한국산학기술학회논문지
Vol. 12, No. 1 pp. 459-466, 2011

침입탐지시스템의 경보데이터 분석을 위한 데이터 마이닝 프레임워크

신문선[1*]
[1]안양대학교 교양학부

An Alert Data Mining Framework for Intrusion Detection System

Moon-Sun Shin[1*]
[1]**Division of Liberal Arts, Anyang University**

요 약 이 논문에서는 침입 탐지시스템의 체계적인 경보데이터관리 및 경보데이터 상관관계 분석을 위하여 데이터 마이닝 기법을 적용한 경보 데이터 마이닝 프레임워크를 제안한다. 적용된 마이닝 기법은 속성기반 연관규칙, 속성기반 빈발에피소드, 오경보 분류, 그리고 순서기반 클러스터링이다. 이들 구성요소들은 각각 대량의 경보 데이터들로부터 알려지지 않은 패턴을 탐사하여 공격시나리오를 유추하거나, 공격 순서를 예측하는 것이 가능하며, 데이터의 그룹화를 통해 고수준의 의미를 추출할 수 있게 해준다. 실험 및 평가를 위하여 제안된 경보데이터 마이닝 프레임워크의 프로토타입을 구축하였으며 프레임워크의 기능을 검증하였다. 이 논문에서 제안한 경보 데이터 마이닝 프레임워크는 기존의 경보데이터 상관관계분석에서는 해결하지 못했던 통합적인 경보 상관관계 분석 기능을 수행할 뿐만 아니라 대량의 경보데이터에 대한 필터링을 수행하는 장점을 가진다. 또한 추출된 규칙 및 공격시나리오는 침입탐지시스템의 실시간 대응에 활용될 수 있다.

Abstract In this paper, we proposed a data mining framework for the management of alerts in order to improve the performance of the intrusion detection systems. The proposed alert data mining framework performs alert correlation analysis by using mining tasks such as axis-based association rule, axis-based frequent episodes and order-based clustering. It also provides the capability of classify false alarms in order to reduce false alarms. We also analyzed the characteristics of the proposed system through the implementation and evaluation of the proposed system. The proposed alert data mining framework performs not only the alert correlation analysis but also the false alarm classification. The alert data mining framework can find out the unknown patterns of the alerts. It also can be applied to predict attacks in progress and to understand logical steps and strategies behind series of attacks using sequences of clusters and to classify false alerts from intrusion detection system. The final rules that were generated by alert data mining framework can be used to the real time response of the intrusion detection system.

Key Words : Intrusion detection system, Alert data, Data mining, Alert correlation analysis, Alert data mining framework

1. 서론

개방형 네트워크의 특징을 가지고 있는 인터넷의 구조상 침입행위를 차단하는 것이 용이하지 않다. 따라서 침입탐지시스템을 설치하여 각 조직의 자원이나 시스템을 보호하고 있으나 침입 탐지 시스템이 공격의 탐지 결과로서 실시간으로 생성하는 대량의 경보 데이터를 분석하기조차 용이하지 못하다.

이러한 문제점을 해결하기 위해 경보 데이터 상관관계 분석(alert correlation analysis)에 대한 연구가 진행되고

*교신저자 : 신문선(msshin9@anyang.ac.kr)
접수일 10년 12월 22일 수정일 11년 01월 05일 게재확정일 11년 01월 13일

있다. 경보 데이터의 상관관계 분석은 침입 탐지 시스템에 의해 생성되는 경보 데이터를 분석하여 침입 탐지 시스템의 효율성을 향상시키기 위한 방법이다. 경보 데이터 간의 연관성 분석을 통해, 새로운 공격의 탐지나 보다 정확한 탐지를 위한 침입 탐지 모델을 구축하고, 보안 관리자에게는 용이한 정보를 제공하는 것이 가능하다.

이 논문에서는 침입 탐지 시스템에서 생성되는 대규모의 경보데이터 관리와 경보데이터 상관관계 분석을 위한 데이터 마이닝 프레임워크를 제안하고 프로토타입을 구현하며 실험 및 평가를 통하여 제안된 프레임워크를 검증한다.

이 논문의 구성은 다음과 같다. 제 2 장에서는 관련 연구로써 침입 탐지시스템과 경보 데이터의 상관관계 분석에 관한 연구들을 고찰한 후 기존 연구의 문제점들을 제시한다. 제 3 장에서는 경보 데이터 마이닝 프레임워크를 제안하고 각 구성요소를 기술하며 각 구성요소들의 알고리즘을 설계한다. 제 4 장과 제 5 장에서는 프로토타입의 구축을 통해 제안한 프레임워크의 실험 평가를 기술하며 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

침입탐지시스템에서는 침입 혹은 공격에 대하여 경보데이터를 발생시킨다. 따라서 보안 관리자는 침입탐지시스템의 경보를 항상 모니터링 하여야 한다. 그러나 대량의 경보 데이터 발생과 오경보의 발생 등은 침입탐지시스템의 성능을 저하시키는 결과를 초래하게 된다. 따라서 경보데이터의 통합 관리와 경보 상관성 분석, 오경보 감소 등을 위해서 경보 데이터를 분석하고, 이를 이용하여 공격의 시퀀스를 추출하고 오경보를 분류하거나 경보데이터 필터링 등을 수행한다.

