네트워크 공격 분석을 위한 마이닝 프로토타입 시스템 구현

김 은 희[†] · 신 문 선[††] · 류 근 호[†††]

요 약

네트워크 공격은 인터넷의 발달과 함께 유형도 다양하고 새로워지고 있다. 기존의 침입탐지 시스템들은 알려진 공격의 시그네처를 기반으로 탐지하기 때문에 알려지지 않거나 변형된 공격을 탐지하고, 대응하기 위해서는 많은 노력과 비용이 필요하다. 본 논문에서는 네트워크 프로토콜 속성 분석을 통해 알려지지 않거나 변형된 네트워크 공격을 예측할 수 있는 마이닝 프로토타입 시스템을 설계 하고 구현 하였다. 네트워크 프로토콜 속성을 분석하기 위해서 연관규칙과 빈발에피소드 기법을 사용하였으며, 수집된 네트워크 프로토콜은 TCP, UDP, ICMP와 통합된 형태의 스키마로 저장한다. 본 실험을 통해서 각 프로토콜별로 발생 가능한 네트워크 공격 유형을 예측할 수 있는 규칙들을 생성한다. 마이닝 프로토타입은 침입탐지 시스템에서 새로운 공격에 대응하기 위한 보조적인 도구로서 유용하게 사용될 수 있다.

An Implementation of Mining Prototype System for Network Attack Analysis

Eun Hee Kim[†] · Moon Sun Shin[††] · Keun Ho Ryu[†††]

ABSTRACT

Network attacks are various types with development of internet and are a new types. The existing intrusion detection systems need a lot of efforts and costs in order to detect and respond to unknown or modified attacks because of detection based on signatures of known attacks. In this paper, we present a design and implementation for mining prototype system to predict unknown or modified attacks through network protocol attributes analysis. In order to analyze attributes of network protocols, we use the association rule and the frequent episode. The collected network protocols are storing schema of TCP, UDP, ICMP and integrated type. We are generating rules that can predict the types of network attacks. Our mining prototype in the intrusion detection system aspect is useful for response against new attacks as extra tool.

키워드 : 네트워크 프로토콜(Network Protocol), 네트워크 공격(Network Attack), 데이터 마이닝(Data Mining), 연관규칙(Association Rule), 빈발 에피소드(Frequent Episode)

1. 서 론

네트워크 공격은 인터넷 발달과 함께 그 유형도 새롭고 다양해지고 있다. 기존의 공격 유형들은 방화벽이나 침입탐지 시스템과 같은 보안 시스템에서 탐지를 하지만 이들 보안 시스템에서는 알려진 공격유형들에 대해서만 탐지를 하기 때문에 새로운 공격을 탐지하고, 대응하기 위해서는 많은 노력과 비용이 요구된다. 특히, 인터넷 애플리케이션의 빠른 발전은 지속적인 새로운 공격을 만들어 내고 있으며, 침입탐지시스템이 이를 대응하는 데는 한계가 있을 수밖에 없다. 그래서 최근에 침입탐지시스템의 성능향상을 위한 연구들이 활발히 진행되고 있다.

※ 이 연구는 한국전자통신연구원의 정보보호 연구단 및 한국 과학재단 RRC (청주대 ICRC)의 연구비 지원으로 수행되었음.
† 준 회 원 : 충북대학교 대학원 전자계산학과
†† 정 회 원 : 충북대학교 대학원 전자계산학과
††† 종신회원 : 충북대학교 전기전자및컴퓨터공학부 교수
논문접수 : 2004년 3월 18일, 심사완료 : 2004년 6월 18일

침입탐지 시스템에 관한 연구로는 자동화된 침입 탐지 모델 구축[1-4], 감사데이터 분석[1-3], 거짓경보 감소[5]에 관한 연구 그리고 침입탐지 시스템의 경보데이터 상관관계 분석[6]등과 관련하여 많은 연구들이 이루어지고 있다. 초기에는 침입 탐지 시스템의 효율을 증가시키기 위해 기계학습[7, 8] 등을 적용한 연구로 시작을 했으나 최근에는 증가하는 데이터 크기와 고차원의 데이터 특성 때문에 데이터 마이닝 기법을 적용한 연구가 많이 진행되고 있다[1-4, 9, 10]. 침입탐지 시스템은 보안을 위한 필수 보안도구 이기는 하지만 새로운 공격에 대응하기 위해서는 다른 부가적인 도구를 필요로 한다. 이러한 도구는 네트워크 또는 시스템의 비정상적인 사용을 일반 관리자가 쉽게 알아 낼 수 있도록 도와주는 형태의 솔루션이 되어야 할 것이다. 또한 관리자가 침입 탐지 시스템 또는 네트워크 기록을 정확히 해석하고 적절한 조치를 취할 수 있도록 도와줄 수 있는 기능이 있어야 한다. 하지만 아직까지 이러한 기능을 제공하는 솔루션은 존재하

지 않는다.

이 논문에서는 네트워크 프로토콜 분석을 통하여 네트워크 공격을 예측할 수 있는 마이닝 프로토타입 시스템을 설계 및 구현한다. 프로토타입 시스템은 두 부분으로 구성되어 있다. 첫 번째는 전 처리 프로세서 모듈로서 수집된 네트워크 프로토콜을 각 프로토콜별로 저장하는 역할을 수행한다. 두 번째는 마이닝 엔진 부분으로서 각 프로토콜별로 저장된 데이터를 분석하여 네트워크 공격을 예측하는 역할을 수행한다. 프로토타입에서 마이닝 엔진 부분에 사용된 기법은 연관규칙과 빈발에피소드 프로그램으로 구성되어 있다. 연관규칙을 통해서 프로토콜 속성간의 연관성을 분석하고, 빈발에피소드를 통해서 프로토콜 속성간의 시퀀스를 분석한다. 또한 실험 결과를 통해서 새로운 네트워크 공격 패턴과 유형을 예측하여 보안 관리자가 새로운 공격에 대해서 적절한 조치를 취할 수 있도록 도와줄 것이다. 또한 침입탐지 시스템의 보조적인 도구로서의 역할에도 기여하게 될 것이다.

