情報保護學會論文誌
第16卷 第6號, 2006. 12

과탐지 감소를 위한 NSA 기반의 다중 레벨 이상 침입 탐지*

김 미 선,† 박 경 우, 서 재 현‡
목포대학교

Negative Selection Algorithm based Multi-Level Anomaly Intrusion Detection for False-Positive Reduction

Mi-Sun Kim,† Kyung-Woo Park, Jae-Hyun Seo‡
Mokpo National University

요 약

인터넷이 빠르게 성장함에 따라 네트워크 공격기법이 변화되고 새로운 공격 형태가 나타나고 있다. 네트워크상에서 알려진 침입의 탐지는 효율적으로 수행되고 있으나 알려지지 않은 침입에 대해서는 오탐지(flase negative)나 과탐지(false positive)가 너무 높게 나타난다. 또한, 네트워크상에서 지속적으로 처리되는 대량의 패킷에 대하여 실시간적인 탐지와 새로운 침입 유형에 대한 대응방법과 인지능력에 한계가 있다. 따라서 다양한 대량의 트래픽에 대해서 탐지율을 높이고 과탐지를 감소할 수 있는 방법이 필요하다.

본 논문에서는 네트워크 기반의 이상 침입 탐지 시스템에서 과탐지를 감소하고, 침입 탐지 능력을 향상시키기 위하여 다차원 연관 규칙 마이닝과 수정된 부정 선택 알고리즘(Negative Selection Algorithm)을 결합한 다중 레벨 이상 침입 탐지 기술을 제안한다. 제안한 알고리즘의 성능 평가를 위하여 기존의 이상 탐지 알고리즘과 제안된 알고리즘을 수행하여, 각각의 과탐지율을 평가, 제시하였다.

ABSTRACT

As Internet fastly grows, network attack techniques are transformed and new attack types are appearing. The existing network-based intrusion detection systems detect well known attack, but the false-positive or false-negative against unknown attack is appearing high. In addition, The existing network-based intrusion detection systems is difficult to real time detection against a large network pack data in the network and to response and recognition against new attack type. Therefore, it requires method to heighten the detection rate about a various large dataset and to reduce the false-positive.

In this paper, we propose method to reduce the false-positive using multi-level detection algorithm that is combine the multidimensional Apriori algorithm and the modified Negative Selection algorithm. And we apply this algorithm in intrusion detection and, to be sure, it has a good performance.

Keywords : *Intrusion Detection System, False-Positive, Association Rule Mining, Negative Selection Algorithm, Anomaly Detection*

접수일: 2006년 10월 16일 ; 채택일: 2006년 11월 29일
* 본 연구는 2005년 목포대학교 학술 연구비 지원에 의하여 수행하였습니다.

† 주저자, misun@mokpo.ac.kr
‡ 교신저자, jhseo@mokpo.ac.kr

Ⅰ. 서 론

인터넷의 급속한 확장으로 인해 네트워크 공격기법의 패러다임이 변화되고 새로운 공격 형태들이 나타나고 있으나 대부분의 침입 탐지 기술은 오용 탐지 기술을 기반으로 하는 시스템이 주를 이루고 있어 알려진 공격 유형만을 탐지하고, 새로운 공격에 능동적인 대응이 어려운 실정이다.

정상적인 행위 정보를 축적하고, 이를 기반으로 침입 여부를 결정하는 이상 탐지 기법은 알려지지 않은 새로운 침입을 탐지할 수 있다는 장점을 가지고 있지만, 정상 행위 정보의 불완전성으로 인하여 정상 행위 패턴을 침입으로 간주하는 과탐지(False Positive) 오류를 야기한다. 침입 탐지 시스템은 과탐지의 가능성에도 불구하고 일단 이상 행위가 발생하면 경고를 발생하게 된다. 따라서 침입 탐지 시스템은 하루에도 아주 많은 양의 경고를 발생시키게 된다. 어떠한 행위에 대한 로그 기록들이 계속 이루어지기 때문에, 관리자는 모든 로그에 대해 분석하여, 그에 대응하는 판단 및 대응책을 제시하기에는 많은 어려움이 있다[1].

본 논문에서는 네트워크 기반에서 과탐지의 최소화 및 침입 탐지 능력을 향상시키기 위하여 대량의 패킷 데이터에 대하여 이상 침입 탐지를 위한 패턴 생성 및 탐지를 위한 방법으로 다차원 연관 규칙 마이닝과 수정된 부정 선택 알고리즘(NSA, Negative Selection Algorithm)을 결합한 다중 레벨 이상 탐지 알고리즘을 제안한다.

제안된 시스템의 다중 레벨 이상 탐지 과정은 크게 다음의 두 단계로 구성된다.

첫 번째 레벨에서는 데이터 전처리 과정을 통해서 정제된 데이터에 대해 데이터 마이닝 기법 중 연관 규칙을 적용하여 정상 행위 프로파일 및 정상 규칙을 생성하고, 생성된 정상 행위에 대해 일차적으로 이상 여부를 판별한다.

두 번째 레벨에서는 면역 시스템 기반의 부정 선택, 복제 선택을 수행하여 이상 탐지기를 생성하고, 생성된 이상 탐지기에 의해 이상 탐지과정을 수행한다.

면역 시스템에 기반한 기존의 연구는 일정기간 모아진 정적 데이터를 사용하여 자기 공간을 생성하고, 이에 대한 이상 탐지기를 생성함으로써 동적으로 변화하는 관찰대상에 대한 능동적인 대처가 어려웠다[2,3]. 반면 본 논문에서는 생성된 정상 패턴과 탐지 규칙에 대해서도 지속적인 모니터링을 수행하여 이상 탐지기 셋을 지속적으로 갱신하여 동적으로 변화하는 네트워크 환경에 적합한 이상 탐지가 수행될 수 있도록 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문의 기반 기술이 되는 이상 탐지 알고리즘에 대해 설명한다. 3장에서는 NSA 기반의 다중 레벨 이상 탐지 기술에 대해 제시한다. 4장에서는 제안한 다중 레벨 이상 탐지 기술의 성능을 평가하기 위하여 KDD99 데이터를 사용하여 각 알고리즘별 시뮬레이션을 수행하고 결과를 도식화하여 보여준다. 5장에서는 결론과 미래 연구방향에 대해 제시한다.

