네트워크 침입탐지를 위한 복제 선택 알고리즘의 적용

김정원(University College London) 최종욱(상명대), 정길호(성균관대)

요약

외부침입탐지 시스템(IDS: Intrusion Detection System)은 컴퓨터의 외부 침입을 자동으로 탐지하는 시스템이다. IDS의 주요목표는 외부사용자들이나 내부 사용자들에서 권한이 없는 사용자 컴퓨터 오용(misuse) 혹은 잘못된 사용(abuse)를 탐지하는 것이다. 파이어 월(Firewall)이나 암호화와 같은 침입 방지 시스템에 관한 연구와 병행하여 최근 IDS에 대한 다양한 연구가 이루어지고 있다. 침입탐지와 바이러스 탐지에 대한 새로운 접근 방법으로서 면역학적 방법이 동원되고 있다.

이 연구에서는 인간의 인체면역 시스템으로부터 얻어진 몇 가지 주요한 Feature들을 외부침입 탐지에 적용하여 기존의 침입탐지 방법에서 오는 한계점을 극복하여 경고 오류(alarm error rate)을 줄이고자 한다. 따라서 본 연구에서는 외부침입을 탐지하고 시스템을 치유하는 인간의 인체 면역에 대해 기초적인 연구를 진행하였으며 이러한 인체면역 기저들을 네트워크 환경에서 어떻게 실제적으로 적용할 것인 지를 연구하였으며 실제 네트워크 데이터를 적용하여 본 연구에서 제안한 모델에 대한 성능을 테스트하였다.

1. 인체 면역 시스템과 네트워크 침입 탐지

최근의 IDS에 관한 대부분의 연구들이 Host기반으로 개발되었으나 본 연구에서는 네트워크 기반의 IDS시스템(Mykerjee et al, 1994)으로 개발되었다. 호스트 기반의 IDS는 한 대의 호스트 컴퓨터에 기록된 audit trails을 분석하고 이를 기반으로 침입을 탐지하는 기술이다. 네트워크 기반의 IDS는 복수의 호스트 컴퓨터와 네트워크 통신을 감시함으로써 여러 대의 호스트 컴퓨터를 감시하는 침입탐지 기술이다.

호스트 기반의 IDS와 네트워크 기반의 IDS 모두 이상 탐지(anomaly detection)와 오용탐지(misuse detection)를 사용한다 (Mykerjee et al, 1994). 이상 탐지방법은 일반사용자, 시스템 사용자, 시스템 자원, 네트워크 통신(Traffic)이나 서비스 등의 정상적인 프로파일을 결정하여 이와는 현저히 다른 모습의 시스템 사용이나 서비스의 형태, 혹은 통신량을 보이는 경우 이를 침입으로 간주하는 방법이다. 오용탐지 방법은 이와는 달리 이미 알려져 있는 시스템의 침입에 대한 의심스러운 패턴을 사용하여 시스템을 감시하고 만약 알려진 패턴과 유사한 패턴이 발견되는 경우 이를 침입으로 간주하는 방안이다. 이 접근 방법에서는 오용 패턴이 audit trail에 존재하는지 여부가 중요한 관건이 된다. 이러한 두 가지 접근 방법은 각기 장점과 취약점이 잇으며 상호 보완적인 관계를 가지고 있다(Mykerjee et al, 1994).

본 연구에서는 인체면역 시스템과 네트워크 기반의 IDS시스템이 상당한 유사성을 갖고 있다는 점에 초점을 맞추어서 새로운 모델과 기술을 개발하고 이를 기반으로 실제 시스템을 개발하였다. Somayaji *et al*(1997)은 그들의 연구에서 인체면역을 이용한 일반적인 컴퓨터 침입탐지 시스템에 대한 일반적인 원칙과 몇 가지 가능성을 제안하고 있다. 이들의 연구와는 달리 본 연구에서는 특정한 문제를 염두에 두고 인체 면역의 몇 가지 뚜렷한 특징을 갖는 기능을 이용하여 네트워크 기반의 경쟁력이 있는 시스템을 설계하는데 주안점을 두고 있다.

1.1 네트워크 기반 IDS의 요구사항

인체면역 시스템을 논하기 이전에 경쟁력있는(competent) 네트워크 기반의 IDS이 가져야 할 기능에 대해 논의하는 것이 필요하다. 현재까지 IDS에 관한 논문들에서 제시되어 있는 네트워크 기반 IDS의 중요한 기능들은 다음 일곱 가지 정도로 요약할 수 있다.

강인성(Robustness): 일반적으로 네트워크 IDS 시스템은 여러 탐지점 (detection

points)를 가지고 있어야하고 각 탐지점들은 공격이나 시스템 결함에도 견딜수 있어야한다 (Balasubramaniyan et al., 1997), (Forrest et al., 1997). IDS 시스템에 있어 가장 중요한 약점은 외부공격에 대한 실패 혹은 시스템 붕괴이다. 만약 외부공격자가 IDS의 존재를 알고 있고 그것을 전복시킬 수 있다면 IDS 개발 자체가 필요 없게 된다.

변형성(Configurability) IDS는 자체적으로 각 호스트 컴퓨터 혹은 각 네트워크의 필요조건을 만족시킬 수 있도록 쉽게 변형이 가능해야 한다 (Balasubramaniyan et al., 1997), (Somayaji et al., 1997). 네트워크에서 각 호스트 컴퓨터는 모두 다른 기종으로 구성되어 있는 경우가 많고 이 때문에 각 기 다른 보안 요구 사항을 가지고 있다고 볼 수 있다. 이외에도 네트워크가 달라지면 라우터, 필터, DNS, 파이어 월 그리고 네트워크 서비스가 달라지기 때문에 보안 요구조건도 다양하게 달라질 수 있다.

확장성(Extendibility) 단순히 운영체제가 달라져도 새로운 호스트시스템에 의한 IDS 감시의 범위가 쉽게 확장될 수 있어야 한다(Balasubramaniyan et al., 1997), (Somayaji et al., 1997). 이는 새로운 호스트 컴퓨터가 기존의 네트워크 환경에 더해지거나 새로 붙여진 호스트 컴퓨터가 다른 Audit 데이터를 갖는 운영체제위에서 돌아가더라도 현재의 IDS와 같은 방식으로 감시가 이루어져야한다.

형상성(Scalability) 분산된 호스트를 통해서 대량의 Audit데이터를 정확하게 수집하고 분석할 수 있도록 신뢰성있는 규모의 확장성을 가지고 있어야한다 (Balasubramaniyan et al., 1997). 단일 IDS의 경우에도 audit trail 데이터의 수집은 분산적이지만 분석 자체는 중앙집중식이 된다(Mykerjee et al, 1994). 하지만 이 경우에도 모든 audit 데이터를 단 하나의 IDS에 손실없이 보내기는 어렵다. 만약 모든 감시 데이터를 바르게 수집 분석할 수 있도록 시스템이 규모의 적응성을 가진다 하더라도 네트워크의 성능에 심각한 장애가 발생할 가능성은 있다.

적응성(Adaptability): 네트워크 외부침입의 변화에도 역동적으로 적응할 수 있어야한다(Balasubramaniyan et al., 1997), (Somayaji et al., 1997). 컴퓨터 시스템의 환경은 계속적으로 변화한다. 사용자와 시스템 판매자, 그리고 시스템 관리자는 계속적으로 변화를 하고 있으며 외부침입자들의 행태역시 끊임없이 변화한다. 따라서 네트워크 IDS는 이처럼 끊임없이 변화하는 네트워크의 환경과 외부 침입에 대해 적응력을 갖추고 있어야한다.

전역적 분석력(Global Analysis) IDS는 네트워크 침입을 탐지하기위한 충분한 증거를 수집하기 위해서는 각 호스트 컴퓨터에서 발생하는 여러 가지 이벤트 데이터를 제대로 수집하여야하며 여러개의 이벤트들 사이의 상관관계를 잘 파악할 수 있어야한다 (Balasubramaniyan et al., 1997), (Mykerjee et al, 1994). 대부분의 네트워크 침입은 네트워크의 여러 곳을 파고든다. 따라서 어느 한 호스트에서는 정상적인 실수처럼 보이지만 여러 곳의 호스트들을 통해 종합적으로 판찰하게 되면 명백히 하나로 연계된 침입일 수가 있다.

효율성(Efficiency) IDS가 설치되는 호스트 컴퓨터나 네트워크에 무리가 가지 않을 정도로 단순하면서도 가벼운 시스템이 되어야 한다(Balasubramaniyan et al., 1997), (Forrest et al., 1996), (Somayaji et al., 1997) 하나의 IDS로 시스템 감시 (Monitoring), 데이터 수집, 데이터 조작, 의사결정을 수행해야 한다. 이러한 복합적인 기능 때문에 IDS가 시스템에 상당한 부담(overhead)이 되거나 특히 CPU와 I/O(입출력)에 부담이 될 수 있으며 심각한 시스템의 기능저하 및 네트워크 부하로 작용할 수 있다. 따라서 IDS의 개발과 선정에는 효율성이 중요한 변수로서 작용한다.