논문의 구성은 다음과 같다. 2장에서는 침입탐지 시스템에 대한 기존 연구들에 대해서 살펴본다. 3장에서는 네트워크 프로토콜을 분석하여 네트워크 공격을 예측할 수 있는 마이닝 프로토타입 시스템을 설계한다. 4장에서는 실험을 통해서 각 프로토콜별로 발생되는 네트워크 공격을 분석한다. 마지막으로 5장에서는 결론과 향후연구로 끝을 맺는다.

2. 관련 연구

지금까지 침입탐지 시스템의 성능향상을 위한 많은 연구들이 수행되어 왔다. Bro[11], EMERALD [12] 시스템은 대표적인 네트워크 기반의 침입탐지모델로서 침입탐지의 주요 설계 목적을 확장하여 만든 것이다. Bro는 명백한 침입 탐지 규칙들을 요약화 하기 위한 고 수준의 스크립트 언어를 제공하며 특정 사이트 모델을 자동적으로 계산하고 설치할 수 있는 프레임워크를 제공한다. EMERALD는 침입 탐지자의 환경설정과 폭넓은 기업형 개발을 손쉽게 할 수 있는 아키텍처를 제공했다. 이런 침입 탐지에서 학습 처리를 자동화하기 위해서 학습기반 메커니즘을 사용했다. 학습 기반 방법들은 [13]에서 규칙 감소를 위해 머신 러닝 접근법을 사용하고 있다. [14]에서는 실행시간동안 시스템 콜의 시퀀스를 식별하여 구성하는 학습기법을 사용하고 있다. 학습된 모델들은 이상 탐지를 정확하게 식별할 수 있다. [15]에서는 감사 데이터로부터 프로그램 행위 데이터를 사용하여 탐지를 하였고, 신경망에서는 시스템 프로그램을 위한 이상 탐지 모델과 오용 탐지 모델을 학습하는데 사용할 수 있다. [16]에서는 사용자 쉘 명령어와 이상 탐지를 분석하기 위한 알고리즘을 개발하였다. 이러한 알고리즘들은 메타러닝 기법을 적용하여 침입 탐지 모델을 학습하였다. Wenke는 1998년도에 자동화된 침입 탐지 모델을 구축하기 위해서 침입탐지 모델에 데이터 마이닝 기법을 적용하기 시작 하였다. 대표적인 시스템으로 MADAMID(Mining Audit Data for Automated Models for Intrusion Detection)이 있다[3]. MADAMID는 이상 탐지를 위해 개발된 규칙들에 대해 데이터 마이닝 기법을 적용한 것이다. 현재 시스템들이 이상 탐지를 위한 광범위한 개발 규칙들을 수동적으로 보고하는 것을 요구하기 때문이다. MADAMID는 침입 행위의 패턴들과 정상적인 행위를 정확하게 획득하여 모델을 계산하기 위해서 감사 데이터에 데이터 마이닝 기법을 적용한 것이다. MADAMID는 학습 분류자와 메타 분류자, 연관규칙, 빈발 에피소드 프로그램으로 구성되어 있다. 지금까지 살펴본 바와 같이 침입탐지의 성능향상을 위한 많은 연구들이 있었다. 하지만 침입탐지 시스템이나 네트워크 모니터링 기록을 정확히 해석하고 적절한 조치를 취할 수 있도록 도와줄 수 있는 기능이 미흡하다.

본 논문에서는 네트워크 프로토콜을 분석하여 네트워크 공격을 분석할 수 있는 마이닝 프로토타입을 제안한다. 제안된 프로토타입에서는 마이닝 기법을 사용하여 네트워크 프로토콜을 분석한다.

데이터 마이닝 기법[16]은 데이터로부터 이전에 잘 알려지지는 않았지만, 묵시적이고 잠재적으로 유용한 지식을 추출하는 기술로 정의되며 전자상거래, 의사결정 지원, 의료 등의 다양한 분야에서 유용하게 활용 될 수 있다. 데이터 마이닝에 대한 기법으로는 연관 규칙, 순차 패턴(시퀀스 마이닝), 분류, 클러스터링 등 다양한 기법들이 연구되고 있다. 이러한 기법 들 중에서 우리는 연관규칙과 순차 패턴(시퀀스 마이닝) 기법을 사용한다.

연관규칙[19]는 어떤 항목 집합(I)이 주어지고, 데이터베이스(D)는 트랜잭션(T)의 집합이고, 각 트랜잭션(T)은 항목집합(I)의 부분집합으로 정의 될 때, 전체 트랜잭션 중에서 빈발하게 발생하는 항목을 찾아내는 기법이다.

