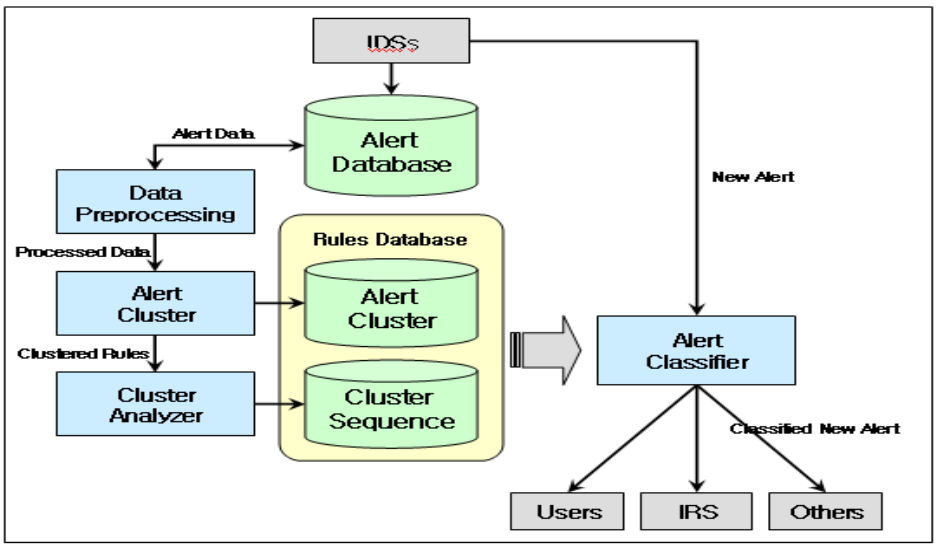
3. 클러스터링 기반 경보데이터 축약

클러스터링은 잠재적인 데이터에서 그룹들을 탐사하거나 관심있는 분포를 확인하는데 유용한 방법이다. 이 기법은 개체들의 집합을 개체의 클래스들로 그룹화하는 절차이다. 이때, 동일한 클러스터에 속하는 개체들은 유사성을 가지고, 다른 클러스터에 속하는 개체간에는 상서성을 가진다.

1) 데이터 수집 : 잠재된 데이터 소스로부터 관련된 데이터 개체를 추출한다.
2) 데이터 정제 : 소스로부터 추출된 데이터를 적절히 조작한다.
3) 데이터 가공 : 이 단계에서 유사도 측정 기준을 선택하고 데이터의 특성이나 차원을 검토한다.
4) 데이터 검사 :주어진 데이터가 클러스터링 하기에 적당한지를 검사한다.
5) 클러스터링 전략의 수립 : 클러스터링 알고리즘과 그에 맞는 초기값들을 선택하는 과정이다.
6) 결과 확인 : 클러스터링 결과를 확인하는 과정으로 주로 수동적인 방법이나 시각적인 표현을 이용한다.
7) 결과 분석 : 결론을 이끌어내고 향후 분석을 위해 다른 부분과 클러스터링된 결과를 통합한다.

경보 데이터의 유사성을 분석하는 시스템은 Data Preprocessor, Alert Cluster, Cluster Analyzer, Alert Classifier로 구성된다. Data Preprocessor는 입력된 데이터 집합에 대해 Alert Cluster가 클러스터링을 수행할 수 있도록 전처리를 한다. 여기에서 효율적이고 보다 정확한 클러스터링을 위하여 도메인 지식에 의한 확장 속성을 추가하고 선택된 속성에 대해 정규화를 수행한다. Alert Cluster는Data Preprocessor에 의해 처리된 데이터에 대해 실제 클러스터링을 수행한다. 이 모듈의 최종 결과는 그룹화된 데이터의 집합들이다.