Ⅱ. 이상 탐지 알고리즘

이상 침입 탐지 시스템은 비정상으로 보이는 행위 패턴을 탐지한다. 이상 행위는 오용 침입과 합법적인 사용으로 명확히 알려진 것 이외의 행위를 말한다. 이상 탐지 기술은 비정상으로 보이는 행위 패턴을 탐지하는 기술로 알려지지 않은 새로운 공격에 대한 탐지 가능성이 있어서 최근 이상 탐지 기술에 대한 연구가 활발하게 이루어지고 있다. 이상 탐지 기술에서 사용되어 지고 있는 방법들로는 통계적 분석, 규칙 기반 분석, 신경망, 데이터 마이닝, 베이지안 네트워크, 인공 면역 시스템 등이 있다.

네트워크 기반 침입 탐지는 네트워크를 통해 수집하는 데이터의 양이 방대하기 때문에 대량의 데이터를 빠르고 효율적으로 분석하기 위한 방법이 요구된다. 따라서 본 논문에서는 이상 탐지 알고리즘 중 대량의 데이터에 대해 쉽게 드러나지 않는 유용한 정보를 추출하는 과정으로 데이터 마이닝 기법을 사용하며, 이상 탐지 알고리즘에서 나타나기 쉬운 과탐지를 감소하기 위한 방법으로 인공 면역 시스템 기법을 적용한다.

데이터 마이닝 기법 중 연관 규칙(Association Rule Mining)은 데이터 집합에서 아이템 사이의 관계를 찾는 방법이다[4]. 따라서 사용자가 사용하는 명령어와 명령어 사이의 연관성, 호스트와 호스트 사이의 연관성 등을 찾아서 침입탐지에 이용할 수 있다. 연관규칙을 탐사하는 문제는 기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 갖는 항목집합들의 모든 집합들인 빈발항목집합(large itemset)을 찾아내어 연관규칙을 생성하는 단계로 이루어진다[4].

본 논문에서는 다중 레벨 탐지 과정의 첫 번째 레

벨에서 정상 규칙 및 정상 행위 프로파일링을 생성하기 위하여 연관 규칙 생성 알고리즘 중 Apriori 알고리즘을 사용한다.

Apriori 알고리즘은 데이터베이스에서 개념 일반화(generalization)에 상관없이 강한 규칙성, 즉 강한 연관성을 갖는 항목들을 발견하는데 초점을 두고 있으며, 사용자가 정의한 최소 지지도를 이용하여 빈발 항목 집합들을 구성하고, 빈발 항목 집합에 대해서는 신뢰도를 이용하여 강한 연관성을 발견하는 방식이다[4].

본 논문에서는 정상 패킷의 다중 속성들간의 연관성을 추출하기 위하여 다차원 연관 규칙을 적용한다. 이진 연관 규칙이 아닌 다차원 연관 규칙이기 때문에 모든 빈발 속성 집합을 탐색하기 위하여 Apriori 알고리즘을 수정하여 빈발 K-속성 집합을 탐색한다.

다중 레벨 탐지 과정의 두 번째 레벨에서 이상 행위 탐지를 위해 사용되는 이상 탐지기 생성을 위해서는 인공 면역 시스템을 모델링한다.

면역 시스템은 그 특성에 따라 자연 면역시스템(Innate Immune System)과 적응 면역 시스템(Adaptive Immune System)이 결합되어 구성되어진다[5]. 자연 면역 시스템은 항원과 상관없이 이미 존재하고 있으며, 침입한 미생물들에 대하여 즉시 반응할 수가 있기 때문에, 항원을 일차적으로 제거해 주는 역할을 한다. 적응 면역 시스템은 항원을 인식할 수 있는 항체의 생성으로부터 시작되며 특별한 항원을 인식할 수 있는 임파구가 활성화되어 항원과의 바인딩을 통해 항원의 제거가 이루어지는 반응으로 자연 면역반응에 의하여 항원들이 완전히 제거되지 않았을 때, 유도되어 나타나는 이차적인 방어체계가 된다.

적응 면역 시스템을 구성하는 기본 요소는 두 가지 형태의 임파구로 B 세포와 T 세포이다. B 세포는 특별한 항원을 인식하여 바인딩 함으로써 항원을 제거하는 항체를 생산, 분비하는 역할을 하며 T 세포는 항원을 제거하거나, B 세포의 성장을 억제하거나 도와주는 역할, 또는 항원을 정상적으로 인식하지 못하는 B 세포를 제거하는 역할을 담당한다. 적응 면역 시스템에서 항원은 부정적 선택 (Negative Selection), 긍정적 선택 (Positive Selection), 복제 선택 (Clonal Selection) 모듈들을 통해 항체와 상호 작용하여 면역 응답을 발생한다[6].

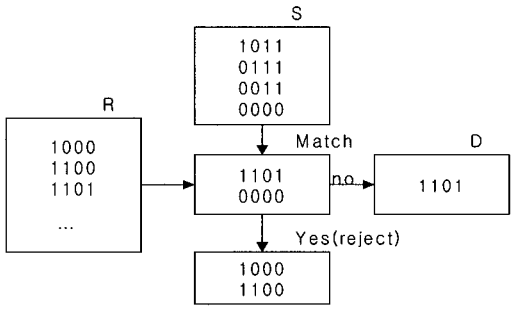
부정선택은 인공 면역 시스템에서 침입에 대해 정상적으로 반응을 보이지 못하는 셀을 신속히 제거하는 역할을 하며 항원의 인식에 있어서 자기를 항원으로 인식하는 것을 배제하기 위한 방법이다[7]. 생체 면역계의 면역 세포 생성원리 중의 하나인 부정선택을 이용하여 자기 인식 알고리즘을 구현한 경우가 S. Forrest의 NSA (Negative Selelction Algorithm)이다.

Forrest의 부정 선택 알고리즘은 크게 세 단계로 수행된다[3].

i. 모니터링된 정상 패턴에 대해 자기를 정의한다. 일정한 길이의 문자열을 가진 패턴으로 자기공간을 생성한다.

ii. 탐지자 집합(detector set)을 생성한다. [그림 1] 에 도시된 대로 생성된 자기공간에 대해 매칭기법을 이용, 매칭되지 않은 스트링열에 대해 탐지자(detector)를 생성한다.

iii. 자기공간을 지속적으로 모니터링하며 탐지자와의 매칭을 시도, 매칭되는 경우 비자기로 검출한다.