개연적 경보 상관관계 분석은 경보 데이터의 속성의 유사성을 이용하여 경보 데이터 간의 상관관계를 분석하는 기법이다. 속성간의 유사성을 이용하여 공격 타입 간의 유사성을 정의하고 이를 이용하여 공격 타입 간의 연관 관계를 추출한다. 그러나 이 방법은 선택된 속성에 의존적이며, 경보 데이터 간의 인과 관계를 완벽하게 탐사하기에 적당하지 못하다는 단점을 가지고 있다. [2]에서는 발견 학습을 이용한 접근 방법을 “포트탐지공격(stealthy portscan)”을 탐지하기 위해 적용하였다. 비록 발견 학습을 경보 데이터 상관관계 분석에 이용하였지만, 이 방법 또한 경보 데이터 간의 인과 관계를 완벽하게 분석하지 못하였다.

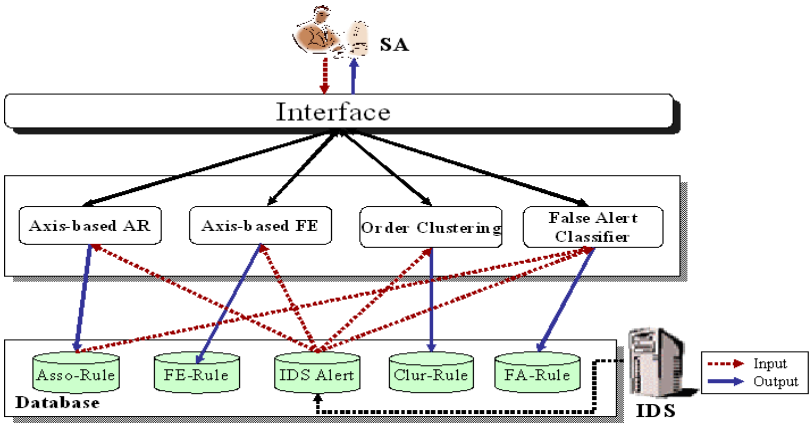
[3]은 경보 데이터의 통합과 상관관계 분석 기법을 제안하였다. 특히, [3]에서 제안된 상관관계 분석 방법은 어떤 타입의 경보가 주어진 경보 유형의 다음에 오는지를 기술하기 위한 결과 메커니즘을 이용하였다. 이것은 오용 탐지 기법과 유사하다. 그러나 이 결과 메커니즘은 단지 경보의 유형과 경보에 의한 프로브, 보안 레벨, 결과 정의에 포함된 두 경보 간의 시간 간격만을 사용하며, 가능한 모든 경보 데이터들이 서로 관련되기 위한 충분한 정보를 제공하지 않는다는 단점이 있다. 게다가 공격자가 어떻게 공격의 시퀀스를 조정할 것인지 예측하는 것 또한 쉽지 않다. 또 다른 접근 방법으로 공격 시나리오가 포함되어 있는 학습 데이터 집합에 기계 학습 기법을 적용하여 경보 상관관계 모델을 학습하는 기법이 있다. 이 방법은 경보 데이터의 상관관계 분석을 위한 모델을 자동적으로 생성할 수 있는 장점이 있지만, 매 적용마다 학습이 필요하며, 결과 모델이 학습 데이터에 의존적이므로 학습 데이터에 포함되지 않은 공격 시나리오는 탐지할 수 없는 단점이 있다.

프랑스의 CERT에서는 경보데이터관리를 위한 데이터베이스 스키마를 설계하고 XML형태의 경보들을 통합하여 클러스터링과 병합과정을 거쳐 상관관계분석을 하고 공격에 대한 광역 진단을 결정하는 프레임워크를 마련하는 프로젝트[7]를 진행하였다. 이 경우 다양한 침입탐지 시스템으로부터의 경보를 통합하고 관리하는 기능을 제공하고 있다. 그러나 여전히 표준화된 경보메시지의 부재로 XML형태로 변환하는 과정이 필요하다.