현재까지 여러가지 기술과 접근 방법이 제시되었지만(Balasubramaniyan, 1997), (Mykerjee et al, 1994), 이러한 요구사항을 완벽하게 만족시킬 수 있는 기술은 개발되지 않고 있다.

1.2 인체 면역학 개관 (Overview of human immune systems)

효율적인 네트워크 기반의 IDS를 설계하는데 있어 IDS에서 유용하게 사용될 수 있는 주용 기능을 찾아내기 위해서는 인체 면역 시스템의 중요 기능을 조사할 필요가 있다. 이 본문에서 제시된 인체 면역학에 대한 개괄적인 내용은

주로 (Paul, 1993), (Tizard, 1995)에서 밝혀진 것들이다. 인체면역 시스템의 대부분은 어느 한개의 특정 기관 때문이 아니라 엄청난 수의 각기 다른 형태의 타고난 원형질적인(선천적인) 세포 혹은 후형질적인 세포 때문에 가능해진다. 수없이 많은 각기 다른 세포들 중에서 임파구(백혈구)는 중심적인 역할을 한다. 이 임파구의 중요한 기능은 인체세포로 이루어지는 자기 세포로부터 위험스러운 타세포를 구분해내는 역할을 한다. 각 임파구는 제한된 수의 구조적으로 해로운 항원(antigen)이라고 불리는 세포들의 활동에 특별히 대응하도록 되어 있다. 임파구는 특정의 속박(binding)영역(수용체)이 있어 항원 결정자의 모양에 보완적인 형태를 갖게 된다. 특정 항원은 임파구 항체 수용체에 속박을 시켜주는 항원결정자 (determinant of antigens)를 보고 알 수 있다.

임파구는 주로 B-cell과 T-cell로 나누어진다. B-cell은 항체 세포이며 T-cell은 항원을 죽이거나 B-cell의 발달을 돕거나 방해하는 역할을 한다. B-cell과 T-cell은 공통적으로 그들 자신의 고유한 유전학적 구조를 가지며 모두 여러 개의 DNA연결고리로 표현된다. 이들 각 연결고리는 가변적인 영역(domain)과 고정적인 영역으로 이루어져 있다. 가변적인 영역에 들어있는 유전자는 각기 아주 가변적으로 속박(binding)할 항원의 특정 부위를 결정하는 역할을 한다. 고정적인 영역에 존재하는 유전자는 변동이 없으며 B-cell 항체 수용체가 항원 결정자에 속박될 때 다양한 생물학적 효과를 나타낸다. B-cell과 T-cell은 뼈의 골수와 흉선에서 각기 만들어진다. 골수와 흉선에서는 B-cell과 T-cell의 영역에 고유하게 상응하는 여러 개의 유전자 라이브러리는 B-cell과 T-cell 수용체를 표현하기 위한 후보를 내포하게 된다. 특정의 수용체는 유전자 라이브러리로부터 임의적으로 유전자 segment를 선택하거나 그들을 결합함으로서 만들어진다. 더구나 다양한 수용체를 만들어 내기 위해서는 이들이 만들어지는 과정에서 여러개의 단계적인 유전자 조작(genetic operators)를 취하게 된다. 유전자 조작은 유전자 재배치, 결합하는 각기 다른 사이트들의 선택, 체세포 돌연변이, 분류 변이 등이다(Tizard, 1995).

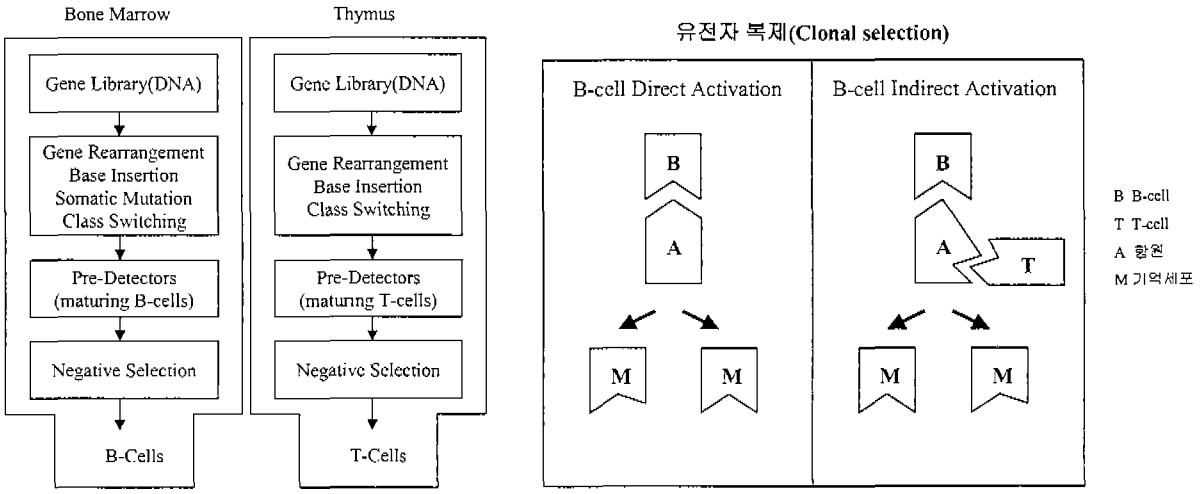

Figure 1 T-cell과 B-cell의 생성(왼쪽). 복제 선택(오른쪽).

B-cell과 T-cell은 골수와 흉선을 떠나기 이전에 성숙해져가는 B-cell이나 T-cell은 마지막 테스트 절차인 부정적인 선택 단계를 거치게 된다. B-cell과 T-cell의 생성기에는 완전히 다른 수용체가 여러 가지 유전자 조작기를 통해 만들어진다. 따라서 임의적으로 만들어진 수용체가 자기 세포 epitopes(항원 결정자)에 속박을 하게 될 가능성이 있다. 이것을 방지하기 위해 성숙되어가는 B-cell과 T-cell이 골수와 흉선을 돌고 있는 자기 세포에 속박을 하게 될 때 인체 속으로 방출되는 대신 죽게된다. Figure 1(왼쪽)은 골수와 흉선에서의 B-cell과 T-cell 의 생성을 보여주고 있다.

부정적인 선택과정을 거친 성숙한 B-cell과 T-cell은 골수와 흉선으로부터 방출된다. B-cell과 T-cell 모두 피 속에 있는 인체를

끊임없이 돌면서 항원을 만나서 활동을 하거나 진화를 하게 된다.

2. 인공 면역학 시스템의 복제 선택

부정적인 선택 과정에서 살아남은 새로운 항원들이 자기 내성(self-tolerant)을 가져야하지만 항원을 찾아내는 효율성(efficacy)에 대해서는, 즉 이들이 언제 흉선이나 골수에서 방출되는지 알려져 있지 않다. 이는 새로운 항체들이 무작위로 생성되고 이들이 단지 자기 세포가 아니라는 것만을 증명하기 때문일 것이다. 그들은 비자기(non-self) 패턴을 갖지만 항원의 패턴은 갖고 있지 않을 것이다. 이같은 비효율적인 탐색기를 제거하기 위해 인체면역 시스템에서는 현존하는 항원 패턴을 향해 항체가 점차 진화하는 것으로 이해하고 있다(Tizard, 1995). 이 진화 과정에서 인체면역 시스템은 복제 선택과정의 하나로서 항체의 일반성과 다양성을 유지하는 독특한 전략을 구사한다(Forrest et al, 1993).

2.1 인체 면역시스템의 복제 선택

항체의 숨어 있는 세포인 B-cell과 T-cell은 여러 가지 유전자를 섞는 메커니즘을 가지고 있고 이러한 발전 단계에서 항체의 다양성을 유지하기 위한 체세포 돌연변이(somatic mutation)을 사용한다. 이러한 메커니즘외에도 인체면역 시스템은 다른 전략도 갖추고 있다. 종족이 자연의 선택과정을 통해 진화하면서 인체면역 시스템은 복제 선택(Clonal selection)이라는 과정을 통해 진화한다. 항체의 특징을 결정짓는 유전자들은 어느 순간에라도 만연한 병원균을 찾아낼 수 있는 능력을 갖도록 끊임없이 진화하고 이러한 병원균에 대해 친화력을 가질 수 있도록 새로운 림프구를 재생산한다(Playfair, 1996), (Roitt and Brostoff, 1998), (Tizard, 1995).

B-cell이 골수에서 발달하면 오직 한정된 에너지만이 공급된다. B-cell은 골수에서 방출되자 말자 순식간에 활성화되지만 이들은 빠르게 탈진되고 죽어간다. 하지만, 이들이 만약 특정 항원과 결합되기만 하면 간단히 죽어없어 지지는 않는다. B-cell은 항원에 의해 직접적으로 간접적으로 활성화되지만 이 세포들은 없어지기 이전에 항원의 숨은 세포, 플라즈마 세포돌의 복제로서 쪼개지고 차별화된다. 플라즈마 세포는 부모 B-cell의 수용체와 동일한 항원-결합(anti-gen binding) 특성을 가지거나 항체 결합특성을 돌연변이 시키는 능력을 가진다. B-cell이 특정 항원을 많이 결합시킬수록 복제시 선택될 확률이 높아진다. 이와 비슷하게 이들이 특정 항원과 결합을 적게할수록 복제시 선택될 확률이 낮아지고 따라서 없어지게 된다. 현존하는 특정 항원의 다른 각기 다른 집단에 대해 가장 적합한 항원만이 살아남는다.