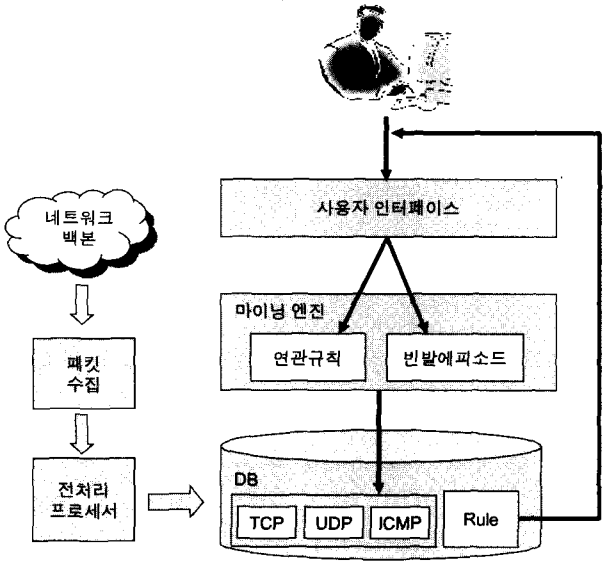
시퀀스 마이닝[18]은 한 트랜잭션 안에서 발생하는 항목들 간의 연관 규칙에 시간의 변이를 추가한 것이다. 연관 규칙은 트랜잭션 안에서 어떤 항목을 함께 사는가 하는 문제로 트랜잭션 내의 문제인 반면 시퀀스를 발견하는 것은 트랜잭션 상호간의 문제인 것이다.

3. 마이닝 프로토타입 설계

이 장에서는 수집된 네트워크 프로토콜을 처리하여 네트워크 공격을 예측할 수 있는 마이닝 프로토타입 시스템 설계에 대해서 기술한다. 네트워크 공격 분석을 위한 마이닝 프로토타입 시스템은 크게 두 부분으로 구성되어 있다. 수집된 네트워크 프로토콜을 처리하는 전처리 부분과 처리된 프로토콜을 분석하는 마이닝 엔진 부분으로 되어 있다.

3.1 시스템 구조

전체 시스템에 대한 구조는 아래 (그림 1)에서 묘사한 것과 같다. 수집된 네트워크 프로토콜은 전처리 과정을 거쳐 네트워크 공격을 예측하기 위해서 마이닝을 수행하게 된다. 각 단계별로 수행되는 과정에 대해서 다음절에서 자세히 기술한다.



(그림 1) 네트워크 공격 분석을 위한 마이닝 프로토타입 구조

3.2 전 처리 프로세서

tcpdump 유틸리티를 사용하여 네트워크 프로토콜을 수집한다. 수집된 네트워크 프로토콜은 원시 데이터이기 때문에 네트워크 분석을 위해서는 먼저 가능한 형태인 ASCII 형태로 변환을 해야 한다. 변환된 패킷 데이터는 타임 스탬핑(패킷 도착 시간)에 의해 정렬되어 있으며 네트워크 접속 정보를 포함하고 있다. 네트워크 프로토콜에 대한 전 처리 프로세서는 tcpdump 유틸리티에 의해 만들어진 원시 패킷 파일을 입력으로 하여, 시간 정보 및 패킷 정보를 전달한다. 이때 패킷 구조체 포맷으로 전달하며, 전달된 값을 데이터베이스로 저장하여 처리하게 된다. 또한 프로토콜별로 발생할 수 있는 공격을 분석하기 위해서 패킷 구조체로부터 전달된 속성들은 데이터베이스에 각 프로토콜별로 저장한다. 프로토콜 저장을 위한 데이터베이스의 스키마는 아래 <표 1>~<표 4>와 같으며 데이터를 체계적으로 관리하기 위해서 관계형 데이터베이스를 사용하였다. <표 1>은 TCP 프로토콜에 대한 속성을 추출하여 생성한 스키마이다. <표 2>는 UDP 프로토콜에 대한 속성을 추출하여 생성한 스키마이다. <표 3>은 ICMP 프로토콜에 대한 속성을 추출하여 생성한 스키마이다. 이때 IP에 대한 정보는 이들 프로토콜이 모두 공통적으로 가지고 있다. 프로토콜에 대한 속성의 데이터 타입은 모두 varchar2 타입으로 정의하였다. 네트워크 접속 정보에 대한 분석을 하기 위한 통합 스키마를 <표 4>와 같이 설계하였다.

〈표 1〉 TCP 데이터베이스 스키마

속성명	데이터타입	속성명	데이터타입
Ip_src	varchar 2	Tcp_sport	varchar 2
Ip_dst	varchar 2	Tcp_dport	varchar 2
Ip_ver	varchar 2	Tcp_seq	varchar 2
Ip_hlen	varchar 2	Tcp_ack	varchar 2
Ip_tos	varchar 2	Tcp_off	varchar 2
Ip_len	varchar 2	Tcp_urp	varchar 2
Ip_id	varchar 2	Tcp_res	varchar 2
Ip_flags	varchar 2	Tcp_flags	varchar 2
Ip_off	varchar 2	Tcp_win	varchar 2
Ip_ttl	varchar 2	Tcp_csum	varchar 2
Ip_proto	varchar 2	Ip_csum	varchar 2

〈표 2〉 UDP 데이터베이스 스키마

속성명	데이터타입	속성명	데이터타입
Ip_src	varchar 2	Ip_off	varchar 2
Ip_dst	varchar 2	Ip_ttl	varchar 2
Ip_ver	varchar 2	Ip_proto	varchar 2
Ip_hlen	varchar 2	Ip_csum	varchar 2
Ip_tos	varchar 2	Udp_sport	varchar 2
Ip_len	varchar 2	Udp_dport	varchar 2
Ip_id	varchar 2	Udp_len	varchar 2
Ip_flags	varchar 2	Udp_chk	varchar 2

〈표 3〉 ICMP 데이터베이스 스키마

속성명	데이터타입	속성명	데이터타입
Ip_src	varchar 2	Ip_ttl	varchar 2
Ip_dst	varchar 2	Ip_proto	varchar 2
Ip_ver	varchar 2	Ip_csum	varchar 2
Ip_hlen	varchar 2	Icmp_type	varchar 2
Ip_tos	varchar 2	Icmp_code	varchar 2
Ip_len	varchar 2	Icmp_id	varchar 2
Ip_id	varchar 2	Icmp_seq	varchar 2
Ip_flags	varchar 2	Icmp_csum	varchar 2
Ip_off	varchar 2		