따라서 이 논문에서는 데이터 마이닝 기법 기반 경보 상관 관계 분석과 오경보 분류 기능을 지원하는 경보 데이터 마이닝 프레임워크를 제안한다. 데이터 마이닝은 " 다량의 저장된 데이터로부터, 이전에 잘 알려지지 않았지만, 묵시적이고, 잠재적으로 유용한 정보를 추출하는 일련의 작업[5]으로 정의된다.

특히 데이터 마이닝 기법은 유사도가 높은 데이터들을 같은 그룹으로 분류하여 주어진 데이터의 분포나 패턴을 찾아내거나 혹은 그룹화를 통하여 데이터 필터링의 효과를 가질 수 있다.

또한 연관규칙이나 순차 패턴, 빈발에피소드 등의 데이터마이닝 기법들은 데이터 간의 상관관계 분석에 유용하게 활용될 수 있다[1]. 즉 자주 발생하는 패턴들이나 규칙들은 서로 상관도가 높은 것으로 예측되어 경보 데이터의 분석에 적용할 수 있으며, 대량의 경보 데이터들로부터 알려지지 않은 패턴을 탐사하여 공격시나리오를 유추하거나, 공격 순서의 예측이 가능하며, 데이터의 그룹화를 통해 고수준의 의미를 추출할 수 있다. 데이터 마이닝 기법 기반 경보데이터 분석은 경보데이터 량의 감소와 알려지지 않은 데이터간의 연관성 분석을 통하여 침입 탐지율을 향상 시킬 수 있다. 또한 결정트리기반의 오경보 분류 모델은 경보데이터의 오



[그림 1] 경보 데이터 마이닝 프레임워크

경보율을 감소시킴으로서 침입탐지시스템의 탐지율을 향상시키게 된다.

3. 경보데이터 마이닝 프레임워크

이 논문에서 제안한 경보데이터 마이닝 프레임워크는 속성기반 연관규칙, 속성기반 빈발 에피소드, 순서 클러스터링, 오경보 분류모델 등 4개의 컴포넌트로 구성된다. 각 구성요소들은 경보데이터의 특성에 맞게 기존의 알고리즘을 확장설계하여 구현되었다. 그림 1은 제안된 경보데이터 마이닝 프레임워크의 각 구성요소와 구성요소간의 관계를 보여준다.

각 구성요소들을 살펴보면 속성기반 연관규칙 컴포넌트는 경보데이터 속성 간의 연관성을 추출하는 기능을 수행하게 되며 속성기반 빈발에피소드 컴포넌트는 주어진 타임윈도우내에서 경보간 연관성을 탐사한다. 순서기반 클러스터링 컴포넌트는 경보데이터를 클러스터링하여 경보축약을 지원하며 또한 각 클러스터간의 순서를 이용하여 경보 시퀀스를 유추하여 경보 이후의 가능한 경보를 예측하는 기능을 수행한다.

오경보 분류 컴포넌트는 훈련데이터를 이용하여 오경보 분류 모델을 생성하여 테스트 데이터가 주어지면 생성된 결정트리를 이용하여 오경보인지 아닌지를 분류하는 기능을 수행한다. 이는 침입탐지시스템을 지원하여 오경보 필터링에 활용될 수 있다.

3.1 속성기반 연관규칙

속성기반 연관규칙 컴포넌트는 APRIORI 알고리즘을 확장하여 항목 선택 시 사용자가 관심 있는 항목을 선택하여 마이닝을 수행함으로써 규칙의 수를 최소화할 수 있다. 관심 있는 속성의 선택을 속성제약조건으로 선택된 속성에 대해서만 연관규칙을 탐사하는 것이다.

속성들 간의 연관성을 탐사하여 많은 양의 경보데이터 속성간의 연관성을 분석할 수 있으며 항목제약사항에 따라 관심 있는 속성들 간의 연관성을 분석하며, 불필요한 규칙의 생성을 줄일 수 있다.

3.2 속성기반 빈발에피소드

속성기반 빈발에피소드 컴포넌트는 경보 데이터간 빈발에피소드 탐사를 위한 시퀀스 패턴을 찾기 위해서 WinEPI 알고리즘을 기반으로 하여 주어진 윈도우 크기만큼씩 투플들을 정렬하여 빈발하는 시퀀스 패턴을 탐사한다. 전체 윈도우 테이블 수는 전체 투플의 끝나는 시간에서 시작하는 시간을 뺀 후 주어진 윈도우 시간(즉, 타임단위 x 윈도우크기)을 더하면 전체 윈도우 테이블 수가 계산이 된다. 그리고 투플 간의 상관관계를 고려하도록 참조속성이라는 항목 제약사항을 추가한다. 참조 속성은 어느 한 속성 값이 다른 속성의 값을 참조 할 수 있다는 것이다. 예를 들어, 네트워크 데이터에서 사용자와 행위에 대한 속성이 있다고 하자. 여기서 행위는 사용자에 의해서 얻어지는 속성이다. 그러므로 동일한 사용자에 대해서는 행위 속성에 대한 값을 참조할 수 있다.