더구나 B-cell이 항원에 의해 활성화되었을 경우 동일한 항원에 대한 재사용을 위해 메모리 세포를 만들어 낸다. B-cell과 그 복제 세포인 플라즈마 세포의 생명은 비교적 짧다. 하지만, B-cell의 복제 세포들 중의 일부는 메모리 세포로서 살아남는다. 메모리 세포중의 일부는 항원에게 노출되어 차별화되어 체세포 변이를 일으키지 않은 채 플라즈마 세포가 되거나 다른 메모리 세포들은 체세포 변화를 거쳐 차별화되어 플라즈마 세포로 바뀐다. 특히 체세포 변이를 일으키지 않고 생성된 플라즈마 세포는 면역 시스템의 이차적 반응을 허용한다. 1차적 반응과 비교하여 메모리 세포에 의한 2차적 반응은 매우 빠르며 효율적이다. 더구나 메모리 세포는 연관 메모리 능력을 부여한다(Dhaeseler, 1997). 이같은 새로운 항원들은 이전에 탐색된 항원과 같지는 않지만 우사한 구조를 갖게 되고 메모리 세포에 의해 검출되어 진다. 이것은 항원과 항체의 결합이 근사에 의해 이루어지기 때문이다. 예를 들면, 인체가 우두에 감염되면 면역 시스템은 항원을 검출하고 제거하기 위해 시간을 돌여 노력한다. 그러나 만약 어떤 사람이 우두에 감염된 두에 다시 감염되었다면 그 사람은 면역 시스템의 제2차적인 반응에 의해 대단히 빠르게 치료가 된다. 이것은 우두와 천연두가 비슷하여 메모리 세포에 의해 제 2차적인 반응을 유도하기 때문이다.

2.2 복제 선택 알고리즘

전절에서 상술한 것처럼 인체 면역 시스템의 복제 선택은 항체의 다양성과 일반성을 유지하기 위해 여러 가지 중요한 메커니즘을 가지고 있다. 본 연구에서는 이중에서 인공 면역 시스템을 위해서는 부정적 선택 기능에서 설명된 비효율성이라는 역점을 보완하기 위한 자연적 선택 메커니즘을 집중적으로 사용하도록 하였다. 이 경우 복제 선택 기능의 진화적인 과정을 사용함으로써 부정적인 선택 과정의 무작위적인 생성으로 인한 컴퓨터 계산

시간이 줄어들 수 있다. 더구나 적합한 수의 검출기를 찾아내기 위한 조절문제(tuning problem)는 적소 전략에 의한 다양한 모드의 수렴에 의해 해결된다.

Forrest et al(1994)은 연구에서 인체 면역 시스템과 유사한 인공 면역 시스템의 적소 전략을 보여주고 있다. 그들은 (1) 무작위로 생성 나타난 항원들의 공통 패턴을 찾아 낼수 있는 지, (2)다양한 항원 집단을 구분해내고 유지해 나갈 수 있는 지에 대해 연구하였다. 그들의 모델에서는 한 집단의 항원을 만들어내고 한 집단의 항체를 만들어 냈다. 그들은 유전자 알고리즘을 이용하여 늘 같은 크기를 갖는 항원 집단환경에서 항체 집단을 진하하도록 하였다. 이 알고리즘은 인체 면역 시스템의 다음과 같은 독특한 특성들을 따라서 만들어졌다.

1. 특정 항체의 입장에서는 일반적으로 항원은 순차적으로 탐색된다.
2. 항원의 입장에서는 몇몇 항체의 숨어 있는 세포들에만 반응한다.
3. 항체들의 수겨진 세포들 사이에는 항원과 결합하기 위한 경쟁이 있다.
4. 항체의 숨겨진 세포들은 체세포 변이에 의해 진화한다.

각 세대별로 복제 선택에 의해 생겨난 인체 면역시스템의 적소 전략에 일치하도록 그들은 유전자 알고리즘을 수정하여 임의적인 크기의 무작위 샘플과 항원 집단으로부터 무작위적인 단일 항원을 선정하였다. 샘플에서의 각 항체가 선정된 항원에 대해 매치(match)되면 그 중에서 가장 높은 점수를 갖는 단 한개의 항체에 대한 접합 점수(fitness score)가 높아지고 다른 항체에 대해서는 접합 점수가 변동이 없게 된다. 결과적으로 인체면역 시스템의 특성들이 다음의 복제 선택 알고리즘으로 정리된다.

1. 하나의 단독 항원이 선택된다.
2. 고정된 크기의 항체 집단이 무작위로 선택된다.
3. 샘플에서의 각 항체는 무작위로 선정된 항원과 비교된다.
4. 가장 높은 매치 점수를 갖는 샘플에 속한 항체는 매치 점수를 접합 점수값(fitness value)에더한다. 다른 항체들의 접합 점수는 그대로 둔다.
5. 이 과정을 여러개의 항원들에 대해 반복한다.

이러한 알고리즘을 사용하여 Forrest et al (1993) 등은 점차 일반적인 성격을 갖도록 진화하여 대부분의 항원에는 어느 정도 매치가 된다고 설명하고 있다. 이러한 결과에 대한 이들의 분석은 항체가 여러 개의 항원들이 공유하고 있는 특질을 찾아내는 방향으로 진화한다는 것을 보여준다. 여러 가지 실험을 통해 그들은 이 알고리즘이 복수의 일관성이 없는 항체 패턴이 유지되고 이들은 탐색 공간에서 여러개의 피크를 갖는 것처럼 할 수 있고, 그리고 항원들 사이의 유사성은 이러한 능력에 영향을 미치지 않는다는 것을 보여주었다. 한 걸음 더 나아가 그들은 인체 면역시스템의 적소 전략과 접합 공유 알고리즘(fitness sharing algorithm)을 비교하였다. 이 비교에서 그들은 항체의 샘플링 메커니즘의 결과로서 인체면역 시스템의 적소 전략은 항체의 샘플 크기를 통해 그 일반성을 제어한다는 것을 보여주었다. 더 정확하게 말한다면 샘플 크기가 줄어들면 선택에 대한 압력이 더욱 일반적인 항체 집단을 형성하는 쪽으로 옮겨간다는 것이다.

이 부정적인 선택 알고리즘이 네트워크 침입에 대한 여러 가지 장점을 제공하지만 무작위적인 생성에 따르는 계산의 지나친 부하를 줄일 수 있도록 하여야한다. Dhaeseleer (1997)은 훨씬 더 효율적인 시간-선형 알고리즘과 greedy 알고리즘이라는 탐색 알고리즘을 소개하고 있다. 기본적인 아이디어는 모든 가능한 후보를 만들어 내어 부정적 선택 알고리즘으로 하여금 유효한 검출기를 선택하도록 효율적인 방법을 제공하는 데 있다. 하지만 이 알고리즘은 단순한 r-continuous-bits 매칭 알고리즘을 사용하기 때문에 다만 이진(binary) 면역 시스템에만 사용될 수 있다는 문제점이 있다. 이것은 자가 세포와 매치되지 않는 r-continuous-bits의 가능한 모든 이진 스트링(string)에 대해 재발생 횟수를 세어 가능한 유효 검출기를 찾아내어 배열하기 때문이다. Dhaeseleer는 또한 향후 중요한 연구과제로서 이진이 아닌 알파벳 면역 시스템을 사용할 것을 제안하였다. 이는 비이진 알파벳 면역 시스템이 성격상 더욱 자연스럽기 때문이다. 더구나 이 공식이 이진 코딩과 r-continuous 매칭 함수의 경우 적용된다 하더라도 이것을 실제 네트워크에 적용할 경우 그 처리량이 너무 많아지기 때문에 적용은 어렵게 된다.

결론적으로 이러한 알고리즘들 중의 하나를 선택해서 쓰기 보다는 본 연구에서 소개된 네트워크 기반의 IDS시스템을 구현하기 위한 부정적 선택 알고리즘은 유효한 탐색기를 생성하기 위한 Smith, Forrest, and Perelson's (1993)의 적소 전략 알고리즘을 사용하여야 한다.