〈표 4〉 통합된 프로토콜 스키마

속성명	데이터타입	길이	설 명
no	number	2	세션 식별번호
start date	date		세션을 위해 시작된 날짜
start time	date		세션을 위해 시작된 시간
duration	varchar 2	15	세션을 위한 존속 기간
service	varchar 2	15	세션에 의해 사용된 서비스 이름
src_port	varchar 2	15	세션을 위한 근원지 포트
dst_port	varchar 2	15	세션을 위한 목적지 포트
src_ip	varchar 2	15	세션을 위한 근원지 주소
dst_ip	varchar 2	15	세션을 위한 목적지 주소
score	varchar 2	15	세션에 할당된 공격 회수
name	varchar 2	15	세션에 사용된 공격 이름

3.3 마이닝 엔진

앞 절의 전 처리 프로세서를 통해서 데이터베이스에 저장된 데이터는 마이닝 엔진을 통해서 공격 패턴들을 규칙 형태로 생성한다. 마이닝 엔진은 연관규칙과 빈발에피소드로 구성되어 있다.

3.3.1 연관규칙

네트워크 프로토콜 속성간의 연관성 탐사를 위해서 연관규칙 알고리즘들 중 Apriori 알고리즘을 기반으로 설계하였다. 또한 네트워크 프로토콜의 특성 상 일반적인 트랜잭션 데이터베이스와는 다르기 때문에, 기존의 Apriori 알고리즘에 키 속성 개념을 적용함으로서 불필요한 규칙이 많이 생성되는 문제를 해결하였다. 키 속성개념은 규칙을 탐사하는 데 있어서 관심 있는 속성을 포함한 규칙만을 탐사하는 것을 말하며, 이것은 사용자가 임의로 선택할 수 있도록 하였다. 연관규칙 알고리즘은 아래 (알고리즘 1)에서 보여주고 있다.

```
AlgorithmAxis-basedAssociationRule
Input
 Data Set : A set of each protocol, integrated protocol data
 Threshold : min_supp, min_conf (minimum support, minimum
             confidence)
 Axis_attr : A set of added attribute
Method
  Relation : each protocol Table
  ID : protocol ID
  ItemList : selected attribute(s) + [Axis attr]
  Call Association( )
      Call GenerateAssociationRule (Relation, ID, ItemList, min_supp,
      min_conf)
          While (item supp >= min_supp) {
               Create Cand_set (1 <= i <= n-1)
               For each (cand item supp >= min_supp)
                  Create Large_set (1 <= j <= m-1)
                   For each (large item supp >= min_supp)
                      Create Rule_set (1 <= k <= l-1)
           }
           Create Final_rule set
           Return Final_rule set ;
Output
  Final_rule set : association rule set
```

(알고리즘 1) 연관규칙 알고리즘

연관규칙은 분석하고자 하는 프로토콜 데이터 셋과 임계치로서 미리 정의된 최소 지지도와 최소 신뢰도를 입력으로 받는다. 또한 사용자가 관심 있는 속성만을 고려하기 위해서 키 속성 개념을 추가할 수 있다. 키 속성 개념으로 추가한 속성은 규칙 생성 시 반드시 포함시키는 것으로 연관규칙을 통해 생성되는 후보항목의 수를 줄일 수 있다. 키 속성 개념은 사용자의 선택사항이다. 입력된 속성들은 속성리스트에 적재되어 최소 지지도와 최소 신뢰도를 만족하는 빈발 항목들을 찾아서 규칙을 생성한다. 기본적으로 규칙을 탐사하는 과정은 Apriori와 유사하며, 탐사된 규칙들의 예는 아래와 같이 생성된다.

- 예 1) *src_host* = 192.108.001.010 ==> *service* = *domain/u*, *dst_ host* = 192.108.001.020 [511, 83]

이 규칙의 의미는 src_host(근원지 호스트)속성 값이 192.108.001.01이고 dst_host(목적지 호스트) 속성값이 192.108.001.020이며, 서비스가 domain/u인 경우가 전체 511회이며, src_host(근원지 호스트) 192.108.001.01에서 dst_host(목적지 호스트) 192.108.001.020로 domain/u 서비스를 수행하는 것이 전체 83% 나타났다는 것을 의미한다.

3.3.2 빈발에피소드

빈발에피소드는 네트워크 프로토콜 속성간의 시퀀스 분석을 통해서 공격 패턴을 찾아내기 위해 사용되었다. 빈발에피소드는 주어진 시간범위 내에서 빈발하게 발생하는 시퀀스를 찾는다. 빈발에피소드에서도 연관규칙과 마찬가지로 키 속성 개념을 적용한다. 빈발에피소드에 대한 간단한 알고리즘은 아래 (알고리즘 2)에서 보여주고 있다.

```
Algorithm Axis-based Frequent Episode
Input
 Data Set : A set of each protocol, integrated protocol data
 Threshold : min_supp, min_conf
 Window slide : time_granularity, window_width
 Axis_attr : A set of added attribute
Method
  Relation : each protocol Table
  ID : ProtocolID
  ItemList : selected attribute(s) + [Axis attr]
  Call FrequentEpisode( )
      Call GenerateFrequentEpisode (Relation, ID, ItemList, min_supp,
                                    min_conf, time_granularity,
                                    window_width)
          Create window_set in order to sort protocol data according
          to window_slide
          While (window_set item supp >= min_supp) {
              Create Cand_set (1 <= i <= n-1)
              For each (cand item supp >= min_supp)
                 Create Large_set (1 <= j <= m-1)
                  For each (large item supp >= min_supp)
                     Create Rule_set (1 <= k <= l-1)
          }
          Create Final_rule set
          Return Final_rule set ;
Output
  Final_rule set : Frequent Episode Rule set
```