3.3 순서기반 클러스터링

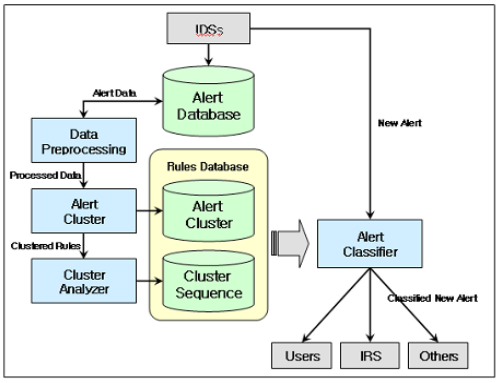
경보 데이터 간의 유사성을 분석하고, 이를 이용하여 유사한 경보 데이터를 그룹화 한다. 클러스터링에서 개개 데이터 개체 간의 유사성은 근접 지수에 의해 정의된다. 이 근접 지수는 두 데이터 개체 간의 유사성이나 연관성을 측정할 수 있는 함수이다. 개체 간의 유사성을 측정하는 방법에는 여러 가지가 있지만 주로 거리개념을 이용하여 측정한다. 본 논문에서는 개체 간의 유사성을 정의

하기 위해서 유클리드 거리 함수를 이용한다. 이는 동일한 속성 값들을 가지는 데이터 개체는 유사하다는 가정을 기반으로 한다. n개의 속성을 가지는 데이터 개체 x ,y 가 주어지고 각각의 속성의 값이 x_i ,y_i ($0 \le i \le n$)로 정의될 때 두 개체 간의 거리를 추출하기 위한 함수는 아래 식 1과 같이 정의된다.

$$dis-inst(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \qquad \text{(식 1)}$$

저장된 경보 데이터와 정의된 유사도 측정 함수를 이용하여 클러스터링을 수행하기 위한 절차는 아래와 같은 모듈에서 각각 수행되어진다.

경보 데이터의 유사성을 분석하는 순서기반 클러스터링 컴포넌트 아키텍쳐는 그림 2와 같다. 그림 2의 순서기반 클러스터링 컴포넌트 아키텍쳐는 Data Preprocessor, Alert Cluster, Cluster Analyzer, Alert Classifier로 구성된다.



[그림 2] 순서기반 클러스터링 컴포넌트 아키텍쳐

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 Cluster는 Data 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 클러스터의 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.

3.4 오경보 분류

침입탐지시스템은 공격이라고 판단되면 경보를 발생하여 보안 관리자에게 알려주거나 자체적으로 대응을 하게 된다. 실제 침입과 침입이 아닌 정상행위에 대해 4가지로 판단 할 수가 있다. 1) 정상행위를 정상행위로 판단하는 경우(TN), 2) 침입행위를 침입으로 정확히 탐지하는 경우(TP), 3) 정상행위를 침입으로 오인하는 경우(FP), 4) 침입을 정상행위로 오인하는 경우(FN)이고, 오경보라는 것은 3)과 4)의 경우를 말하고 있으며 이중 3)False Positive가 전체 오경보의 대부분을 차지하고 있다. 즉, False Positive는 정상 행위를 침입행위로 오인하여 판단하는 것을 의미한다. 경보들 중 이러한 오경보들은 네트워크 전반에 걸친 보안 서비스의 질을 하락시키는 원인이 된다.

오경보중 False Positive가 생기게 되는 원인으로는 패턴 매칭 방식을 이용하기 때문에 침입탐지시스템의 시그네처와 유사한 형태의 패턴을 공격으로 탐지하는 경우이다. False Positive 는 경보데이터 발생 후에 관리자의 분석을 통해 판단할 수밖에 없는 데이터이다.

따라서 False Positives를 분석하기 위해 이미 정상 및 공격으로 검증된 네트워크 패킷 데이터의 속성을 추출하고, 이들 패킷 데이터 중에서 정상으로 판정된 패킷을 침입탐지 시스템에 통과시켜 공격으로 오인하는 데이터를 False Positive로 정의한다. 또한 저장된 네트워크 패킷 데이터는 데이터 마이닝 기법을 통해 분류모델 구축에 사용된다.

하지만 네트워크 패킷데이터는 원시 데이터이기 때문에 관계형 데이터베이스에 저장하기 위해서는 데이터 가공이 필수적이다. 따라서 원시 데이터에서 속성을 추출하여 데이터베이스에 저장하기 위해 전처리 프로세서 모듈을 이용하여 아스키 형태로 변환하였다[1].