2.3 네트워크 침입탐지를 위한 복제 선택 알고리즘

복제 선택알고리즘의 핵심은 자기/비자기 프로파일을 구축하는데 있다. 본 연구에서는 네트워크에서 송수신되는 통신망의 원시 자료를 수집하여 패킷을 분석하고 그 연결 수준에 따라 자기/비자기 프로파일에 저장하였다.
이러한 프로파일은 이미 식별된 필드를 가지고 있어 정상적인 네트워크 활동과 비정상적인 활동을 구별할 수 있다. 그리고 이 프로파일은 적합한 데이터 표현으로 인코딩 된다. 두 번째 단계에서는 모든 자기/비자기 구분 프로파일이 인코딩되면 복제 선택 알고리즘은 탐색기 생성을 시작한다. 본 연구에서는 Forrest et al의 복제 선택 알고리즘을 부정적 선택을 하나의 조작으로 더해줌으로서 수정하여 사용한다. 타색기 생성을 위한 이 알고리즘의 두 번째 단계는 다음과 같이 정리할 수 있다.

모든 연결 프로파일과 상응되는 탐색기에 대해:

1. D 탐색 패턴이 무작위적으로 생성되고 이들의 접합치(fitness value)는 1으로 시작한다.
2. N 개의 탐색 패턴은 생성된 D개의 탐색 패턴으로부터 무작위로 선정된다.
3. 단일의 침입패턴이 비자기 프로파일로부터 선택된다.
4. 샘플에서의 각 탐색기는 모든 자기 패턴과 자기 프로파일과 비교되고 유사도가 측정된다. 만약 어떤 탐색기라도 자기 패턴과 정확히 일치한다면 이 탐색기는 무작위적인 유전자를 가지는 새로운 탐색기에 의해 교체된다.
5. 샘플속의 각 탐색기는 선택된 침입패턴과 비교되고 유사도가 측정된다.
6. 샘플속에 있는 단일 탐색기의 접합(fitness)치는 가장 큰 유사도를 갖는 탐색기의 접합도가 증가하고 다른 탐색들의 접합도는 그대로 남는다.
7. 과정 2-5를 되풀이한다 (대부분 항체 수의 3배정도: (Smith, Forrest, and Perelson, 1993)).
8. P_b % 탐색기 패턴이 부모 패턴으로 crossover, mutation등의 유전자 조작기들은 새로운 유전자 조작기 생성에 사용된다.
9. P_w % 자식 유전자에게 공간을 물려주기 위해 탐색기 패턴을 삭제한다.
10. 선정된 부모 탐색기와 8항에서 만들어진 자손 탐색기를 포함함으로써 새로운 탐색 집단이 선정되었다.
11. 접합치가 변화를 멈출 때까지 과정 2에서 9까지를 되풀이 한다.

이러한 과정을 거쳐 2단계가 끝나면 복제 선택 알고리즘은 새롭게 가져온 패킷을 분해(parsing)하여 새로운 자기 프로파일을 만들어 나간다. 3단계에서는 각 탐색기 세트에 저장된 탐색 패턴이 상응하는 새로운 자기 프로파일과 비교된다. 만약 어느 탐색기라도 새로운 자기(self)패턴과 매치된다면 그 알고리즘은 경고 메시지를 보내게 된다.

이전의 절에서 논의된 것처럼 적소 전략은 탐색기 크기에 따라 각 탐색기들의 일반성을 제어한다. 실제적인 이유로 이러한 알고리즘이 더욱 많은 일반적인 탐색기를 만들어 각 탐색기들이 하나 이상의 외부침입과 매칭이 되도록 하여야한다. 이것은 각 탐색기가 한개의 침입에 대해 정확히 접합은 되지 않지만 여러 개의 침입에 대해 어느 정도 접합이 가능함을 의미한다. 이러한 접근 방법은 네트워크 침입탐지에 적합한 방법이다. 이는 본 연구에서 사용하고 있는 각 자기 염색체의 길이와 이러한 자기 염색체들이 형성하는 탐색 공간의 길이가 대부분의 단순한 부정적 선택 알고리즘(Dhaeseleer, 1997)에서 사용한 탐색공간보다 아주 길기 때문이다. 더구나 복제 선택의 계산시간이 무작위 탐색이 아니라 진화적인 접근 방법을 사용하기 때문에 적어질 것이라고 예상한다. 마지막으로 탐색기의 적합한 수는 진화과정의 복수 모드 집중에 의해 자연적으로 결정될 것으로 믿기

때문이다.

2.4 복제 선택을 위한 통신망 데이터

이전의 절에서 설명한 데이터 세트를 대상으로 부정적 선택 알고리즘이 구현되었다. 이러한 데이터를 사용하여 첫 번째 시제품을 만들 수는 있지만, 데이터 량이 연결성 네트워크의 행태를 나타내기에는 충분치 않다. 이는 데이터들이 아주 짧은 시간 동안에 수집된 것으로 대체적으로 15분 정도에 해당한다. 이 때문에 더 많은 데이터 세트를 찾게 되었다. 두 번째 데이터 세트는 DARPA의 외부침입 평가 프로그램 (http://www.ll.mit.edu/IST/ideval/index.html)에서 입수하였다. 이 프로그램은 MIT Lincoln연구소에서 외부침입 탐지 기술을 조사하고 평가하기 위해서 수집하고 관리하는 프로그램이다. 이 프로그램에서는 약 4기가 바이트 정도의 7주 분량의 압축된 데이터로서 여러 가지 네트워크 침입을 실험한 데이터이다. 여기에서 수집된 데이터에는 Denial-of-service (서비스 거절), 원거리에서 권한이 없는 사용자의 접근, 지역적으로 권한이 없는 수퍼사용자들이 수퍼 사용자의 권한을 사용하는 침입, 그리고 경계 및 조사 공격의 4가지 유형으로 나뉘어 진다. 다시 말해 각 침입 타잎에 대해 비슷한 공격 시나리오를 사용하지만 여러 가지 다른 침입들이 모의 실험되어 있다. 첫 번째 데이터 세트와 비교하여 이 데이터 세트는 훨씬 더 폭넓은 공격 패턴을 가지고 있다.

이러한 두개의 데이터 세트의 데이터는 기본적으로 네트워크 패킷을 수집하기위한 필드, 예를 들어 time stamp, source IP address, source port, destination IP address, destination port 등등이 포함되어 있다. 하지만 수집된 네트워크 패킷의 필드들만 가지고는 의미있는 프로파일을 만들기 어렵다. 결론적으로 더욱 의미 있는 프로파일을 추출하기 위해서는 정상적인 데이터와 비정상적인 데이터를 판별하는 데이터 프로파이링 프로그램을 개발하여야한다. 여러 연구에서 TCP프로토콜의 보안 허점을 논의하였고(Porras and Valdes, 1998) 따라서 본 연구에서 사용한 프로파일은 이러한 연구에서 논의된 문제점들을 깊이 연구한 뒤 결정되었다. 이 프로파일들은 대부분 각 단일 연결활동을 나타내기 위해(describe) 정의된다.

이 자동 프로파일 프로그램은 TCP의 원시 자료에서 얻어진 연결 수준의 정보를 추출하기 위해 개발되었으며 첫 번째 데이터 세트에서도 의미 있는 필드를 추출하기 위해 사용되었다. 두번째 세트에 대해서는 Lee(1999) 가 DARPA의 데이터로부터 의미 있는 필드를 추출하기 위해 사전 처리된 데이터 세트를 제공하고 있어 이를 사용하였다.

각 TCP연결을 위해서는 다음과 같은 필드들이 두개의 데이터 세트로부터 추출되었다:
-연결 구분(Connection identifier): 각 연결은 다음 4가지 필드로 구성 된다: initiator address, initiator port, receiver address and receiver port. 이 4가지 필드는 각 연결을 구분하기 위해 우선적으로 프로파일에 포함된다.
- 알려진 포트 취약점(Known port vulnerabilities): 많은 네트워크 침입이 여러 가지 포트 취약점을 공격한다. 출발지 포트 혹은 도착지 포트들에서 잠재적으로 취약점을 가지는 몇 가지 필드 정보가 있다.
- 제3자 교환(3-way handshaking): TCP프로토콜에서는 안전한 통신을 위해 3자 교환을 사용한다. 어떤 네트워크 침입이 있을 경우 이 3자 교환의 규칙이 깨어지는 경우가 많다. 따라서 3자 교환 오류를 체크할 수 있는 필드들이 있다.
- 통신량(Traffic intensity): 네트워크 활동은 단일 연결상에서 생겨나는 통신량을 측정함으로서 관측할 수있다. 예를 들면 어느 특정 연결상에서 관측되는 패킷의 수와 데이터량은 그 연결에서 일어나는 일반적이고 정상적인 활동의 프로파일이 될 수 있다.
- 연결 컨텐츠(Connection Content): 이러한 필드는 네트워크 연결활동을 잘 기술한다. 한 예로 실패한 로그인 횟수나 수퍼 사용자 계정을 사용하여 성공한 경우이다.

따라서 전체적으로 본다면 네트워크 프로파일이라는 측면에서 첫 번째 데이터 세트에서는 35개 필드를, 두 번째 데이터 세트에서는 41개의 각기 다른 필드가 나타났다. 첫 번째 데이터 세트에서는 필드의 첫 번째 네가지 타잎이 나타났고 두 번째 세트에서는 Connection Content가 추가되었다.