[그림 1] 자기공간과의 매칭을 통한 탐지자 생성

이 알고리즘은 랜덤하게 탐지자를 생성하고, 자기를 탐지하는 탐지자들은 제거하여 남아있는 탐지자들이 비자기를 탐지할 수 있도록 하는 유사한 개념들에 영향을 미친다. 만일 어떤 탐지자에 의해서 탐지되었다면 비자기로 인식한다[2].

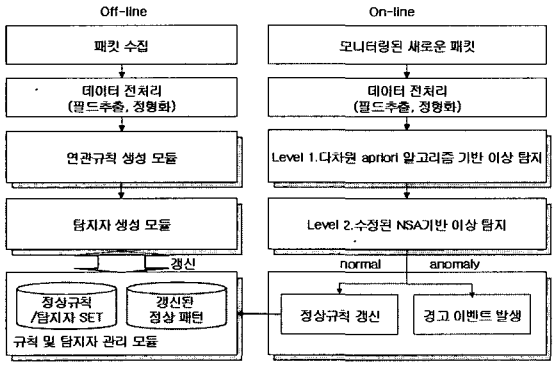
본 논문에서는 NSA 기법을 확장하여 비자기 즉, 이상 행위를 탐지하는 알고리즘을 제안한다. 기존의 NSA 와 달리 다중 탐지자를 제공하고, 온라인상에서 모니터링 된 정상 데이터를 추가하여 정상 프로파일을 갱신하고 또한 복제선택을 통하여 탐지자의 지속적인 갱신과 메모리화 과정을 수행하여 최적화된 탐지자를 제공한다. 동적이고 최적화된 탐지기를 생성하기 위하여 면역 시스템의 복제 선택이론을 반영한다. 복제 선택 이론 (Clonal Selection Theory)은 특정 항원을 인식할 수 있는 세포를 선택하여 활성화시켜서 클론의 수를 증식하는 과정으로 특정한 항원에 신속하고 효율적으로 반응할 수 있게 한다. 즉,

효과적인 탐지능력을 지닌 cell을 복제한다[5].

Ⅲ. NSA 기반의 다중 레벨 이상 탐지

이상 행위 탐지 방법을 이용한 침입 탐지 시스템에서 기초가 되는 정보는 사용자의 정상 행위 패턴이다. 네트워크 기반 침입 탐지는 네트워크를 통해 수집하는 데이터의 양이 방대하기 때문에 대량의 데이터를 빠르고 효율적으로 분석하기 위한 방법이 요구된다. 본 논문에서는 대량의 네트워크 데이터에 대한 빠른 분석과 이상 탐지 수행을 위하여, 데이터 마이닝 기법 중 다차원 연관 규칙과 수정된 NSA 알고리즘을 결합한 다중 레벨의 이상 탐지 알고리즘을 제안한다. 본 논문에서 제안하는 다중 레벨 이상 탐지 알고리즘은 첫 번째 레벨에서 대량의 데이터에 대해 쉽게 드러나지 않는 유용한 정보를 추출하는 과정으로 데이터 마이닝 기법 중 다차원 연관 규칙 기법을 사용하며, 두 번째 레벨에서 이상 탐지 알고리즘에서 나타나기 쉬운 과탐지를 감소하기 위한 방법으로 인공 면역 시스템 기법을 모델링한 NSA 알고리즘을 수정하여 적용한다.

본 논문에서 제시하는 시스템의 구성 및 전체적인 구조는 [그림 2]와 같다.

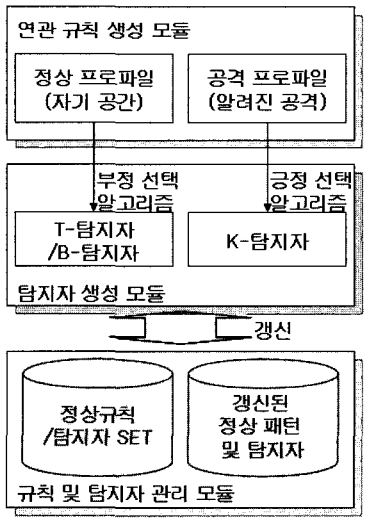


[그림 2] 다중 레벨 이상 탐지 시스템 구조

정상 행위를 모델링하기 위한 과정인 프로파일링은 정상 행위들로만 구성된 자료로부터 정상 규칙을 생성하고, 정상 규칙 데이터베이스에 유지한다. 생성된 정상 규칙은 이상 탐지 알고리즘의 첫 번째 레벨에서 일차적으로 이상 행위를 탐지한다. 기존의 연관 규칙 기반 이상 탐지 시스템은 이 단계에서 바로 이상 탐지를 발생하는 것과 달리 본 논문에서는 과탐지를 줄이기 위한 방법으로 첫 번째 레벨에서 이상으로 발견된 데이터에 대하여 두 번째 레벨에서 수정된 NSA 기반 이상 탐지 알고리즘을 적용하여 이상 여부를 판별한 후 이상 탐지 경고를 발생한다. 기존 이상 탐지 시스템과 달리 두 단계의 이상 탐지 알고리즘을 사용함으로써 과탐지를 감소하고자 한다.

1. 정상 프로파일 및 탐지자 생성

네트워크 기반의 침입 탐지는 네트워크 데이터인 패킷 헤더 정보를 이용하여 이상이나 오용침입을 탐지한다. 따라서 네트워크 기반의 이상 탐지를 위해서 기초가 되는 정보인 사용자의 정상행위 패턴을 생성하는 과정이 중요하다. 본 논문에서 제시한 다중 레벨 이상 탐지 알고리즘을 수행하기 위해서는 탐지과정에서 사용하기 위한 정상 행위 프로파일, 정상 규칙, 탐지자가 필요하다. 따라서 오프라인 상에서 미리 수집된 패킷에 대한 데이터 전처리, 다차원 연관 규칙을 적용한 정상 행위 프로파일 및 정상 규칙 생성, 부정선택과 긍정선택을 사용한 탐지자 생성 과정을 [그림 3]과 같이 수행한다.