먼저 네트워크 패킷 데이터를 수집하여 전처리 프로세서를 거친 후 연관규칙 기반과 통계적인 분석 기법을 사용하여 유용한 속성들을 선택한다. 구축된 모델은 실험 데이터를 이용하여 분류 규칙을 분석한다.

선택된 속성들을 가지고 분류 모델을 구축하기 위해

먼저 루트 속성 결정을 위한 방법으로 정보이득을 계산하고 엔트로피와 정보예측치를 구하여 각 속성에 대한 값을 계산하여 각 노드를 결정하여 모델을 구축한다.

저장된 학습데이터를 이용하여 분류 모델을 구축하기 위한 단계별 수행 내용은 다음과 같다.

1) 속성 선택

의사결정트리의 뿌리마디부터 끝마디까지 사용할 속성들을 결정하는 부분이다. 속성을 선택하는 방법에는 여러 가지가 있을 수 있으나 이 연구에서는 데이터 마이닝 기법 중의 하나인 연관규칙의 빈발집합을 이용한 방법과 통계적 분석방법인 상관분석을 이용하여 사용한다. 연관규칙[5]이란 어떤 사건이 발생하면 다른 사건도 따라 발생하는 사건 사이의 강한 연관성을 의미하는 것이다. 따라서 이 연관규칙을 이용하여 빈발항목을 찾아냄으로써 각각의 패킷들 간에 강한 연관성을 가진 속성들을 지지도를 기반으로 선택할 수 있는 것이다. 이 논문에서는 마디가 될 속성을 선택하여 속성목록에 저장한다. 상관관계 분석은 둘 또는 그 이상의 변수들 간에 존재하는 상관 관계정도를 분석하는 것을 말한다. 변수들 간의 상호 관계정도를 분석하는 통계적인 기법으로써 이 연구에서는 다중 속성들 간의 분석을 위해 셋 또는 그 이상의 변수간의 관계 정도를 밝히는 방법인 다중 상관관계 분석 방법을 이용하였다. 상관분석을 통하여 각 속성들 간의 상관계수를 계산하여 상관계수가 양적 선형관계에 있는 속성들을 선택하여 결정트리의 마디로 결정한다.

2) 훈련 데이터를 이용한 분류규칙 생성

속성 선택 단계에서 저장된 속성목록의 속성 값들을 마디로 하여 분류모델을 생성하는 과정이다. 먼저 연속적인 값을 일정한 도메인을 가지는 이산적인 값으로 변환하는 작업을 수행한다. 의사결정트리에서 하위노드로 가지를 생성할 때 연속적인 값을 그대로 사용한다면 불필요한 개수의 가지가 생성되게 된다. 그래서 일정한 범위를 정해 이산적인 값으로 변환하는 것이다. 두 번째로 뿌리 노드 결정 작업을 수행한다. 각 속성들의 정보값을 구하고 불순도/순수도를 측정한다. 불순도 함수로는 엔트로피 지수를 이용하였다. 각각의 마디들이 가지는 내부 항목, 즉 속성들이 가지고 있는 이산적인 값들에 대해서도 불순도를 측정하고 측정된 각각의 엔트로피 값을 이용해 최종적으로 정보 이득율을 계산한다.

각 속성들에 대해 정보 이득율(Information Gain)을 계산한 후 가장 높은 속성을 뿌리 마디로 결정한다. 재귀적인 방법으로 더 이상 분할이 일어나지 않을 때까지, 즉 끝마디까지 마디가 분할이 이루어지면 의사결정 트리가 종료된다. 트리의 분할이 종료되면 부모마디부터 끝마디까지 이어지는 규칙들이 생성되고 이 규칙은 규칙 데이터베이스에 삽입하여 두 번째 단계를 마치게 된다.

3) 실험 데이터를 이용한 분류 모델의 정확도 평가

훈련 데이터를 이용해 생성한 분류 규칙의 정확도를 평가하는 단계로써 훈련데이터의 일부분을 추출하거나 또는 소량의 실험용 데이터를 이용하여 수행한다. 이를 통해 관리자는 생성된 분류 규칙의 정확도 및 과적용 등을 분석하고 필요한 경우 가지치기를 통해 의사결정트리의 정확도를 높인다.

4) 실제 테이터를 적용한 데이터 분류

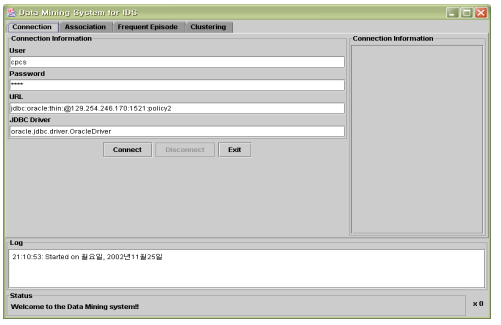
이 단계에서는 분류할 데이터의 집합을 구축된 분류모델에 적용시키는 과정으로써 데이터베이스에 저장된 분류 규칙들과의 패턴 매칭을 통해 데이터를 분류한다. 결정트리의 규칙 집합은 IF ~ THEN 형태로 출력된다.