그 결과는 룰 데이터베이스에 저장이 되고, 이는 이후에 새로운 경보의 자동적인 분류나 생성된 클러스터 간의 연관 관계 분석에 이용된다. Cluster Analyzer는 클러스터링의 수행을 통해 생성된 클러



[그림 1] 클러스터링을 이용한 경보데이터 축약 시스템

스터의 생성 원인을 분석한다. 이것에 의해 수행된 결과는 클러스터의 시퀀스로 표현된다. 이를 이용하여 우리는 클러스터간의 연관 관계를 분석할 수 있으며, 특정 경보에 대한 차후 가능한 경보의 집합의 예측에 이용할 수 있다. Alert Classifier는 Alert Cluster에 의해 생성된 클러스터 모델을 이용하여 새로운 경보를 적절한 클러스터로 분류하고, Cluster Analyzer의 결과로서 생성된 시퀀스를 이용하여 차후 발생 가능한 경보들을 추출하는 역할을 수행한다. 경보 데이터의 유사성을 분석하기 위한 시스템의 구조는 [그림 1]과 같다.

4. 프로토타입 구축및 실험

이 논문에서 소개하는 시스템의 구현 환경은 Compac Server에서 JAVA와 Pro*C로 구현하였다. 실험한플랫폼은 Linux 7.1와 Solaris 7에서 수행하였으며, DBMS는 Oracle 8.1.7을 사용하였다.

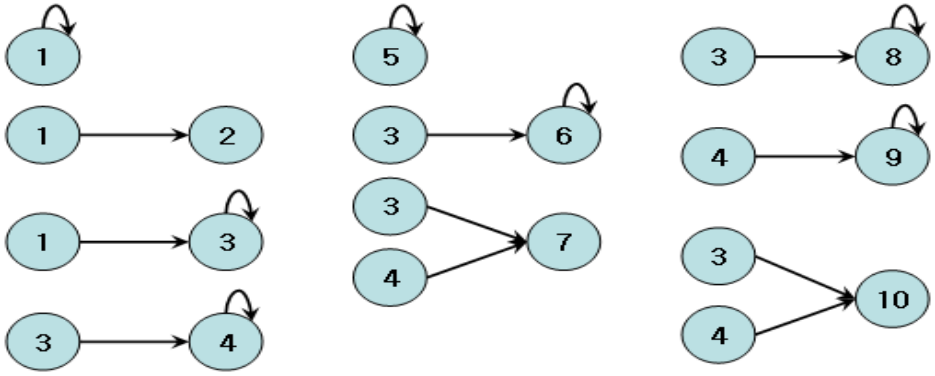
이 논문에서 제안된 방법에 대해 두 가지 측면에 대해 평가를 수행한다. 첫 번째는 구현된 시스템의 클러스터링 성능을 평가하기 위한 실험이다. 이 실험은 구현된 시스템에 의해 생성되는 각 클러스터의 정확도를 평가하는 것이다. 두 번째는 생성된 클러스터에 대해 각 클러스터의 이전 클러스터를 정의하고 이를 기반으로 클러스터의 시퀀스를 생성할 수 있는지의 여부를 평가하기 위한 실험이다.

실험을 위해 이 논문에서 사용한 실험 데이터는 KDD Cup 1999 데이터 집합이다[KDD99]. 이 데이터 집합은 DARPA 1998를 이용하여 몇가지 속성을 추가하여 생성한 데이터 집합이다[Darp98]. DARPA 1998의 트레이닝 데이터는 7주간의 네트워크 트래픽으로 구성된 TCP Dump 데이터이다. 이 데이터 집합은 약 5,000,000 개의 데이터 인스턴스로 구성되어 있으며, 네트워크 환경 상에서 가능한 다양한 형태의 침입을 포함하고 있다.

입력 데이터에 대해 클러스터링을 수행하기 위해서는 먼저 사용자 입력 변수를 결정하여야 한다. 이 시스템에는 사용자가 입력하여야 하는 변수가 2개 있다. 선택된 포인터를 대표값으로 변환하기 위한 수축률과 클러스터의 대표값의 개수이다. 실험을 위한 수축률과 대표값의 개수를 선택하기 위하여 우리는 임의로 선택된 10%의 트레이닝 데이터 셋에 대해 클러스터링을 수행하였다. 실험을 한 결과 대표값의 수는 비교적 많을 경우 일 때 클러스터링의 결과가 향상 되었지만, 대표값의 수가 너무 많을 경우에 실행 시간이 너무 커진다는 문제점이 발생하게 된다는 것을 알 수 있었다.