[그림 3] 정상 규칙 및 탐지자 생성

본 논문에서는 KDD-CUP99 데이터 집합에 대한 SVM(Support Vector Machine)의 척도로 제시된 41개의 척도 중 개별 TCP 연결의 기본 사양 9개를 사용하여 정상 행위 프로파일링을 수행한다[8]. 선택된 9개의 척도는 [표 1]과 같다.

〔표 1〕 정상 프로파일을 위한 특징

선택 특징	설 명
duration	연결 시간
protocol_type	프로토콜 타입 (TCP,UDP, ICMP)
service	서비스 종류 (HTTP, FTP 등)
flag	정상 또는 에러 플래그
src_bytes	소스로부터 데이터 길이
dst_bytes	목적지로부터 데이터 길이
land	1:같은 소스/목적지주소,0
wrong_fragment	"Wrong"fragment 개수(0-3)
urgent	Urgent 패킷 개수(0-3)

정상 규칙 및 정상 프로파일은 데이터 마이닝 기법 중 연관 규칙을 사용하여 정상 행위 패턴들이 가지고 있는 연관성을 추출하여 생성한다. Apriori 알고리즘은 각 패스에서 빈발항목 집합들의 후보 항목집합을 구성하고 난 후에 각 후보 항목집합의 발생 빈도수를 계산하고, 사용자가 정의한 최소 지지도와 최소 신뢰도를 기초로 하여 빈발항목 집합들을 결정한다. Apriori 알고리즘은 분석하고자 하는 네트워크 데이터 셋과 임계치로서 미리 정의된 최소 지지도와 최소 신뢰도를 입력으로 받는다.

본 논문에서 사용되는 데이터 셋인 패킷 데이터들은 그 특성상 각각의 트랜잭션이 두 개 이상의 속성을 가지고 있기 때문에 다차원 연관 규칙을 적용한다[4]. 정상 프로파일을 위한 데이터들이 (duration, service, flag, src_bytes, dst_bytes)라는 5개의 속성을 가지고 이들은 규칙 내에서 한 번씩만 사용된다. 다차원 연관 규칙을 마이닝하기 위해서는 빈발 항목 집합을 탐색하는 대신에 빈발 속성 집합을 탐색해야 한다. k-속성 집합이란 k 개의 논리곱으로 연결된 속성들을 가지는 집합을 말하며 본 논문에서는 5-속성 집합을 사용한다. 다차원 연관 규칙이기 때문에 모든 빈발 속성 집합을 탐색하기 위하여 Apriori 알고리즘을 수정하여 빈발 K-속성 집합(Lk)을 탐색한다. 수정된 다차원 Apriori 알고리즘은 〔표 2〕에서 보여주고 있다.

〔표 2〕 연관 규칙 탐사를 위한 다차원 Apriori 알고리즘

```
Input : database Đ, support ϕmin,
  confidence δmin
Output : Rc all association rules
begin
   for (k=1; D.degree; k++) do begin
      Fk={frequent l-attributesets};
   Ll=Fl;
   for (k=2; D.degree; k++) do begin
      Ck = multi_apriorig_gen(Lk-1, Fk, ϕ
  min );
       //generate new candidates from
         Lk-1⋈Fk
   for all transactions T∈D do begin
      Ct=subset(Ck,T); //candidates
        contained in T
      for all candidates X∈Ct do
         Count(X)=Count(X)+1;
      end
      Lk={X∈Ct | count(X)≥ϕmin×|D|}
   end
   Lf=UkLk;
   Rc=GenerateRules(Lf,δmin)
end
```

기존의 연관 규칙 기법은 특정 시점의 데이터베이스 상태 전체를 대상으로 규칙을 탐사함으로써, 연관 규칙 탐사 시점 이후에 발생하는 네트워크 정보에 대해서는 탐사 대상으로 고려할 수 없기 때문에 네트워크와 같은 동적 환경에서 연관 규칙의 실시간 탐사가 어렵다는 문제점을 갖는다.

본 논문에서는 데이터베이스 트리거 기능을 적용하여 이상 탐지 과정에서 정상으로 판단된 데이터가 발생되면 이를 구축된 규칙 DB에 새로운 트랜잭션으로 발생시켜, 업데이트 된 네트워크 데이터에 대한 정보를 반영하여 정상 규칙이 자동으로 갱신될 수 있도록 한다.

본 논문에서 다차원 Apriori 알고리즘을 수행하여 탐사된 규칙들은 최소지지도 5%, 신뢰도 50%로 설정하여, 초기에 15개의 규칙을 생성하였으며, 탐사된 규칙의 예는 아래와 같다.

duration = 1 ∧ service= http ∧ flag=SF ==〉 src_bytes=141..150 ∧ dst_bytes=0..1000 〔190, 62%〕

이 규칙의 의미는 duration 속성값이 1이고, service 속성값이 http이며, flag 속성값이 SF이고, src_byte 속성값이 141에서 150사이값이며, dst_bytes 속성값이 0에서 1000사이값을 갖

는 경우가 전체 270회이며, duration 이 1이고, service가 http, flag가 SF인 패킷중에서 src_byte 속성값이 141에서 150 사이값이며, dst_bytes 속성값이 0에서 1000 사이값을 갖는 경우가 62% 나타났다는 것을 의미한다.

연관 규칙들이 탐사되면 규칙들을 저장하기 위하여 고정된 길이의 2진 문자열로 변환하는 코드 변환 작업을 수행한다. 코드 변환을 위하여 속성 값에 대한 코드 정보를 2진 문자열로 정의한 후 이를 데이터 사전에 저장하여 정상 행위 프로파일 과정에서 사용한다.

변환된 코드의 예는 아래와 같다.

duration = 1, service= http, flag=SF, src_bytes=141..150, dst_bytes=0..1000
==〉 01 0010 01 00100 001

코드 변환 작업을 수행함으로써 규칙 중 필요한 정보만을 용이하게 추출할 수 있으며, 문자열 패턴에 비하여 탐지 과정의 비교연산이 효율적으로 수행된다.