(알고리즘 2) 빈발에피소드 알고리즘

빈발에피소드에서는 분석하고자 하는 프로토콜 데이터와 임계치 값으로 미리 정의된 최소 지지도와 최소 신뢰도를 입력한다. 그리고 윈도우 폭을 결정하기 위해 시간 단위를 입력을 한다. 기본 시간 단위는 초로 설정하였다. 빈발에피소드에서는 윈도우 단위로 시퀀스 패턴을 분석한다. 그래서 시간단위와 윈도우 폭을 곱하여 시퀀스 윈도우 폭를 계산한다. 계산된 윈도우 폭을 가지고 시퀀스 별로 재 정렬을 한 후 그 안에서 후보 시퀀스와 빈발 시퀀스를 탐사하여 규칙을 생성한다. 또한 사용자가 관심 있는 속성을 추가한다. 기

본적으로 규칙을 탐사하는 과정은 기존의 빈발에피소드와 유사하며, 탐사된 규칙들의 예는 아래와 같이 생성된다.

- 예 2) (*src_host* = 203.255.74.102, *dst_host* = 210.155.167.10, *src_port* = 9156, *dst_port* = 21, *service* = *tcp*) ==> (*src_host* = 203.255.71.102, *dst_host* = 210.155.167.10, *src_port* = 9156, *dst_port* = 21, *service* = *tcp*) [10, 60, 10]

위 규칙의 의미는 10초 동안 "(src_host = 203.255.74.102, dst_host = 210.155.167.10, src_port = 9156, dst_port = 21, service = tcp) ==> (src_host = 203.255.71.102, dst_host = 210.155.167.10, src_port = 9156, dst_port = 21, service = tcp)" 이런 형태의 시퀀스 패턴이 전체 10회 이상 발생하였고, (src_host = 203.255.74.102, dst_host = 210.155.167.10, src_port = 9156, dst_port = 21, service = tcp) ==> (src_host = 203.255.71.102, dst_host = 210.155.167.10, src_port = 9156, dst_port = 21, service = tcp) 시퀀스가 전체 60%이상 나타났다는 것을 의미한다. 다음 절에서 마이닝 프로토타입의 실험 평가를 통해서 검증한다.

4. 실험 및 결과분석

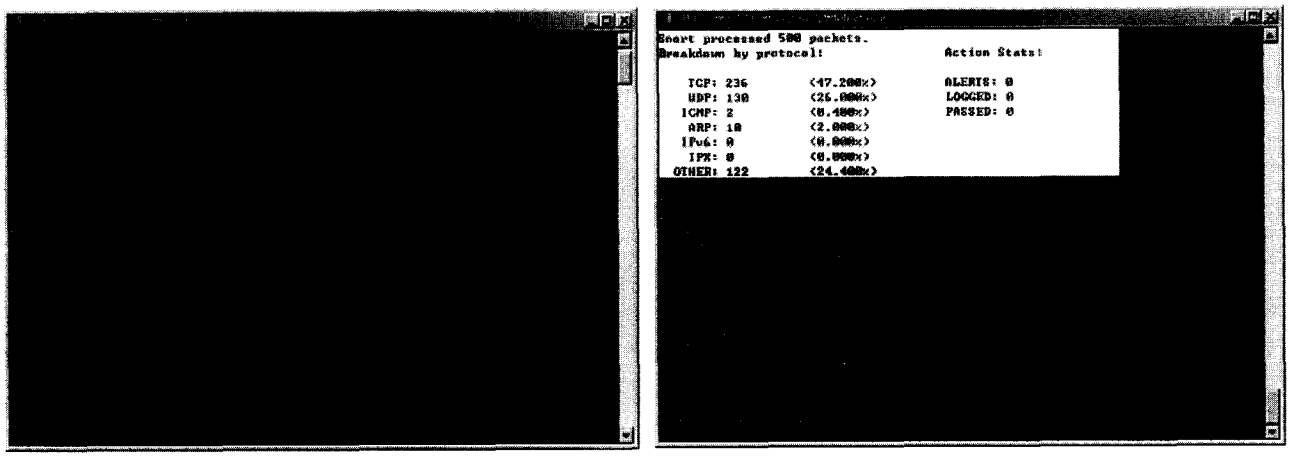
4.1 실험 환경

네트워크 공격 분석을 위한 마이닝 프로토타입 시스템을 위한 구현 환경은 OS로 Linux7.1과 DBMS로는 오라클 8.1.7을 사용하였다. 사용한 언어는 Java를 사용하여 구현하였으며 데이터베이스와 연동을 위해 JDBC 드라이버를 사용하였다. 실험을 위해 사용된 시스템은 Pentinum PC 1.3GHz 256Mbyte이며, DBMS로서 오라클 8i를 사용하였고, 오라클과의 연동을 위해 JDBC드라이버를 사용하였다. 그리고 실험에 사용된 데이터 셋은 실제 TCPDUMP 유틸리티를 통해 수집된 네트워크 패킷데이터와 DARPA 평가에 사용된 데이터를 가지고 실험하였다.

4.2 실험 결과 및 분석

4.2.1 전 처리 프로세서

전 처리 프로세서는 수집된 네트워크 패킷 데이터 500개



(그림 2) 전 처리 프로세서 수행 결과

를 수행 시켰다. (그림 2)에서는 전 처리 프로세서의 수행결과로서 각 프로토콜별로 TCP 236개, UDP 130개, ICMP 2개로 각각 분류되었음을 확인했다. 프로토콜별로 분류된 패킷 헤더 정보들은 각 프로토콜에 해당하는 IP를 가지고 해당 프로토콜 테이블로 저장된다.