2.5 구현

이절에서는 본 연구에서 제안된 복제 선택 알고리즘이 상세히 설명된다. 이 단계에서는

초기 유전자 라이브러리가 만들어진다. 이를 위해서는 유전자형(genotype)과 표현형(phenotype)에 대한 표현을 어떻게 할 것인지, 유전자 조작자 (operator)와 탐색기와 자기 패턴사이의 유사도를 측정하기 위한 접합함수(fitness function)에 대해 설명한다.

2.5.1 유전자형과 표현형(Genotypes and Phenotypes)

복제 선택 알고리즘은 판별 규칙에 대한 진화적인 메커니즘을 사용하고 있으며 이러한 규칙은 자기 세포와 비자기 세포를 구분하는데 사용된다. 판별 규칙을 표현하는 자연스러운 방법의 하나는 일련의disjunctive normal form (DNF) 규칙을 사용하는 것이다. 각 규칙의 *if-part*는 하나 혹은 두개 이상의 조건이 합동이며, ***then*** 부분은 그 규칙에 할당된 클래스 수준을 나타낸다. 이 연구에서는 생성된 단일 탐색기는 여러개의 규칙을 하나로 묶은 합동 규칙을 표현형(phenotype)으로 가지게 된다. 따라서 생성된 탐색기에 의해 탐색된 패턴은 이러한 합동규칙들의 비접합(disjunction)이 된다.

최근 유전자 알고리즘을 이용하여 Classification의 진화를 구현한 연구가 있었다(Bentley, 2000a), (Bentley, 2000b). 본 연구팀과 공동작업을 하고 있는 영국 UCL(University College of London)의 Bentley교수c팀이 진행한 전자상거래 Frauds 탐색을 위한 연구로서 사용자의 행태 데이터로부터 Fraud 여부를 구분하는 연구이다. 이 연구에서는 여러 가지 함수 세트, 즉, negation, larger-than, mutation 등에 적응할 수 있도록 유전자 알고리즘의 유연성을 이용하였다. 이러한 유연성 덕분에 표현형(phenotype), 분류 규칙등을 사용하여 좀 더 복잡한 조건까지 표현할 수 있게 되었다. 하지만, 이러한 접근 방법은 유전자 알고리즘의 강력한 표현력과 비트 스트링을 그 유전자 타잎으로 사용하는 표준형 유전자 알고리즘의 단순성을 희생하도록 하는 것이다. 본 연구의 상당히 중요한 면모는 기존에 제안된 인공면역학 알고리즘들이 약점의 분석과 보완에 있다. 그렇게 하기 위해서는 주어진 표현형(phenotype)을 표현하기 위해서는 강력하면서도 충분히 경제적인 동시에 기존 연구의 유전자 표현 타잎을 가장 가깝게 따라야 한다. 이 연구에서 사용된 유전자형과 표현형은 최선의 것이라고 주장하지는 않는다. 이 연구에서 제안된 표현 타잎은 Bentley등의 연구에서 제안된 것으로서 이 연구의 목적을 위해서는 적합한 것이라고 생각한다.

<u>유전자형 인코딩과 표현형 해석</u>

유전자형은 41개의 유전자로 이루어지고 각 유전자는 탐색기의 각 특징(feature)을 나타낸다. 이전 기술된 것처럼 본 연구를 위해 만들어진 프로파일은 41개 특징(feature)으로서 이루어지고 이 숫자는 탐색기에 포함된 이에 대응되는 유전자의 전체 숫자를 결정한다. 각 유전자는 현존하는 특징 값으로 이루어진다. 예를 들어 Protpcol-Type 특징에서는 이 특징점의 유효한 값돌로서는 tcp, udp and imcp돌이 있다. 따라서, 이러한 특징은 3가지의 가능한 값 중의 하나를 가지게 되며 이에 상응하는 유전자들은 오직 3개의 핵산 성분만을 갖게 된다. 그림-2 나타난 것과 같이 각 핵산 성분은 이진비트로서 그 값이 1인 것은 판별 규칙에서의 조건부분에 포함되어 있는 상응하는 특징을 나타내는 것이며 그 값이 0인 경우는 반대의 경우이다. 인공면역 시스템에서 유전자형을 이처럼 표현하게 되면 각 탐색기의 단독 특징점은 하나 이상의 값을 가지게 되며 각 특징값들에 별도의 비트를 할당함으로써 OR 연산자에 의해 특징값들을 결합하여 나타낼 수도 있다. 더구나 각 탐색기의 유효한 유전자들은 AND 연산자에 의해 연결될 수있다. 표 -2 나타나 있는 유전자 타잎은 "IF (field1 belongs to ([6..10] or [41..50]) and field2 belongs to ([1..2] or [5..9]) and .... and field5 belongs to [53..66])THEN it is intrusion" " 만약 필드-1의 값이 [6,,10]에 속하고, 필드-2 의 값이 [41,, 50]에 속하고, 필드-3의 값이 [....], ... 필드-5의 값이 [53,6,66]에 속한다면, 그렇다면 그것은 외부침입이다" 이라는 규칙으로 해석된다. 그리고 주의할 점은 한 단독 유전자의 핵산성분이 모두 같은 값을 갖는 경우 그 값은 모두 1일되던지 혹은 모두 0이 된다는 점이다. 그렇게 되면 이러한 특징들을 주어진 분류 규칙의 조건 부분에서 제외시킨다는 것을 의미한다. 이는 양쪽 모든 경우 (전자는 이 유전자의 모든 값이 그리고 후자의 경우 유전자의 어떤 값도 결코) 유전자의 어떤 값도 클래스를 결정하는데 연관성이 없는 것으로 해석될 수 있기 때문이다.

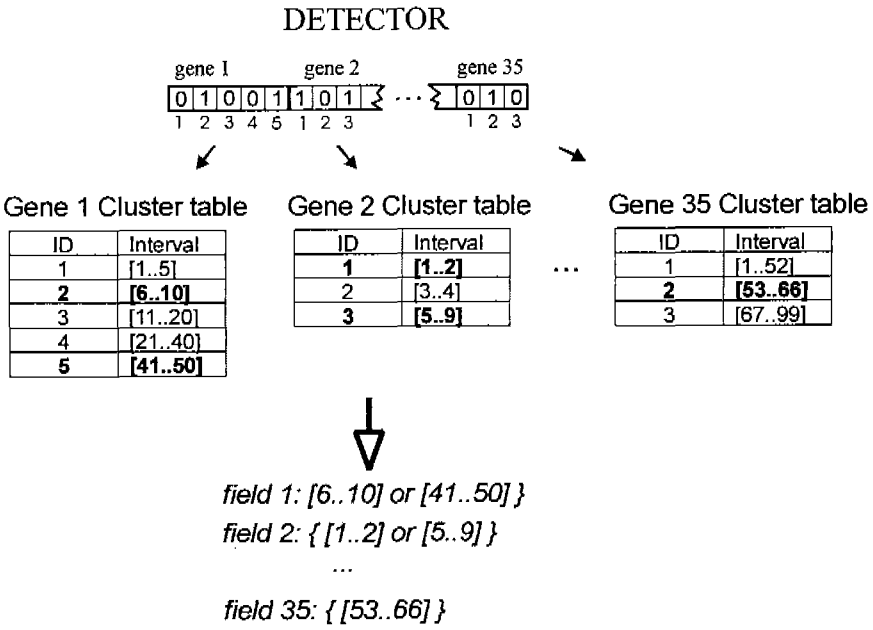

Figure 2 복제 선택에 대한 탐색기 표현

이러한 종류의 유전자형 표현은 새로운 것은 아니다. 이 같은 인코딩 방법은 매우 오랫동안 광범위하게 사용되어 왔으며 그 표현력은 실제적으로 규칙의 진화를 잘 다룰 있을 만큼 강력한 것으로 이미 증명이 되어 있다. De Jong(De Jong et al., 1993)에 의하면 유전자 알고리즘을 사용하는 이러한 표현 방법은 개념 학습에 사용할 수있을 것으로 제안하고 있는데, 개념 학습은 속성에 기반을 둔 분류 규칙의 진화를 배우는 것을 목표로 하고 있다. 이 접근 방법은 같은 분야의 다른 고전적인 유전자 인코딩 메커니즘과 비교된다. 유전자 알고리즘을 대안의 하나로서 언급한 것처럼, 판별 규칙을 진화로 표현하기 위한 여러 가지 다른 접근 방법이 있다. 예를 들어 실수값 기반의 인코딩, 퍼지 규칙 인코딩 등은 매우 강력하면서도 유연한 인코딩 방법이다. 하지만, 이미 언급한 다른 이유에서처럼, 유전자형의 인코딩 메커니즘을 이러한 방법들과 직접적으로 비교하는 것은 행해지지 않았다. 우리는 이 연구에서 분류규칙의 진화를 위해 미시간 접근방법에서 사용된 이진 인코딩 방법에만 집중하려고 한다.