다차원 Apriori 알고리즘을 사용하여 생성된 정상 규칙과 코드 변환 작업을 거친 정상 프로파일은 〔그림 4〕와 〔그림 5〕에서 보이고 있다.

다차원 연관 규칙에 의해 정상 프로파일이 생성되면, 두 번째 레벨의 수정된 NSA기반 이상 탐지 과정에서 사용하기 위한 탐지자를 생성한다. 두 번째 레벨의 수정된 NSA기반 이상 탐지 과정에서는 면역 시스템 기반의 탐지자를 생성하여 이상 탐지 여부를 판별하게 되는데 기존의 NSA 알고리즘에서는 T셀에 대한 부정 선택 기법을 적용하여 단일 탐지자만을 사용하는데 반하여, 본 논문에서는 면역 시스템의 T셀과 B셀의 다중 방어 메카니즘을 적용한 다중 탐지자 기법을 사용하기 위하여 인공 면역 시스템의 부정 선택, 복제 선택을 수행하여 T/B 탐지자, 기억 탐지자를 생성한다. 또한 이상 탐지 알고리즘에 오용 탐지의 개념을 삽입하여 알려진 공격에 대해서도 빠르게 인식할 수 있도록 하기 위하여 비정상 데이터로부터 긍정 선택 기법을 적용하여 K-탐지자를 생성한다.

인공 면역 시스템에서 항원의 인식은 B셀이 항원의 특정 부분을 인식하고, T셀이 B셀의 항원 인식을 도와주는 역할을 수행한다. 이러한 면역 시스템의 기법을 이상 탐지 알고리즘에 적용하기 위하여 특정 위치를 인식하기 위한 B-탐지자와, 항원을 정확하게 인식할 수 있도록 하기 위한 일정 윈도우 사이즈를 가진 T-탐지자를 〔그림 1〕 에서 보여준 부정 선택 알고리즘을 적용한 자기 공간과의 매칭과정을 사용하여 생성한다.

K-탐지자는 알려진 공격에 대한 패턴을 기반으로 생성된 탐지자로서 비정상 데이터로부터 생성된 패턴에 의해 생성한다. K-탐지자는 비정상 데이터로부터 생성된 패턴에 대해 긍정 선택 기법을 사용하여 생성한다.

T-탐지자는 정상 패턴으로 구성된 자기 공간에서 각각의 정상 패턴을 r-연속비트 매칭을 사용하기 위해 일정한 윈도우 사이즈를 갖는 조각으로 나눈다. 이 나눠진 조각에 대해 〔그림 6〕의 a)와 같이 매칭되지 않는 경우를 T-탐지자로 생성한다. r-연속비트 매칭 기법은 부분 매칭 기법을 이용해서 연속적으로 인접한 r개의 셀들이 매칭되는 경우를 두 문자열 사

association : 테이블

duration	service	flag	src_bytes	dst_bytes
0	auth	SF	0..10	0..100
0	finger	SF	0..10	0..400
0	ftp_data	SF	11..20	0..100
0	http	SF	141..150	0..1000
0	http	SF	301..310	0..1000
0	smtp	SF	2001..2100	0..1000
0	smtp	SF	601..700	0..1000
0	smtp	SF	901..1000	0..1000
1	http	SF	141..150	0..1000
1	smtp	SF	1401..1500	0..1000
1	smtp	SF	801..900	0..1000
2	smtp	SF	1001..1100	0..1000
3	http	SF	151..160	0..1000
3	smtp	SF	801..900	0..1000
4	smtp	SF	1201..1300	0..1000

〔그림 4〕 다차원 연관규칙을 이용한 정상 규칙

Association_self : ...

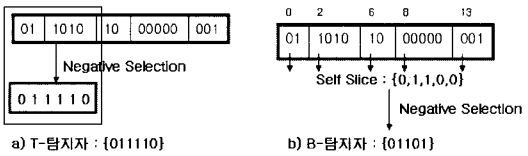
normal_profile
010001010001000010000
010001010001000010000
010000010001000000000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000
010001010001000010000

NUM

〔그림 5〕 정상 프로파일

이의 매칭으로 정의한다.

면역 시스템에서 B-세포 수용체는 항원의 특별한 표면 부분을 인식하여 항원을 제거하는 역할을 한다. 본 논문에서는 B-탐지자 생성을 위하여 정상 패턴에서 각각 특징의 첫 비트 항목만을 추출하여 윈도우 사이즈가 5인 자기 조각을 생성한다. 그리고 이 자기 조각에 대해 부정 선택 알고리즘을 적용하여 〔그림 6〕의 b)와 같이 B-탐지자를 생성한다.



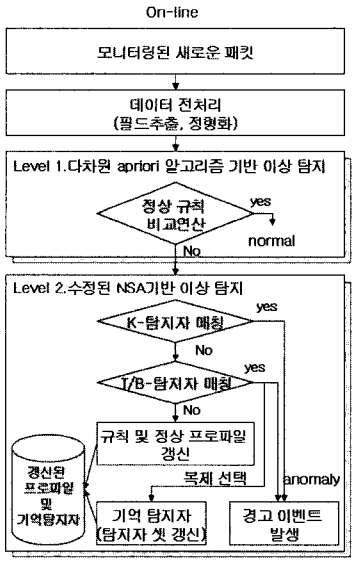
〔그림 6〕 T/B 탐지자 생성

2. 다중 레벨 이상 침입 탐지

정상적인 행위 정보를 축적하고, 이를 기반으로 침입 여부를 결정하는 이상 탐지 기법은 정상 행위 정보의 불완전성으로 인하여 정상 행위 패턴을 침입으로 간주하는 과탐지(False Positive) 오류를 야기한다. 그러므로 이상 탐지 시스템에서는 과탐지율이 시스템 성능의 중요한 평가 기준이 되고, 과탐지를 줄이기 위한 여러 가지 기법들이 연구되고 있다. 따라서 본 논문에서는 이상 침입 탐지 시스템의 과탐지를 감소하기 위한 방법으로 다차원 연관 규칙과 수정된 NAS 알고리즘을 결합한 다중 레벨 이상 탐지 기술을 제안하였으며, 다차원 연관 규칙 알고리즘을 이용하여 정상 규칙 및 프로파일을 생성하고, 면역 시스템 기반의 NSA 알고리즘을 수정, 적용하여 다중 탐지자를 생성하였다.