4.2.2 네트워크 프로토콜 속성간의 연관성 분석

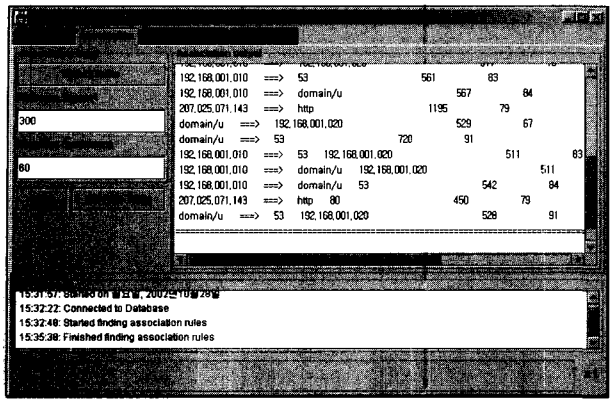
이 실험은 연관규칙을 이용하여 속성간의 연관성을 분석한다. 실험 데이터는 DARPA 데이터를 이용하였고, 프로토콜 속성으로서 근원지, 목적지 주소, 서비스, 근원지 포트, 목적지 포트 속성을 선택하였다. <표 5>는 속성간의 연관성 분석을 위한 샘플 데이터이다. 규칙 생성을 위한 최소 지지도와 신뢰도에 대한 값은 각각 300회, 60% 이상으로 설정하였다. 연관 규칙 탐사 결과는 (그림 3)과 같다.

<표 5> 연관성 분석을 위한 샘플 데이터(tcpdumplist)

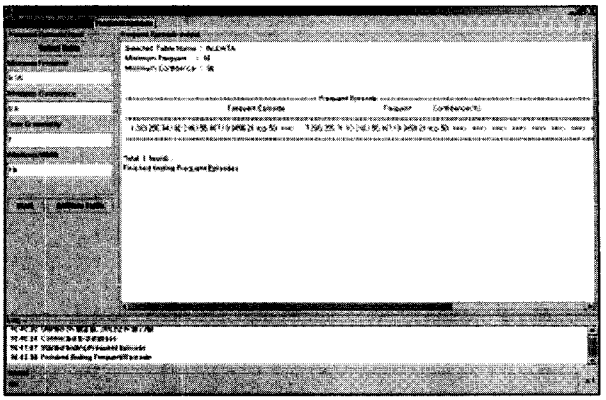
No	saddr	daddr	service	sport	dport
1	172.016.114.207	152.163.214.011	http	2127	80
2	172.016.114.207	152.163.212.172	http	2138	80
3	172.016.114.207	152.163.214.011	http	2128	80
4	172.016.114.207	152.163.214.011	http	2129	80
5	172.016.114.207	152.163.214.011	http	2130	80
:	:	:	:	:	:

위 실험에서 탐사된 연관규칙들 중에서 예를 들어보면 <표 6>과 같다.

<표 6>에서처럼 탐사된 예를 통해서 192.108.001.010 주소에는 53번 포트를 이용하여 대부분이 domain/u 서비스를 이용하고 있으며, 207.025.071.143 주소에서는 주소 http 서비스를 이용한다고 예측할 수 있다.



(그림 3) 연관성 분석 결과



(그림 4) 프로토콜 시퀀스 패턴 분석 결과

〈표 6〉 DARPA데이터에서의 연관성 분석 결과

탐사된 규칙	192.108.001.010 ==> domain/u 192.108.001.020 [511, 83]
의미	192.108.001.020 주소에서 목적지 주소 192.108.001.010 로 접속하여 domain/u 서비스를 수행하는 패턴이 전체 83% 발생했다.

〈표 8〉 프로토콜 시퀀스 패턴의 규칙

탐사된 규칙	203.255.74.102 210.155.167.10 9156 21 tcp ==> 203.255.71.102 210.155.167.10 9156 21 tcp [10, 60, 10]
의미	203.255.74.102 IP에서 목적지 IP 210.155.167.10로 9156번 포트에서 21번 포트로 접근해서 tcp서비스를 수행하는 패턴이 10초 이내에 발생하는 것이 전체 60% 나타났다.

4.2.3 네트워크 프로토콜 시퀀스 분석

프로토콜 시퀀스 분석에서는 목적지 주소에 대해서 언제, 어떤 주소에서 접속하여 어떠한 프로토콜을 수행하는 패턴들이 빈발하게 발생하는 지를 탐사하기 위해서 근원지 주소, 목적지 주소, 근원지 포트, 목적지 포트, 프로토콜 속성을 선택하여 빈발 에피소드를 탐사한다. 프로토콜 시퀀스 분석을 위한 샘플 데이터로서 실제 tcpdump 프로그램을 이용하여 학내의 네트워크 프로토콜을 수집한 것으로 <표 7>과 같다. 단, 샘플 데이터의 개수는 윈도우 테이블 수가 많아지기 때문에 일부만을 추출하여 실험을 한다. 프로토콜의 시퀀스를 탐사하기 위해서 초기 임계치로서 최소 지지도, 최소 신뢰도, 시간 단위, 윈도우 폭은 각각 35%, 60%, 1초, 10초로 설정을 하였다.