다른 방법으로는 유전자 값을 충분히 커버할 수 있도록 적당한 비트수를 할당하는 것이다 (Goldberg, 1989). 예를 들면, 만약 특징1이 세개의 값을 가지고 그 것이 2비트를 필요로 한다면 이 2비트로 세가지 다른 경우, 즉 01, 10, 11을 충분히 표현할 수 있다. 이 방법에서는 각 유전자를 인코드하기 위해 필요한 비트 수는 2**이다. 여기서 **는 현재 존재하는 유전자의 값의 개수이다. 이 방법보다 De Jong이 제안한 유전자 인코딩 접근법이 유리한 점은 각 단독 탐색의 각 특징에 대한 특징 값들이 비연결적(disjoint)이 될 수 있다는 점이다. 하지만 다른 방법들은 탐색기의 각 특징점마다 한가지 값만을 표현할 수 있다. 이처럼 하여 단독 탐색 규칙은 더욱 강력한 표현력을 갖게 되어 더욱 더 일반화되며 그 결과 더욱 많은 침입패턴과 매치가 이루어진다. 결과적으로 주어진 수의 침입패턴을 탐하여야하는 탐색 규칙의 수는 점차 작아지게 된다. 가벼운 탐색기가 네트워크 외부침입 탐지를 위해 필요하게 되었으며 De Jong식 접근 방법의 강함때문에 생성된 탐색기의 가벼운 특징을 선호한다.

이러한 유전자 타입의 표현은 분류규칙의 진화에 대한 통상적인 접근 법이다. 그러나 약간 다른 방법이 개념의 패턴을 인식하기 위해 인공면역 알고리즘에 의해 사용된다. Forrest et al (1994), Porter(1997)는 개념 패턴 학습을 위해 인공면역 시스템을 개발하였으며 그들은 위에 소개된 것돌과는 다른 유전자 인코딩 방법를 사용하였다. 그들은 약간 제한된 유전자 설명을 사용하였지만, 그들은 매치 역치 개념을

사용함으로서 광속의 경량화된 특징을 생성된 탐색기에 부여하였으며 이러한 특징으로 인해 인체면역 시스템의 근사 바인딩에 근접해 보인다. 이 방법의 간단한 묘사를 하자면 각 항체와 항원이 고정된 길이의 이진 스트링으로 표현되고 사전에 정의된 비트 수가 보완적인 경우 이 두개의 스트링은 서로 매치 된다고 믿는다. 주어진 한 개의 항체 스트링은 이 보완적인 비트의 수가 특정하게 정해진 역치(threshold)를 넘어서는 한 하나의 항원이상의 매치를 가질 수가 있다. 하지만 역치의 값을 정한느 것이 어렵다. 이문제를 해결하기 위해 Potter는 하에가 진화하여 복제 선택 알고리즘에 의해 진화될 때는 이 값들이 함께 진화하도록 하였다.

Forrest et al(1994) and Potter(1997)에 의해 사용된 방법은 탐색기의 경량성을 유지한다는서 인체 면역 시스템과 가까워 보이지만 각 탐색기의 규칙들이 갖는 표현력은 De Jong이 사용하였던 인코딩 방식보다 약하다. 예를 들어 표-2 나타나 있는 De Jong의 방식에 의해 인코드된 단독 탐색기는 Forrest et al(1994) 이나 Potter(1997)의 인코딩 구조로 표현이 가능하다. 따라서 이들 방식이 역치(threshold) 매치 방식을 사용하여 생성된 탐색기의 일반성을 높여주지만 De Jong의 유전자 인코딩 방식은 이 본다 더욱 경량화 탐색기를 제공한다. 결론적으로 이 방법은 탐색공간의 크기와 생성된 탐색기의 경량성 특징과는 서로 trading의 관계에 있다. 이 론에서는 역치 매치 함수 방식이 일반적으로 나쁘다는 것은 아니다. 특히 유전자 알고리즘의 특유의 토리 구조는 여러 가지 함수를 사용하는 바, 강력한 표현력을 제공하며 구 역치(sphere threshold) 혹은 선형 역치(linear threshold)와 같은 매칭 역치 개념을 사용하고 있다. 즉 논리적인 구조의 표현 방식이 좋을 수 있다는 점이다. 그러나 본 연구에서는 De Jong의 접근 방식을 선호하고 있다. De Jong이 제안한 표현형(phenotype)은 결과적으로는 논리적인 표현이기 때문에 훨T니 읽기에 자연스럽고, 생성된 탐색기 규칙에 대한 강한 이해력을 제공하기 때문이다. 여기에서 말하는 이해력(intelligibility)는 보안 담당관이 IDS를 운영하기 위해 필요한 중요한 IDS의 특징 의 하나이다.

그러나 이 논문에서 논의된 것처럼 외부침입 탐지를 위한 인체면역 시스템의 전반적인 구조를 생각해보면, 중요한 관심사인 경량화된 IDS 시스템개발은 네트워크에서의 가능한 부담을 Primary IDS로부터 많은 수의 secondary IDS로 이전하여 줄이고자하는 것이다. 따라서 각 탐색기의 경량화라는 특징은 긴 탐색 시간으로 인하여 탐색기를 생성하기 위해 필요한 긴 계산시간보다 훨씬 중요한 것으로 생각된다. 이러한 문제는 Primary IDS들을 병렬 구조로 배열함으로서 해결할 수 있을 것으로 본다.

이산성(Discretisation)

본연구에서 구현한 실험에서는 각 네트워크 활동 프로파일이 41개 필드를 가지고 있다. 이 41개 필드로부터 32개 필드값이 연속적인 값을 가지며 다른 9개 필드값은 이산적인 값을 갖는다. 특히 32개 필드의 연속적인 필드값은 넓게 분포되어 있다. 이처럼 다양하고 넓게 분포되어 있는 값을 유전자형으로 인코딩하기 위해서는 이산화 알고리즘이 필요하다.

부정적인 선택을 위해서는 각 이산화된 클러스터는 동수의 데이터를 갖는 것으로 정의하고 있다. 단순히 이산화 알고리즘에 의해 클러스터가 제공되지만 각 클러스터가 심각한 정보 손실없이 만들어지고 있는지 여부는 확실하지 않다. 결과적으로 본 연구에서 사용한 복제 선택과정은 엔트로피 기반의 이산화 알고리즘을 사용하였다(Witten and Frank, 2000). 이 알고리즘의 기본적인 아이디어는 속성의 수치적인 값을 계속 재귀적으로 분할하여 클러스터의 간격이 엔트로피를 더 이상 최소화할 수 없을 때까지 만든다는 것이다. 다시 말해 이 알고리즘은 정보 이득(information gain)을 최대화할 수 있을 때까지 분할점을 찾는 것이다. 정조 이득이란 분할이 없는 경우와 분할이 있는 경우의 그 차이를 말한다.

하지만, 이러한 단순한 엔트로피 기반의 이산화 알고리즘은 과적합 문제(overfitting Problem)가 지적되고 있다. 엔트로피는 데이터 포인트의 순수성을 측정하는 것이기 때문에 엔트로피 값은 어느 분할 순간의 데이터 포인트가 동일한 클래스에 속할 때 최소가 된다. 더구나 생성된 클러스터들이 오직 학습용 데이터만 반영하고 있는 점이 문제로 지적되고 있다. Fayyad and Irani (1993)는 엔트로피 기반의 이산화 알고리즘은 과적화(overfitting) 문제를 해결하기 위한 수단으로 제의하였다. 최대의 정보 손실을

찾기보다는 이 제안에서는 최소 묘사 길이(Minimum description length: MDL) 원칙을 택하여 재귀적인 분할을 정지할 수 있도록 한 것이다. 이 원칙에서 주장하는 것은 최상의 해결책은 학습된 시스템의 전체적인 비용을 최소화 시키는 점에 놓여 있다는 것이다. 시스템의 비용이란 시스템의 복잡성과 현재 시스템을 적용하였을 경우 에러를 정하기 위해 필요한 정보량을 더한 값으로 정의된다. 다시 말해 훈련기간 동안은 시스템 복잡도가 증가하고 이 비용은 훈련 데이터 세트의 특징을 더욱 밀접하게 반영한다. 따라서 훈련기간 중의 예측 에러는 줄어든다는 점이다. 그 시스템의 진화를 조기에 정지함으로서 시스템의 복잡도는 감소하지만 정보를 부정확하게 예측한 것에 대해서는 더 많이 지불하여야 한다. 따라서 MDL 원칙에서는 훈련 데이터 세트를 정확하게 반영하는 시스템을 만들기 위해 지불해야하는 높은 비용과, 시스템이 복잡하지 않은 경우 추가적인 정보를 보내 잘못된 정보를 수정하는데 필요한 비용과의 균형을 맞추는 점에서 결정된다는 것이다.