정상 규칙 및 프로파일과 다중 탐지자를 사용하는 다중 레벨 이상 탐지 수행과정은 〔그림 7〕과 같이 제안하였다.

〔그림 7〕에서 보이는 바와 같이 네트워크를 통하여 모니터링된 패킷 데이터가 입력되면 이상 탐지 과정을 수행하기 위한 필드 추출 및 데이터 전처리 과정을 수행하여 비교 연산을 수행하기 위한 정형화된 데이터로 변환한다. 데이터 전처리 과정을 거친 패킷 데이터들은 일차적으로 정상 규칙 집합과의 비교 연산을 통하여 이상 여부를 판별한다. 첫 번째 레벨의 다차원 apriori 알고리즘 기반 이상 탐지 과정에서 이상으로 판별된 경우, 두 번째 레벨로 데이터를 전송하여, 수정된 NSA 기반 이상 탐지 알고리즘을 수행한다.



〔그림 7〕 다중 레벨 이상 탐지 수행 과정

〔표 3〕은 〔그림 7〕의 다중 레벨 이상 탐지 수행 과정 중 Level 2.에 적용하기 위하여 수정된 NSA 알고리즘 이다.

K-탐지자와의 매칭을 통하여 이상 행위 여부를 판별하고, 매칭되지 않은 경우 기억 탐지자, T-탐지자와 B-탐지자의 조합을 통하여 이상행위를 판별한다. K-탐지자에 의해 탐지되지 않은 경우, 다음 단계로 기억 탐지자에 대한 탐지 작업을 수행하고, 이 단계에서도 탐지되지 않은 경우, T/B-탐지자에 대해 탐지작업을 수행한다. 또한 이 단계에서는 이상행위에 반응하는 탐지자를 기억 탐지자로 생성하고, 일정 기간동안 탐지 작업을 수행하지 못하는 탐지자는 제거함으로써 탐지자 셋을 최적화 할 수 있도록 한다.

T/B-탐지자에 의해서도 탐지되지 않은 경우의 데이터는 이상 행위가 발견되지 않은 것으로 간주하여 갱신된 정상 규칙 데이터 셋에 저장하고 규칙 및 탐지자 관리 모듈에서 정상 규칙 및 탐지자 셋을 갱신한다. B-탐지자는 탐지과정에서 반응할 경우 탐지 횟수를 증가시킨다. 시스템에서 미리 정의한 탐지 반응 임계값에 도달하면 기억 탐지자로 변환한다. B-탐지자가 일정 기간에 일정 탐지 횟수를 수행하지 못한 경우에 B-탐지자는 제거된다.

〔표 3〕 수정된 NSA 알고리즘

```
Input : K(K-detector-set),
M(Memory-detector-set),
B(B-detector-set), T(T-detector-set),
I(1-level output), dt (Detection
Threshold),
lt(Life time threshold)
output : detect, N(Normal-data)
begin
   dc = 0; // dc (B-detector's
detection count)
   L = 0; // L (B-detector's  Life time)
   M=null
        while(I){
         if(I==K)
            return detect=true;
         else if (M && I==M && I==T)

            return detect=true;
         else if (I==T && I==B)
            dc++;
            If (dt<=dc) M=M+B
               //Memory-detector generate
            return detect=true;
         else
            L++;
            if (lt<=L) B--//B-detector remove
            return detect=false, N;
         }
end
```

제안한 다중 레벨 이상 탐지 기술은 다차원 연관 규칙 마이닝의 사용으로 네트워크를 통해 수집하는 방대한 양의 데이터에 대해 빠르고 효율적인 분석이 가능하며, 이상 탐지 알고리즘으로 널리 사용되고 있는 NSA 알고리즘에 다중 탐지자를 적용한 수정된 NSA 알고리즘은 동적인 탐지자 셋을 사용함으로써 변화되는 네트워크 환경에 능동적으로 대응할 수 있는 장점을 갖는다.

연관 규칙 마이닝을 사용하는 기존의 이상 탐지 알고리즘은 연관 규칙과의 비교 연산을 수행한 후 이상으로 판별된 경우 바로 이상 신호를 발생하는데에 반하여, 본 논문에서 제안한 다중 레벨 알고리즘은 첫 번째 레벨에서 이상으로 탐지된 경우, 두 번째 레벨에서 다시 한번 이상 여부를 탐지하게 된다.

또한, 면역 시스템에 기반한 기존의 연구는 일정 기간 모아진 정적 데이터를 사용하여 자기 공간을 생성하고, 이에 대한 이상 탐지기를 생성함으로써 동적으로 변화하는 관찰대상에 대한 능동적인 대처가 어려웠다. 반면 본 논문에서는 생성된 정상 패턴과 탐지규칙에 대해서도 지속적인 모니터링을 수행하여 이상 탐지기 셋을 지속적으로 갱신함으로써, 동적인 탐지가 가능하다.

기존의 NSA 알고리즘에서는 T셀에 대한 부정 선택 기법을 적용하여 단일 탐지자만을 사용하는데 반하여, 본 논문에서는 면역 시스템의 T셀과 B셀의 다중 방어 메카니즘을 적용한 다중 탐지자를 생성함으로써 탐지율을 높이고자 하였으며 다음 절에서 NSA알고리즘과 다중 레벨 이상 탐지 알고리즘의 실험 평가를 통하여 이를 검증한다.

Ⅳ. 시뮬레이션

1. 실험 데이터 셋

네트워크 기반 이상 탐지를 위하여 본 논문에서는 KDD-99 Cup data set 중 correct 데이터를 사용하였다. 정상 프로파일을 생성하기 위하여 총 데이터 311,029 중 정상 데이터 3,500개로 샘플링하고, 프로토콜별로 분류하였다. 다중 Apriori algorithm에 최소 지지도와 최소 신뢰도를 각각 5%, 50%로 설정하여 초기 정상 규칙을 15개 생성하였다. 정상 프로파일의 각 스트링 길이는 21bit, B-탐지자와의 매칭 작업을 수행하기 위한 자기 조각은 스트링의 각 필드의 첫 번째 위치를 선택하여 윈도우 사이즈 5 인 문자조각을 생성하였다. 또한 오용탐지 개념을 적용하기 위한 K-탐지자는 공격패턴으로부터 비정상 프로파일을 생성하고, 이 비정상 프로파일로부터 프로토콜 타입이 TCP 인 경우에 대해서 20개의 패턴을 생성하였다. 이상 탐지에 사용되는 탐지자의 경우, B-탐지자는 자기 조각의 위치를 인식하기 위한 탐지자이므로 윈도우 사이즈를 자기 조각과 같은 5로 고정하여 생성하였으며, T-탐지자는 윈도우 사이즈를 다양하게 변화시켜 생성하였다.