〈표 7〉 프로토콜 시퀀스 패턴 분석을 위한 샘플 데이터

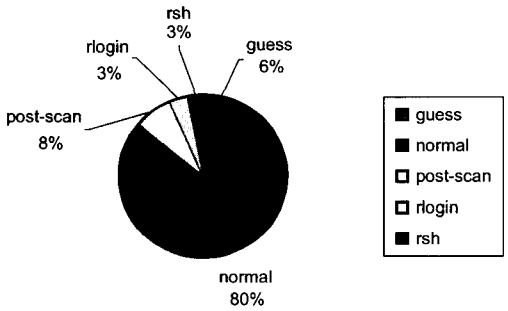
No	saddr	daddr	sport	dport	proto
1	203.255.71.10	210.155.167.10	9158	21	tcp
2	203.255.71.11	210.155.167.10	9159	21	tcp
3	203.255.71.12	210.155.167.10	9160	21	tcp
4	203.255.71.20	210.155.167.10	9161	21	tcp
5	203.255.72.10	210.155.167.10	9162	21	tcp
:	:	:	:	:	:

위 실험을 통해서 발견된 프로토콜 시퀀스 패턴의 규칙을 보면 <표 8>에서처럼 나타났다.

<표 8>에서처럼 탐사된 프로토콜 시퀀스 패턴에서는 203.255.74.102 IP에서 21번 포트로 접근을 가장 많이 시도하므로, TCP와 관련된 SYN 공격등을 예측할 수 있다.

4.2.4 네트워크 공격 예측 율

이번 실험은 DARPA 데이터를 가지고 네트워크 프로토콜 속성간의 연관성 분석과 프로토콜 시퀀스 패턴 분석을 통해서 네트워크 공격의 예측 율을 평가한다. (그림 5)는 실험 결과를 보여준다. 실험에 사용한 데이터는 각 세션별로 거의 공격이 없었고, 공격이 있다고 하더라고 희박하기 데이터이기 때문에 공격 회수가 1회 이상이면 해당 공격이라고 가정한다. 실험 결과를 통해서 보듯이 80% 이상이 normal 상태였고, 그 다음으로 port-scan > guess > rlogin > rsh 순으로 공격이 예측되었다.

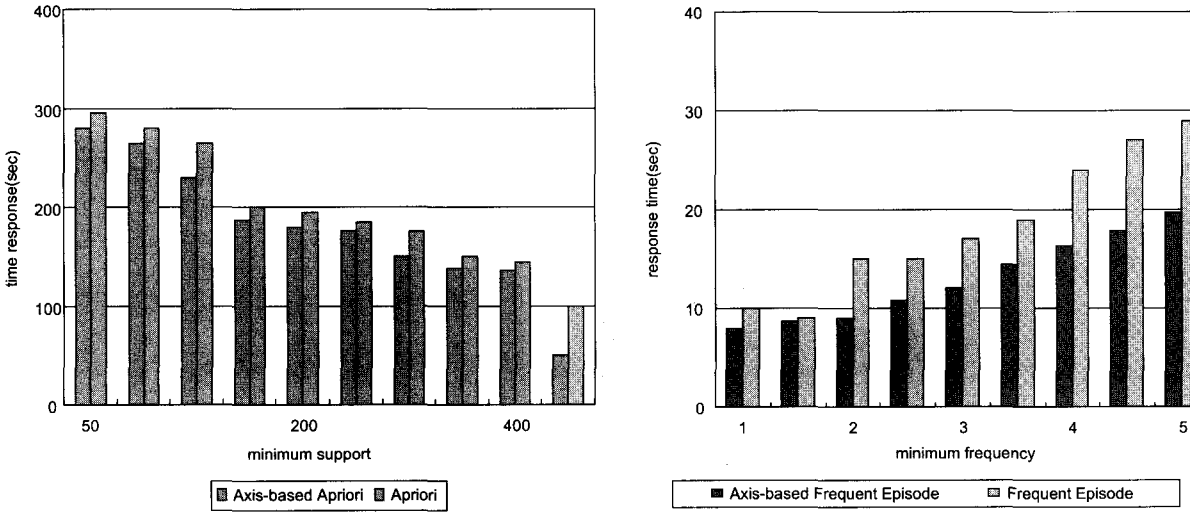


(그림 5) 탐지된 공격 분포도

4.2.5 성능 평가

성능평가에서는 데이터 크기에 따른 응답 시간을 비교해 본 것으로서 기존의 알고리즘과 키 속성 개념을 적용한 알고리즘간의 응답시간을 평가 하였다. (그림 6)은 데이터 크기 변화에 따른 응답시간에 대한 결과를 보여준다. 특히 마이닝을 수행하는 데 있어 초기 임계치인 최소 지지도와 신

뢰도는 분석가의 주관에 따라서 값을 조절할 수 있다. 그 이유는 지지도와 신뢰도를 조절해 가면서 수행한 결과 어느 정도의 값에서 의사 결정을 위해 가장 좋은 규칙들이 생성이 되었는지를 결정하기 때문이다. 우리는 성능 평가를 위하여 DARPA 데이터의 크기에 따라 최소 지지도를 변경하며 결과를 확인하였다. 데이터의 크기는 1,000개, 10,000개, 30,000개, 50,000개, 70,000개씩 증가시켰고, 최소지지도는 10%, 15%, 20%, 30%씩 증가시키면서 성능평가를 수행하였다. 실험 결과에서도 보듯이 데이터 크기에 비해 수행되는 시간이 적절함을 보여주고 있다.