Fayyad and Irani's의 재귀적인 분할의 정지 판단은

$$Gain(A.T,S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T,S)}{N}$$

---- (1),

여기서 N은 set S에 속한 샘플의 수, A는 특징, T는 파티션 경계, 그리고

$$Gain(A,T,S) = Ent(S) - E(A,T,S)$$

T에 의해 만들어진 파티션의 클래스 정보 엔트로피는

$$E(A,T,S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_1|}{|S|} Ent(S_2)$$

그리고 $Ent(S)$는 S의 엔트로피,

$$\Delta(A,T,S) = \log_2(3^k - 2) - [k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)$$

그리고 k_1는 set S_i에 속한 클래스 레벨의 수. 수식(1)에서 첫 번째 구성항은 분할점을 지정하기 위해 필요한 정보이며 두 번째 구성항은 어느 클래스가 상단에 해당하고 어느 정보가 하단에 해당하는 가를 결정하기위해 전송해야하는 정정 비용이다. 위의 방식을 이용하여 각 부할 포인트를 독립적으로 평가하여 각 수치적인 속성값을 재귀적인 파티션으로 분할하기 때문에 어떤 수치적이 데이터는 아주 정교하게 클러스터 되지만 반면 어떤 데이터는 엉성하게 파티션된다.

2.5.2 적합함수(Fitness Functions)

유전자 연산자를 적용하여 유전자형 수준에서 탐색기를 생성하는 동안, 표현형(phenotype) 수준에서는 탐색기 후보들에 대한 접합도를 측정한다. 효과적인 네트워크 침입탐지를 위해서는 인공면역시스템의 두가지 중요한 목표인 정확도(accuracy)와 일반성(generality)이 고려되어야 한다. 생성된 탐색기 규칙에서 정확도란 false negative error(FNE)와 false positive errors(FPE)를 고려하여 측정된다. 인체 면역시스템은 FPE를 줄이기 위해 독특한 방법, 즉 부정적 선택이라는 방법을 사용한다. 본 연구에서 사용된 복제 선택 알고리즘은 이 부정적 선택 메커니즘을 적합도 측정 과정에서 포함하여 사용한다. 일반성이란 항체 샘플링을 통한 독특한 적합도 평가 방법을 사용하여 이루어진다.

상업적인 사기 탐색(financial frauds detection)에서 가장 중요한 것으로 간주되는 명료성(intelligibility)은 복제 선택의 목적함수에서 제외되었다. 명료성은 중요한 특질이기는 하지만 본 연구에서는 정확도와 일반성이라는 측면만을 우선적으로 다루었다. 이 시스템의 명료성은 향후 풀어야할 숙제로 남아 있다. 다만 이전의 상업적인 사기 적발 연구에서 수행된 다목적함수 적합함수(multiobjective fitness function)ffm 상요하면 해결할 수 있을 것으로 생각한다. 따라서, 인공면역 시스템의 복제 선택은 생성된 탐색기 규칙의 정확도를 나타내는 하나의 접합도 함수로서 평가한다.

다수의 진화론적 학습 알고리즘은 생성된 탐색기의 적합도를 정확하게 분류한 샘플의 수를 세어서 정확도를 측정하는데 사용한다. 그런데 본 연구에서는 복제 선택 알고리즘은 적합성 평가 단계에서 항원 샘플링에 의해 완성되는 신생의 접합성 공유(emergent fitness sharing)를 통해 일반성을 유지한다. 신생의 적합성 공유(emergent fitness

sharing)은 샘플에 있어 각 세대당 서로 경쟁하는 항체중에서 에서 오직 선택된 항체만 내보내기 때문에 성취가 가능해진다. 이 경우에 있어 주어진 항원 주위의 특정 반경만을 가진 하이퍼 공간안에 있는 항체들은 지역적 경쟁 그룹을 형성하고 이 그룹에서 가장 적합한 것만이 지역적 탐색 공간을 잘 나타내는 일반적인 항체가 된다.

복제 선택에 대한 신생의 접합성 공유에 있어 중요한 점은 주어진 선정된 항원 주위의 반경을 중심으로 하는 하이퍼 구에 근거를 두고 자연tm럽게 형성되는 경쟁하는 그룹의 존재이다. 즉, 항원 탐색 공간 속에서 다차원 탐색이 후보 항체를 그룹화 함으로서 수행되다. 그룹화하는 작업은 비슷한 형태를 띄게되고 점차 같은 목표 항원을 가지게 되며 결과적으로는 하나의 경쟁하는 그룹이 된다. 따라서, 본 연구에서 사용된 복제 선택 알고리즘은 유사도 측정을 통해서 정확하게 분류된 항목을 세는 대신 정확도 특징을 제고할 수 있게 된다.

더구나 거리 측정 방식은 초기에 진화 속도를 가속화할 수 있다는 장점이 있다. 이는 탐색기들이 초기에는 무작위적인 유전자들에 의해 만들어지고 대부분의 이러한 무작위로 생성된 유전자들은 항원으로부터 멀리 떨어져 있기 때문이다. 만약 시스템이 정확하게 식별된 항원을 접합도로 사용한다면 접합도는 0이 되거나 아주 작은 값이 될 것이다. 이것은 결과적으로 진화를 최기에 천천히 진행하도록 한다. 하지만, 유사도 측정함수는 시스템으로 하여금 각 타색기가 현재의 항원으로부터 얼마나 멀리 떨어져 있는 지, 따라서 현재의 항원을 타겟으로 하는 진화과정이 탐색기가 항원을 찾기 이전이라도 더 빨리 진행될 수 있도록 알려줄 수 있다.

선정된 항원과 항체사이의 유사성을 측정하기 위한 초기의 접근방식이 유전자형 수준에서 이루어진다. 항원의 표현형(phenotype) 값은 적합한 클러스터 번호로 전환되어야 하며 그 번호는 탐색기의 유전자 타입을 정의한다. 만약 상응하는 탐색기의 유전자형 유전자 핵산성분이 On되는 경우, 즉 비트값이 1이 되는 경우, 이 한 쌍은 매치(match)가 된다. 주어진 항원의 유전자와 탐색기 유전자가 매치가 되는 경우 탐색기 유전자는 매치 거리가 0으로 된다. 하지만, 탐색기 유전자의 핵산 성분이 ON되지 되지 않고, 즉 비트값이 0으로 되어있으면, 시스템은 가장 가까운 핵산성분이 ON되어 있는지를 찾는다. 가장 가까운 핵산성분이 ON되어 있는 탐색기사이의 거리 비트를 측정하여 이 두개의 유전자 사이의 거리를 계산한다. 두 번째 단계는 매치되는 필드들을 세고, 매치되지 않는 필드돌 사이의 거리를 측정한다. 이렇게 하여 최종적인 매치 점수 M_{gb}를 A_g 항원와 탐색기 A_d 사이에서 측정할 수 있다.

$$M_{gb}(A_g, A_d) = \sum_g \left(\frac{d_{mg}}{N_{cg}} \cdot \frac{1}{N_g} \right),$$

여기서 d_{mg} g번째 항원 유전자와 g번째 탐색기 사이의 매치이다. 그리고 N_{cg} 은 g번째의 유전자를 구성하는 전체 클러스터의 숫자를, N_g 는 전체 유전자의 숫자를, 가르킨다. 여기서 각 유전자 쌍에 대하여 유전자 매치 거리는 주어진 유전자의 전체 핵산 성분의 숫자로 나누어지고, 이는 유전자 길이에 상대적인 유전자 매치 정보를 줄 수 있다.

2.5.3 유전자 연산자(Genetic Operators)

교차(crossover)와 변이(mutation)라는 두개의 유전자 연산자를 사용하여 자식 규칙이 생성되었다.

교차(Crossover)

본 연구에서 사용된 교차(crossover)는 유전자 알고리즘의 고전적인 단일 점(single-point) 교차방식이다. 이 방식에서는 부모 유전자들 사이의 임의의 두 점을 찾아내고 그 점에서 유전자형(genotype)을 겹쳐서 잇는다. 이처럼 꼬아서 만들어진 유전자형은 교차된 것으로 그 유전자 속에 잇는 이진수 역시 교차된다. 이 연산자는 모든 새로운 자식 탐색기를 만들기 위해 사용된다, 즉 교차는 200%의 확률을 가지고 만들어진다.

변이(Muttion)

복제 선택 알고리즘은 여러 가지 변이 연산자를 사용한다. 우선, 오래된 전통적인 변이를 사용하면 각 유전자속에 있는 이진수를

한 비트씩 무작위적으로 수정한다. 두 번째 변이 연산자는 일반화 변이이다. 이 변이는 보다 일반화된 자손을 만들기 위해 설계되었다. 새로운 탐색기가 보다 일반화되기 위해서는 이 연산자의 한 비트를 변화시키고 결과적으로는 표현형(phenotype)탐색기에서 하나 더 조건부를 가지게된다. 예를 돌면 일반화 연산자를 적용한 후 탐색기는

(A1 = small) and (A2 = good) 새로운 자손을 만들게 될 수 있다.
(A1 = small or medium) and (A2 = good).

이러한 변이의 효과를 자손에 대한 점진적인 변화를 하도록 하기 위해 복제 선택 알고리즘은 하나의 핵삼 성분을 선택하고 그리고 인접한 비트만 On으로 바꾼다. 모든 핵산성분이 Off인 경우 어느 핵산 성분도 On일 수 있다.