2. 성능 척도

이상 탐지 시스템의 중요 성능 척도가 되는 탐지율과 과탐지는 아래와 같이 계산하여 얻을 수 있다[2][9].

$$Detection\ rate = \frac{TP}{TP+FN}$$

$$False\ \ alarm\ \ rate = \frac{FP}{TN+FP}$$

위의 식에서 TP(true positive)는 이상항목을 이상으로 인식한 경우, TN(true negative)는 정상항목을 정상으로 인식한 경우, FP(false positive)는 정상항목을 이상으로 인식한 경우, FN(false negative)는 이상 항목을 정상으로 인식한 경우를 의미한다.

본 논문에서 제시한 이상 탐지 알고리즘은 다양한 파라미터를 갖는다. 파라미터의 조합에 따라 이상 탐지의 결과가 민감하게 반응하기 때문에 각 파라미터들의 값을 변화하여 실험을 수행하였으며, 탐지율과 과탐지의 결과를 ROC 커브로 나타냈다.

3. 실험 결과

실험은 크게 3가지로 나누어 수행하였으며, 첫 번째는 T-탐지자의 윈도우 사이즈와 탐지자의 수를 변화시켜 시뮬레이션을 수행하였고, 두 번째는 대표적인 면역 시스템 기반 이상 탐지 알고리즘인 NSA 알고리즘을 적용한 경우와 본 논문에서 제안한 다중 레벨 탐지 알고리즘을 적용한 경우로 나누어 수행하였으며, 세 번째는 본 논문에서 제안한 두 알고리즘을 각각 적용한 경우와 비교하여 시스템의 성능을 평가하고자 하였다. 각 실험과정은 10회 반복하여 수행하였으며, 각 수행결과의 평균값을 취하였다.

실험 1. T-탐지자 임계값에 따른 탐지율 비교

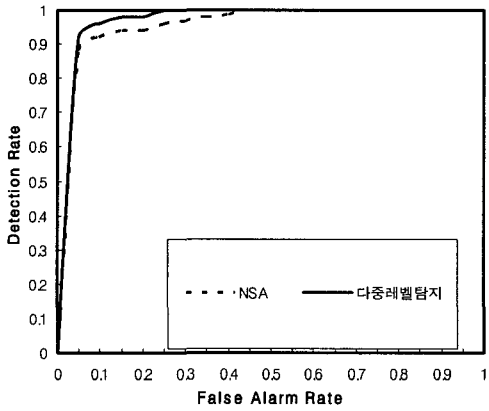
T-탐지자의 윈도우 사이즈(t_w)를 6, 8, 12로 변경하고, 각 경우에 탐지자의 수(Nt)를 30, 50으로 변경하여 알고리즘을 수행한 후 탐지율과 과탐지율을 비교하였다. 과탐지율은 T-탐지자의 윈도우 사이즈가 클수록 감소, T-탐지자의 수가 많을수록 증가함을 보인다. 또한, 탐지율은 T-탐지자의 윈도우 사이즈가 클수록 증가, T-탐지자의 수가 많을수록 증가함을 보인다.

〔표 4〕 T-탐지자 임계값에 따른 탐지율 비교

t_w	Nt	Detection Rate	False Alarm rate
6	30	87.12	22.96
	50	89.45	24.12
8	30	92.16	17.25
	50	94.75	19.57
12	30	96.25	11.78
	50	97.12	12.37

실험 2. NSA와 다중 레벨 탐지 알고리즘 비교

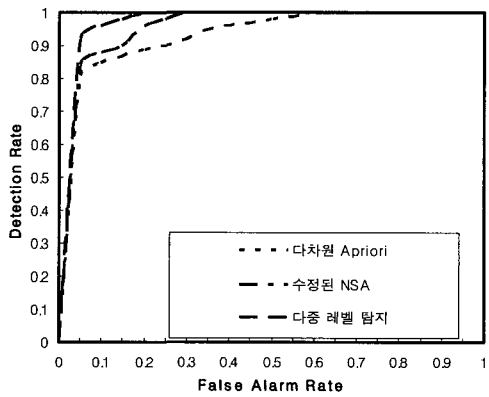
NSA의 경우 다중 레벨 탐지 알고리즘에서 적용하는 T-탐지자와 같은 길이의 윈도우 사이즈를 갖는 단일 탐지자만을 생성하여 이상 탐지를 수행하였으며, 이 경우 실험1의 실행결과 가장 낮은 과탐지율을 보인 T-탐지자의 윈도우 사이즈 12, 탐지자의 수 30개로 설정하였다.



〔그림 8〕 NSA와 다중 레벨 탐지 알고리즘 비교

실험 3. 단일 레벨 탐지와 다중 레벨 탐지 비교

다차원 Apriori 알고리즘과 수정된 NSA를 각각 실행하여 이상 탐지를 수행한 경우와 두 알고리즘이 결합된 다중 레벨 탐지를 수행한 경우의 이상 탐지를 비교하기 위한 실험으로, 단일레벨에서 사용되는 파라미터와 다중레벨에서 사용되는 파라미터의 값은 동일하게 지정하여 사용하였다.



〔그림 9〕 단일레벨 알고리즘과 다중 레벨 탐지 알고리즘 비교

V. 결 론

본 논문에서는 네트워크 기반에서 과탐지의 최소화 및 침입 탐지 능력을 향상시키기 위하여 다차원 Apriori 알고리즘과 수정된 부정 선택 알고리즘(NSA, Negative Selection Algorithm)을 결합한 다중 레벨 이상 탐지 알고리즘을 제안하였다.