(그림 6) (a) 기존의 연관규칙과 키 속성 개념을 적용한 연관규칙간의 응답시간 비교
(b) 기존의 빈발에피소드와 키 속성 개념을 적용한 빈발에피소드간의 응답시간 비교

5. 결 론

본 논문에서는 네트워크 프로토콜 분석을 통해 네트워크 공격을 예측할 수 있는 마이닝 프로토타입 시스템을 설계 및 구현하였다. 구현된 마이닝 프로토타입 시스템에서는 네트워크 공격을 탐지하기 위해서 2단계로 수행된다. 1단계에서는 수집된 네트워크 프로토콜의 전처리 과정이다. 이 단계에서는 원시 데이터 형태의 프로토콜을 아스키 형태로 변환하며 각 프로토콜(TCP, UDP, ICMP) 별로 분류하여 데이터베이스에 저장한다. 2단계에서는 저장된 프로토콜을 가지고 마이닝을 수행하여 네트워크 공격을 탐지할 수 있는 규칙들을 생성한다. 규칙을 생성하기 위해서 연관규칙과 빈발에피소드를 적용하였고, 관심있는 속성을 기준으로 규칙을 탐지하기 위해서 키 속성 개념을 적용하였다. 실험을 통해서 네트워크 공격의 분포를 확인할 수 있었다. 마이닝 프로토타입은 지속적으로 생성된 규칙을 가지고 새로운 네트워크 공격 패턴과 유형을 예측하여 보안 관리자가 새로운 공격에 대해서 적절한 조치를 취할 수 있도록 도와줄 수 있다. 또한 침입 탐지 시스템의 보조적인 도구로서의 역할에도 기여할 수 있다.

참 고 문 헌

[1] Wenke Lee, Salvatore J. Stolfo, "Data Mining Approaches for Intrusion Detection," In Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, January, 1998.

[2] Wenke Lee, Salvatore J. Stolfo and K. W. Mok, "Mining audit data to build introduction detection models," In Proceedings of the 4th International conference on Knowledge Discovery and Data Mining, New York, NY, AAAI Press, August, 1998.

[3] W. Lee, "A Data mining framework for constructing features and models for intrusion detection systems," Ph.D thesis Columbia university, 1999.

[4] Wenke Lee, Wei Fan, "Mining System Audit Data : Opportunities and Challenges," In Proceedings of the ACM SIGMOD special issue 4, New York, NY, December, 2001.

[5] K. Julisch. "Dealing with False Positives in Intrusion Detection," In 3nd Workshop on Recent Advances in Intrusion Detection, http://www.raid-symposium.org, 2000.

[6] Cuppens, F., Miege, A. "Alert correlation in a cooperative intrusion detection framework," In Proceedings of the IEEE Symposium on Security and Privacy, 2002.

[7] Chris Sinclair, Lyn Pierce, Sara Matzner, "An Application of Machine Learning to Network Intrusion Detection," In Proceeding of the 15th Annual computer security applications conference, Phoenix, Arizona, 1999.

[8] W. W. Cohen. "Fast effective rule induction. In Machine Learning," the 12th International Conference, Lake Taho,

CA, Morgan Kaufmann, 1995.

[9] W. Lee, "A Data mining framework for constructing features and models for intrusion detection systems," Ph.D thesis Columbia university, June, 1999.

[10] Salvatore J. Stolfo, Wei Fan, Wenke Lee, "Cost-based Modeling for Fraud and Intrusion Detection : Results from the JAM Project," In Proceedings of the DARPA Information Survivability Conference and Exposition, 2000.

[11] V. Paxson, "Bro : A System for detecting network intruders in real-time," In Proceedings of the 7th USENIX Security Symposium, 1998.

[12] P. A. Porras, P. G. Neumann, "EMERALD : Event monitoring enabling responses to anomalous live disturbances," In National Information Systems Security Conference, 1997.

[13] C. Warrender, S. Forrest, B. Pearlmutter, "Detecting intrusions using system calls : Alternative data models," In Proceedings of the 1999 IEEE Symposium on Security and Privacy, 1999.

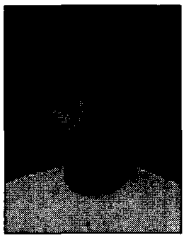
[14] S. Forrest, S. Hofmeyr, A. Somayaji, T. A. Longstaff, "A sense of self for Unix processes," In Proceedings of the IEEE Symposium on Security and Privacy, 1996.

[15] A. K. Ghosh, A. Schwartzbard, "A study in using neural networks for anomaly and misuse detection," In Proceedings of the 8th USENIX Security Symposium, 1999.

[16] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques," Morgan Kaufmann Publishers, 2001.

[17] 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, 1993.

[18] V. Jacobson, C. Leres, and S. McCanne, tcpdump. available via anonymous ftp to ftp.ee.lbl.gov, June, 1989.



김 은 희

e-mail : ehkim@dblab.chungbuk.ac.kr

2001년 삼척대학교 정보통신공학과 학사

2003년 충북대학교 전자계산학과 석사

2003년~현재 충북대학교 전자계산학과 박사과정

관심분야 : 데이터 마이닝, 데이터베이스 보안, 접근 제어, 침입 탐지 시스템

신 문 선

e-mail : msshin@dblab.chungbuk.ac.kr

1988년 충북대학교 전산통계학과 학사

1997년 충북대학교 전자계산교육 석사

1999년~현재 충북대학교 전자계산학과 박사과정 수료

관심분야 : 시공간 데이터베이스, 데이터 마이닝, 데이터베이스 보안, 침입 탐지 시스템

류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr

1976년 숭실대학교 전산학과 이학사

1980년 연세대학교 공학대학원 전산전공 공학석사

1988년 연세대학교 대학원 전산전공 공학박사

1976년~1986년 육군군수 지원사 전산실(ROTC장교), 한국전자통신연구원(연구원), 한국방송통신대 전산학과(조교수)

1989년~1991년 Univ. of Arizona Research Staff(TempIS 연구원, Temporal DB)

1986년~현재 충북대학교 전기전자및컴퓨터공학부 교수

관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 객체 및 지식기반 시스템, 지식기반 정보검색 시스템, 데이터마이닝, 데이터베이스 보안 및 Bio-Informatics