다음으로 적용되는 변이 연산자는 특성화(specialization) 변이이다. 일반화 변이와는 대조적으로 이 변이는 단순하게 한 비트를 끊어내고 한 비트가 적은 자손을 생성해낸다. 예를 들어 탐색기

(A1 = small or medium) and (A2 = good) 새로운 자손을 만들어낼 수 있다.
(A1 = small) and (A2 = good) 특성화 변이가 적용된 이후에,
이 변이가 일반화 변이와 비슷한 방법으로 적용된다. 무작위하게 선택된 핵산 성분은 만약 주변의 비트가 off 되면 Off된다. 모든 핵산 성분이 On된 경우 모든 핵산 성분은 off될 수 있다.

네 번째 변이 연산자는 이동 변이이다. 이 조작에서는 선택된 유전자의 모든 비트들이 오른 쪽으로 혹은 왼쪽으로 한 비트씩 이동을 한다. 이동 방향은 임의적으로 결정된다. 따라서 탐색자는

(A1 = small) and (A2 = good) 는 새로운 자손을 가지게 될 것이다
(A1 = medium) and (A2 = good) 이동 변이 연산자를 적용하였으므로

마지막으로 삭제 변이도 소개된다. 이 연산자는 주어진 탐색기로부터 모든 유전자를 제거하는 것을 목표로 한다. 예를 들어 탐색기는

(A1 = small) and (A2 = good) 자손을 하나 가질수 있다.
(A1 = small) 삭제 변이가 수행된 이후

이 변이는 모든 핵산 성분을 On 시키고 결국 삭제 연산자가 표현형(phenotype)으로 매핑되는 경우 유전자를 삭제 시킬 수 있다. 이 변이는 모든 핵산 성분을 On시키는 경우, 따라서 결과적으로 탐색기를 표현형으로 매핑하는 경우 가능하다. 이 변이는 특징 선발 연산을 위해서 개발되었다. 후보 특징들이 훈련용 세트 속에 포함되어 있으므로 가끔 불필요한 요소들도 포함되어 있다. 이 변이에서는 탐색 방법을 갑자기 멋게할 수도 있으며 복제 선택 알고리즘은 이러한 유전자들을 필터링하기 위해 제거 변이를 상요하기도 한다.

<u>다른 연산자들</u>
복제 선택 알고리즘은 모집단 중첩(overlapping)를 사용하수 있으며, 이 경우 모집단의 제일 나쁜 *n%*는 가장 좋은 *m%*로부터 만들어진 새로운 자손에 의해 대치될 수 있다. 대표적으로 n=80, m=40으로 할 경우 좋은 결과를 얻을 수 있다. 모집단 크기는 보통 100과 200 사이가 된다.

3. 결론(Conclusion)

이 연구에서는 네트워크 기반의 IDS 기술에 대해 조사가 이루어졌으며 문헌 조사를 통해 일련의 일반적인 요구사항을 정리하였다. 이러한 요구조건을 바탕으로 세 개의 중요한 설계 목표가 제시되었다. 단순화된 인체 면역 시스템에 대한 소개와 더불어 경쟁력 있는 네트워크 기반의 외부침입 탐지 시스템을 구축하기 위해서 필요한 중요한 특징돌을 찾아내고 이를 기술하였다. 이러한 분석작업에서 인체 면역 시스템은 몇 가지 정교한 메커니즘으로 이루어져 있음을 이해할 수 있었다. 이들 정교한 메커니즘은 3가지 요구조건을 만족시키는 것으로 이해되었다. 결과적으로, 인체면역학에 기반을 둔 새로운 네트워크 환경의 IDS 설계는 상당히 유망하다는 것을 확인하였다.

본 연구에서는 현재의 네트워크 환경에서의 IDS에 대한 기술조사가 이루어졌다. 이들은 단일구조, 수직적 구조, 협력적 구조를 갖는

3가지 접근 방법으로 구분되었으며 각 접근 방법들에 대한 문제점을 기술하였다. 이러한 문제점들을 해결하기 위해 새로운 인체면역 시스템이 제안되었다. 이 모델에서는 3가지 순차적인 단계를 거쳐 진화가 이루어진다: 유전자 라이브러리 진화, 부정적 선택, 그리고 복제 선택. 이러한 3개의 프로세스는 네트워크 상에서 경쟁력있는 IDS를 구축할 수 있는 3가지 목표, 즉 분산적일 것, 자기 조직적일 것, 그리고 경량성일 것이라는 요구조건을 만족시킨다. 이같은 통일된 진화적 접근방법의 성격을 분석해 보면 현재의 접근 방법들과는 달리 네트워크 IDS가 충족시켜야할 요구조건들을 만족시킨다. 결과적으로 이러한 새로운 모델에 기반을 두고 있는 알고리즘은 향후 IDS개발을 위해 상당한 장래성을 갖고 있다.

인체면역 시스템을 사용한 네트워크 기반의 IDS이 개발되고 있다. 현재의 작업에서는 초기의 자기 프로파일을 구축하여 자기(self)세포와 비자기(non-self) 세포를 구분하는데 초점을 맞추고 있다. 이러한 기능능 실제적인 네트워크 상황에서 수집된 TCP/IP패킷을 정상적인 패킷과 비정상적인 패킷으로 구분하는 작업이 이루어졌다. 이러한 노력의 첫 번째 시도로서 부정적 선택의 단계가 구현되어 이 알고리즘을 실제 네트워크에서 적용할 수있는 지에 대한 실험이 이루어졌다. 그 결과로 본 연구에서는 인체면역 전체를 속에서 부정적 선택에 대한 역할에 대해 새롭게 정의되었다. 최종적으로 복제 선택 단계의 외부침입시스템에 대한 조사가 이루어졌고, 복제 선택단계에 대한 명확한 이해는 부정적 선택에 대한 확실한 이해를 할 수 있도록 되었다.

이연구의 공헌은 실제적으로 분산되어 있는 인공면역 시스템을 이용하여 경쟁력 있는 IDS를 설계하는 데 필요한 방법론을 제시하였다는 점이다. 본 연구에서 제시된 방법론은 분산적이며, 자기 조직적이며, 경량성을 가진 IDS라는 점에서 네트워크 환경에서의 진정한 IDS의 요구조건을 만족시키다.

4. 참고문헌

(Balasubramaniyan et al., 1997) Balasubramaniyan, J. S. et al., "Software Agents for Intrusion Detection", Department of Computer Sciences, Purdue University, 1997. available at http://www.cs.purdue.edu/coast/coast-library.html

(De Jong et al., 1993) De Jong, K. A., Spears, W. M., and Gordon, D. F., "Using Genetic Algorithms for Concept Learning", *Machine Learning*, Vol.13, No.2/3, pp.161-188, 1993

(Dhaeseleer et al., 1997) Dhaeseleer, P. et al, "A Distributed Approach to Anomaly Detection", *ACM Transactions on Information System Security*., 1997. Available at http://www.cs.unm.edu/~patrik

(Fayyad and Irani, 1993) Fayyad, U. M, and Irani, K. B., Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, *Proceeding of The Thirteenth International Joint Conference on Artificial Intelligence*, pp.1022-1027, 1993.

(Goldberg, 1989) Goldberg, D. E., *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley, 1989.

(Mykerjee et al, 1994) Mykerjee, B.; Heberlein, L. T.; Levitt, K. N., "Network Intrusion Detection", *IEEE Network*, Vol.8, No.3, pp.26-41, 1994.

(Paul, 1993) Paul, W. E., "The Immune System: An Introduction", *Fundamental Immunology* 3rd Ed., W. E. Paul (Ed), Raven Press Ltd.

(Playfair, 1996) Playfair, J. H. L., *Immunology at a Glance*, 6th Ed, Blackwell Science, 1996

(Porras and Valdes, 1998) Porras, P. A.; Valdes, A., "Live Traffic Analysis of TCP/IP Gateways", *Proceeding of ISOC Symposium of Network and Distributed System security, 1998.*
Available at http://www2.csl.sri.com/emerald/downloads.html

(Potter, 1997) Potter, M. A., *The design and analysis of a computational model of cooperative co-evolution*, PhD Thesis, George Mason University, Fairfax, VI., 1997.

(Roitt and Brostoff, 1998) Roitt, I., Brostoff, J., and Male, D., *Immunology*, Fifth Ed., Mosby International Ltd, 1998.

(Smith, Forrest and Perelson, 1993) Smith, R. E., Forrest, S., and Perelson, A. S., "Searching for Diverse, Cooperative Populations with Genetic Algorithm", *Evolutionary Computation*, 1(2), 127-149, 1993.

(Somayaji et al., 1997) Somayaji, A.; Hofmeyr, S.; Forrest, S., 1997, "Principles of a Computer Immune System", *Proceeding of New Security Paradigms Workshop, Langdale, Cumbria*, pp.75-82, 1997.

(Tizard, 1995) Tizard, I. R., *Immunology: Introduction*, 4th Ed, Saunders College Publishing, 1995.