기존의 연관 규칙 기법은 특정 시점의 데이터베이스 상태 전체를 대상으로 규칙을 탐사함으로써, 연관 규칙 탐사 시점 이후에 발생하는 네트워크 정보에 대해서는 탐사 대상으로 고려할 수 없기 때문에 네트워크와 같은 동적 환경하에서 연관 규칙의 실시간 탐사가 어렵다는 문제점을 갖는다. 본 논문에서는 다차원 연관 규칙 마이닝 기법에 트리거 기능을 적용하여 연관 규칙 탐사 시점 이후에 이상 탐지 과정에서 정상 데이터가 발생하면 이를 구축된 연관 규칙 DB에 새로운 트랜잭션으로 발생시켜, 업데이트 된 네트워크 데이터에 대한 정보를 반영하여 정상 프로파일이 자동으로 갱신됨으로써 실시간 네트워크 데이터에 대한 빠른 분석이 가능하다.

또한 수정된 NSA 알고리즘은 이상을 정확히 탐지하는 탐지자를 기억 탐지자로 변환하여, 다음에 같은 이상 행위가 발생되었을 때 더 빠르고, 능동적으로 대응할 수 있다는 장점을 갖으며, 탐지 능력이 떨어지는 탐지자는 탐지자 집합에서 제거하고, 갱신된 정상 프로파일에 대해 새로운 탐지자를 지속적으로 생성함으로써 새롭게 발생하는 네트워크 패킷들에 대해 능동적인 탐지가 수행될 수 있다.

다양한 파라미터의 설정과 반복된 실험 과정을 통하여, 기존에 사용되는 단일 레벨 알고리즘에 비하여, 제시한 다중 레벨 탐지 알고리즘이 과탐지를 감소할 수 있음을 보였다. 그러나, 실험 결과에 비추어 제안된 이상 탐지 시스템은 각각의 탐지자에 대한 한계값 및 파라미터의 값에 많은 영향을 받음을 알 수 있다. 따라서 추후 다양한 패킷 데이터군에 대한 실험 및 다양한 한계값의 설정에 따른 실험을 지속적으로 수행하여 알고리즘을 최적화하여 탐지 시간을 줄이고, 효율적이고 신뢰성있는 이상 탐지가 이루어질 수 있도록 하여야 할 것이다.

현재 제안된 알고리즘의 탐지 시간을 줄이기 위한 방안을 연구 중에 있으며, 생성된 침입 경고에 대해 필터링을 수행한 후 경고 시스템에 이벤트를 발생시키기 위한 방안에 대해 연구하고 있다.

참 고 문 헌

[1] P. Tadeusz, T. Axel, "Data mining and machine learning-Towards reducing false positives in intrusion detection", Article Information Security Technical Report, Volume 10, Issue 3, pp. 169-183, 2005.

[2] D. Dasgupta, S. Yu and Majumdar, N., "MILA Multilevel Immune Learning Algorithm," GECCO 2003, LNCS 2723, pp.183-194, 2003.

[3] D. Dasgupta and S. Forrest, "An Anomaly Detection Algorithm Inspired by the Immune System", Artificial Immune Systems and Their Applications 1st edition Part III, Springer, pp.262-275, Dec. 1998.

[4] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann publishers, 2001.

[5] R. Goldsby, T. Kindt, and Osborne, B., Kuby Immunology, 4th Edition, W.H. Freeman & Company, Jan. 2000.

[6] de Castro. L. N. and Von Zuben, F. J, "Artificial Immune Systems: Part I Basic Theory and Applications," Technical Report RT DCA 01/99, 1999.

[7] S. Forrest, Perelson, A., Lawrence Allen, and Rajesh Cherukuri, "Self-Nonself Discrimination in a Computer", In IEEE Symposium on Research in Security and Privacy, pp.202-212, May 1994.

[8] KDD CUP 1999 DATA, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[9] F. Gonzalez and D. Dasgupta, "Anomaly detection using real-valued negative selection", In special issue of the Journal of Genetic Programming and Evolvable Mach

ines, Vol. 4, Issue 4, pp 383-403, Dec. 2003.
[10] DARPA Intrusion Detection Evaluation, MIT Lincoln Laboratory, http://www.ll.mit.edu/IST/ideval.
[11] D. Chowdhury, "Immune Network: An Example of Complex Adaptive Systems", Artificial Immune Systems and Their Applications, 1st edition, Part II, Springer, pp.89-114, Dec. 1998.
[12] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok, "A Data Mining Framework for Building Intrusion Detection Models", IEEE Symposium on Security and Privacy, 1999.
[13] S. A. Hofmeyr, "An Immunological Model of Distributed Detection and Its Application to Computer Security", PhD Thesis, Dept of Computer Science, University of New Mexico, May 1999.
[14] Jianxiong Luo and Susan M. Bridges, "Mining Fuzzy Association Rules and Fuzzy Frequency Episodes for Intrusion Detection", International Journal of Intelligent Systems, Vol. 15, No. 8, pp.687-704, 2000.

〈著者紹介〉

김 미 선 (Mi-Sun Kim) 정회원
1996년 2월: 목포대학교 컴퓨터공학과 졸업
2000년 2월: 목포대학교 컴퓨터공학과 석사
2002년 3월~현재: 목포대학교 컴퓨터공학과 박사과정
<관심분야> 컴퓨터공학, 네트워크보안, 정보보호

박 경 우 (Kyung-Woo Park)
1986년 2월: 전남대학교 계산통계학과 졸업
1988년 2월: 전남대학교 전산통계학과 석사
1994년 2월: 전남대학교 전산통계학과 박사
1995년 3월~현재: 목포대학교 정보공학부 컴퓨터공학전공 교수
<관심분야> 컴퓨터공학, 분산 시스템, 시스템 소프트웨어, 정보 보호

서 재 현 (Jae-Hyun Seo) 종신회원
1985년 9월: 전남대학교 계산통계학과 졸업
1988년 2월: 중앙대학교 전자계산학과 석사
1988년 3월~1996년 6월: 송원대학교 전임강사
1996년 8월: 전남대학교 전산통계학과 박사
1996년 9월~현재: 목포대학교 정보공학부 정보보호전공 부교수
<관심분야> 정보보호, 시스템 및 네트워크보안, 컴퓨터 네트워크