두 번째 실험은 생성된 클러스터에 대해 각 클러스터의 이전 클러스터를 정의하고 이를 기반으로 클러스터의 시퀀스를 생성할 수 있는지의 여부를 평가하기 위한 실험이다. 입력 데이터 집합은 snort에 의해 생성된 실제 경보 데이터를 이용하였으며 사용자 정의 변수는 앞의 실험에 의해 구하여진 값을 사용하였다. 입력 데이터는 약 2시간의 시뮬레이션에 의해 생성된 데이터로서 약 15,000개의 데이터 인스턴스로 구성되어 있다.

원형은 클러스터를 나타내며, 화살표는 공격의 진행 방향을 나타낸다. [그림 2]는 실험 결과 나타난 클러스터의 시퀀스이다.



[그림 2] 클러스터의 시퀀스

이와 같은 클러스터의 시퀀스를 추출함으로써 특정 클러스터에 포함된 경보 이후에 가능한 경보의 그룹을 알 수 있다.

5. 결론

이 논문에서는 데이터 마이닝의 클러스터링 기법을 적용하여 침입탐지 시스템의 경보데이터 축약 기법을 제안하였다. 클러스터링을 수행하기 위해, 침입탐지 시스템에서 생성하는 경보 데이터를 관계형 데이터베이스에 저장하기 위한 경보 데이터의 스키마를 설계하였다. 또한, 효율적인 클러스터링의 수행을 위하여 경보 데이터의 기본적인 속성을 이용하여 확장된 속성을 정의하였다. 클러스터링 알고리즘은 경보 데이터의 특성을 고려하여 다차원의 속성을 가지는 데이터 집합에 대해서도 클러스터링이 가능한 CURE 알고리즘을 변형, 구현하였다.

새로운 경보를 적절한 클러스터에 할당하기 위하여, 클러스터의 대표값을 이용하여 개개 데이터 인스턴스와 생성된 클러스터들과의 유사도를 측정하는 방법을 정의하고 구현하였으며, 클러스터에 포함된 경보의 이전 경보를 분석하여 클러스터간의 시퀀스를 추출하기 위한 방법 또한 정의하였다. 제안된 클러스터링 기반 경보데이터 축약기법을 평가하기 위해 KDD Cup 1999 데이터 집합에 대해 클러스터링을 수행함으로써, 구현된 클러스터링의 성능을 실험하였으며, 클러스터들의 시퀀스 생성 여부를 실험하였다.

이 논문에서 제안한 클러스터링을 이용한 경보 데이터의 분석은 방법은 다음과 같은 장점을 가진다. 먼저, 데이터간의 유사성을 이용한 경보 데이터의 그룹화를 통해 생성된 모델을 이용하여 새로운 경보 데이터에 대한 분류를 자동화할 수 있다. 이것은 과거에 탐지된 공격의 형태뿐만 아니라 새로운 혹은 변형된 경보의 분류나 분석에도 이용할 수 있다. 두 번째로 생성된 클러스터의 생성 원인의 분석을 이용한 클러스터 간의 시퀀스의 추출을 통해 사용자가 공격의 순차적인 구조나 그 이면에 감추어진 전략을 이해하는데 도움을 주며, 현재의 경보 이후에 발생 가능한 경보들을 예측할 수 있으므로 이들을 필요로 하는 보안 분야에 적용할 수 있다.

참고문헌

[1] Moon Sun Shin, HoSung Moon, KeunHo Ryu, JinOh Kim and KiYoung Kim, "Applying Data Mining Techniques to Analyze Alert Data", APWeb2003, LNCS 2642 pp.193-200, SpringerVerlag.

[2] A. Valdes and K. Skinner, "Probabilistic alert correlation", In Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection (RAID 2001), pages 5468, 2001.

[3] P. Ning and Y. Cui., "An intrusion alert correlator based on prerequisites of intrusions", Technical Report TR-2002-01, Department of Computer Science, North Carolina State Univ., Jan. 2002.

[4] D. Curry and H. Debar, "Intrusion detection message exchange format data model and extensible markup language document type definition", Internet Draft, draft-ietf-idwg-idmef-xml-03.txt, Feb. 2001.

[5] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases" In Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 207-216, 1993.

[6] 신문선, 류근호, "침입탐지시스템의 성능향상을 위한 오경보 분류 모델 구현", 정보과학회논문지:데이터베이스 2007년 12월