4. 경보데이터 마이닝 시스템 구현

실험 및 평가를 위해 구현된 프로타입은 그림 3과 같다. 사용자 인터페이스를 통해서 데이터베이스에 접속한다. 연관규칙, 빈발에피소드, 클러스터링 탭을 선택하여 각각의 마이닝 작업을 수행한다.

■ 연관 규칙 탐사

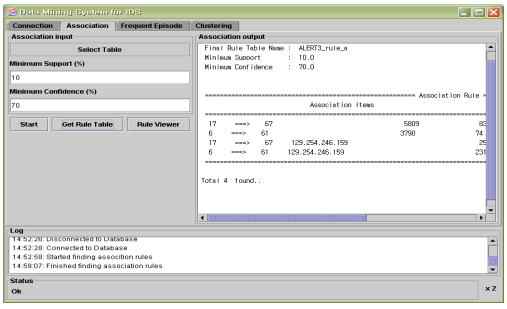
연관규칙을 탐사하기 위해 먼저 연관 규칙 탐사 탭을 선택한 다음 대상 테이블을 선택한다. 그러면 선택된 테이블리스트가 나오게 되며, 이 중에서 속성제약조건이 될 관심 있는 속성들을 선택한다.



[그림 3] 경보데이터마이닝 프레임워크 프로토타입

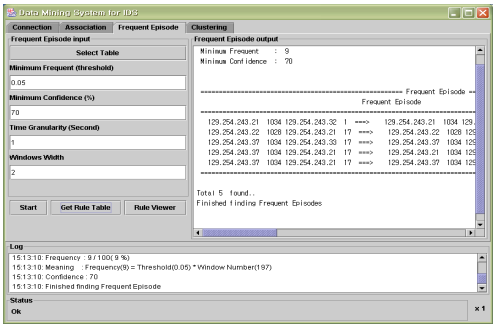
대상 테이블을 선택하고 나면 최소 지지도와 최소 신

뢰도에 대한 값을 입력한 후 시작 버튼을 눌러 연관규칙 탐사를 수행하며 수행이 완료되면 탐사된 규칙들을 볼 수 있다. 그림 4는 탐사된 연관규칙 결과를 보여준다.



[그림 4] 탐사된 연관 규칙 결과

빈발 에피소드를 탐사하기 위해 먼저 빈발 에피소드 탭을 선택한 다음 탐사하고자 하는 대상 테이블을 선택한다. 선택된 테이블리스트가 나오게 되며, 이중에서 속성제약 조건에 해당하는 관심 있는 속성들을 선택한다. 대상 테이블을 선택하고 나면 최소 빈발도와 최소 신뢰도 그리고 타임 단위와 윈도우 폭에 대한 값을 입력한 후 시작버튼을 누르면 로그 창에 수행되는 과정이 나타나게 된다. 수행이 완료되면 규칙 보여주기 버튼을 누르면 탐사된 규칙들을 사용자 인터페이스를 통해 확인할 수 있다. 그림 5에서 탐사된 빈발에피소드 규칙 결과를 볼 수 있다.

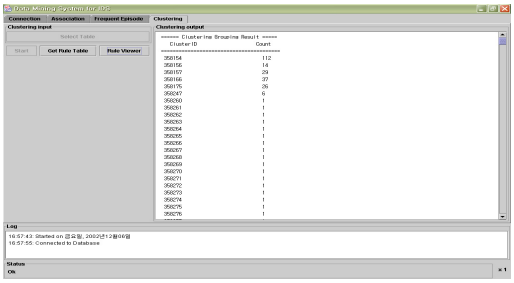


[그림 5] 탐사된 빈발 에피소드 결과

■ 순서기반 클러스터링

순서기반 클러스터링을 수행하기 위해 먼저 클러스터링 패널을 선택한 다음 경보데이터가 저장된 데이터베이스 테이블 선택버튼을 선택한다. 그러면 선택된 테이블리스트가 나오게 되며, 이중에서 관심 있는 속성들을 선택한다.

마이닝 할 테이블을 선택하고 나면 클러스터링을 수행하며 결과 패널에 탐사된 규칙들이 출력된다. 그림 6은 탐사된 클러스터링 결과이다.

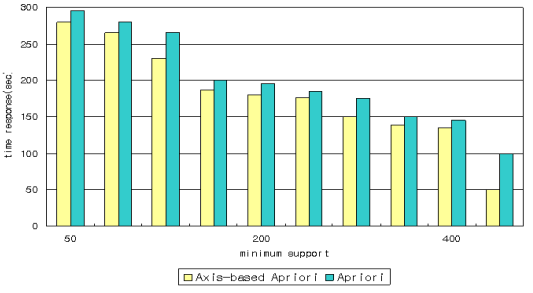


[그림 6] 클러스터링 결과

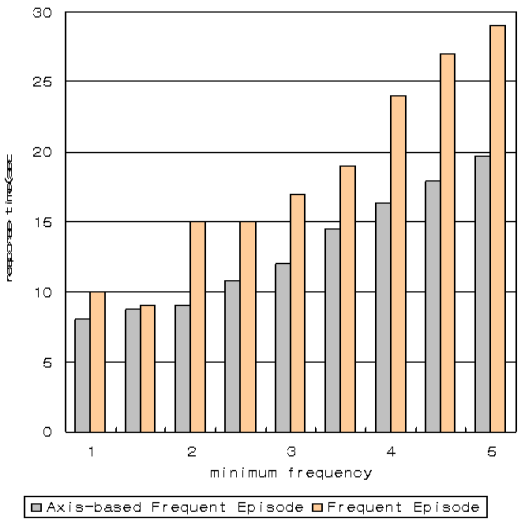
5. 실험 및 평가

이 논문에서 제안한 경보데이터 마이닝 프레임 워크를 위한 알고리즘들과 기존의 알고리즘을 비교 분석한다. 기존의 알고리즘은 연관규칙의 경우 Apriori와 비교를 하였으며 빈발에피소드의 경우 WinEPI 알고리즘과 비교 분석하였다.

이 실험의 제약 조건은 고려한 실험요소가 후보항목집합의 수만을 고려하여 최종 규칙 생성시까지의 시간을 비교 분석하였다. 최종 규칙 탐사시간은 데이터 항목의 구성이나 값 분포 등도 영향을 받을 수 있으나 실제로 후보항목집합이 가장 큰 영향을 미치는 요소이므로 다른 요소들은 실험에서 고려하지 않았다. 그림 7은 기존의 Apriori 알고리즘으로 경보데이터 속성간의 연관규칙을 탐사하였을 경우와 이 논문에서 제안한 프레임워크에 적용된 속성기반 연관규칙 알고리즘을 적용하였을 때의 탐사시간을 비교한 결과이다. 기존의 연관 규칙과 제안한 속성기반 연관규칙을 비교하였을 때 속성기반 연관규칙의 후보항목집합이 훨씬 적게 생성됨으로 연관 규칙 탐사에 필요한 실행시간이 훨씬 작다는 것을 알 수 있다.



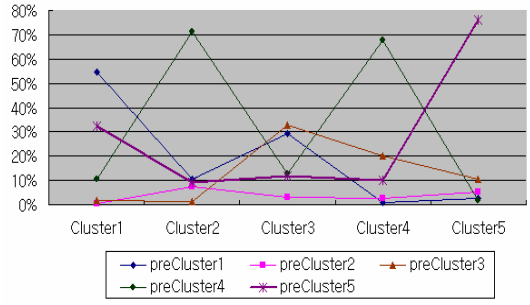
[그림 7] Apriori와 속성기반 연관규칙의 탐사시간



[그림 8] 속성기반 빈발에피소드의 탐사시간

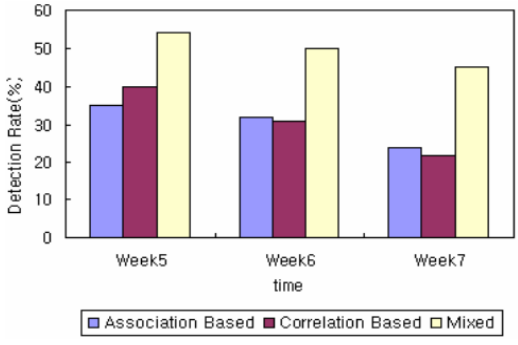
두 번째 실험인 기존의 빈발에피소드알고리즘과 이 논문에서 제안한 속성기반 빈발에피소드에 대한 비교 실험으로 제약조건은 빈발 후보항목집합만을 고려한 빈발에피소드 탐사시간을 측정하였다는 것이다. 그림 8은 기존의 WinEPI알고리즘과 제안된 속성기반 빈발에피소드의 비교 실험 결과이다. 기존의 알고리즘보다 빈발에피소드 탐사시간이 훨씬 작다는 것을 알 수 있다.

순서기반 클러스터의 실험은 생성된 클러스터에 대해 각 클러스터의 이전 클러스터를 정의하고 이를 기반으로 클러스터의 시퀀스를 생성할 수 있는지의 여부를 평가하기 위한 실험으로 실험 데이터는 KDDCup 1999 데이터 집합을 사용하였다. 이 실험은 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 제공할 수 있다는 것을 확인시켜 주었다. 그림 9의 그래프를 보면 cluster 4에 대한 이전클러스터를 보면 preCluster3의 비율이 가장 높다. 그러므로 Cluster4의 precluster는 Cluster3이 된다. 예를 들면, 새로운 경보가 들어오고 그 경보가 Cluster4에 분류된다면 이후 발생할 경보데이터는 Cluster2에 속하는 것이라는 것을 예측할 수 있게 해준다.



[그림 9] 클러스터링 결과 분석 그래프

오경보 분류 모델의 실험에서는 연관규칙기반의 속성선택과 통계적 상관분석 기법 기반 속성선택으로 분류모델의 노드를 결정한 후 두 개의 분류 모델을 생성하였다. 실험데이터는 1998년 DARPA 데이터를 사용하여 5주 100,000개, 6주 46,650개, 7주 55,150개의 데이터를 입력 값으로 분류하였으며 결과인 정상패킷의 탐지율은 상관분석 기반의 분류 모델이 29.79%, 연관규칙 기반의 분류 모델은 37.37%의 평균값을 보였다. 또한 탐지율을 높일 수 있는 방법으로 두 개의 분류모델을 결합한 모델을 사용하여 실험해 보았으며 그림 10의 세 번째 계열 Mixed는 두 개의 분류 모델을 결합한 후 실험한 결과인 혼합모델인 경우도 보여주고 있다. 혼합모델의 경우 평균 48.32%의 탐지율을 보였다.



[그림 10] 분류모델별 정상패킷 탐지율

6. 결론

이 논문에서는 경보 데이터 통합관리 및 경보상관관계 분석을 위한 경보 데이터마이닝 프레임워크를 제안하였다. 아울러 제안된 프레임워크의 프로토타입을 구현하여 제안된 마이닝기법 기반 경보데이터 통합관리 및 경보상

관관계분석에 적용하고 실험 평가하였다. 경보 데이터의 체계적인 통합 관리와 경보 상관관계 분석을 위해 경보 데이터들이 데이터베이스에 저장되어야하므로 전처리 과정을 통해 경보데이터를 저장하고 경보데이터 마이닝 태스크를 수행할 수 있도록 경보데이터 스키마를 설계하고, 데이터마이닝기법 기반 경보데이터 통합 관리 프레임워크 구축하였다. 기존의 데이터마이닝기법의 알고리즘을 경보데이터의 특성에 맞게 확장 설계하였다. 설계된 알고리즘은 경보데이터의 특성을 고려하여 관심있는 항목에 대한 지식탐사를 가능하게 하였으며 경보들중 오경보를 감소시키는 기능도 수행하는 장점을 가진다. 실험을 통하여 경보상관관계분석과 침입탐지시스템의 성능향상 결과를 평가하였다.

제안된 경보데이터 마이닝 프레임워크는 기존의 경보데이터 상관관계분석에서는 해결하지 못했던 통합적인 경보 상관관계 분석 기능을 수행할 뿐만 아니라 대량의 경보데이터에 대한 필터링을 수행하는 장점을 가진다. 즉 침입탐지시스템의 앞부분 혹은 뒷부분에서 오경보를 감소시키는 오경보 분류모델은 침입탐지시스템의 오경보율을 감소시키고 침입 탐지율을 향상시키는 역할을 수행한다는 것을 확인할 수 있었다. 추출된 규칙 및 공격시나리오는 침입탐지시스템의 실시간 대응에 활용될 수 있다.

향후 침입탐지시스템의 시그네쳐나 규칙에 추가할 수 있도록 하는 에이젼트 개발에 대한 연구가 계속되어야할 것이다.

참고문헌

[1] Moon Sun Shin, HoSung Moon, KeunHo Ryu, "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, 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] S. Staniford, J.A. Hoagland, and J.M. McAlerney, "Practical automated detection of stealthy portscans", 2000

[7] Moon Sun Shin, EunHee Kim, Keun Ho Ryu, “ False Alarm Classification Model for Network-based Intrusion Detection System”, IDEAL2004, LNCS, SpringerVerlag,

[8] 신문선, 류근호, “침입탐지시스템의 성능향상을 위한 오경보 분류 모델 구현”, 정보과학회논문지:데이터베이스 2007.

신 문 선(Moon-Sun Shin) **[정회원]**

- 2004년 8월 : 충북대학교대학원 전자계산학과(이학박사)
- 2005년 8월 ~ 2009년 2월 : 건국대학교 컴퓨터응용과학부 강의교수
- 2010년 3월 ~ 현재 : 안양대학교 교양학부 조교수

<관심분야>
데이터베이스, 데이터마이닝, 침입탐지시스템, 정보보안, RFID 보안