2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

유전자 알고리즘을 이용한 침입 패턴 인식에 관한 연구

임명현*, 김두현*, 김원필*, 김판구*
*조선대학교 전자계산학과
e-mail:love95@cspost.chosun.ac.kr

A Study on Intrusion Pattern Recognition Using Genetic Algorithm

Myung-Hyun Lim, Doo-Hyun Kim, Won-Pil Kim, Pan-Koo Kim
*Dept of Computer Science, Chosun University

요약

침입탐지시스템에 대한 수많은 방법들이 제시되고 있지만, 가장 핵심적이라고 할 수 있는 침입에 대한 패턴을 어떻게 정의하고 탐지해 낼 것이며, 알려지지 않은 비정상적인 행위에 대한 패턴을 탐지하는 것에 대한 연구가 미흡하다. 이에 본 논문에서는 알려지지 않은 침입 행위를 판별하는데 있어서 유한오토마타를 이용하여 시스템 콜에 대한 패턴을 정의하고, 정의된 패턴을 유전자 알고리즘을 이용하여 비정상적인 침입 행위를 판별할 수 있도록 새로운 패턴들을 유전자 조작을 통하여 생성하여 알려지지 않은 침입 패턴에 대해서도 침입탐지시스템에 이를 적용하여 침입을 탐지할 수 있는 방안에 대해 연구하였다.

1. 서론

정보화 사회에 즈음해서 우리는 정보의 중요성을 새삼 느끼고 있다. 크랙커의 침입으로 수많은 정보들이 개인의 의사와 무관하게 공개되고 있다. 이는 개인의 문제뿐만 아니라 사회의 문제로까지 야기되고 있다. 또한 크랙커의 침입이 있었는지, 어떤 행위를 했는지를 일반 관리자들은 분석하기 힘들 정도로 더 치밀해 지고 있고, 알려지지 않은 침입방법들이 난재해 있다. 침입탐지시스템이 많이 개발되고 있지만, 침입이 아님에도 침입임을 통보한다든지 침입이 일어나고 있음에도 침입패턴이 분석되지 않아 침입 경보를 알리지 않는 경우가 비일비재하다. 이에 본 논문에서는 침입을 판단함에 있어 정확성을 기하고, 전문가가 정해준 침입 패턴뿐만 아니라 알려지지 않은 패턴들에 대해서도 침입패턴을 찾아 낼 수 있는 방법을 제안하고자 한다.

논문의 구성은 다음과 같다. 제2절에서는 기존의 침입탐지시스템의 유형 및 유전자알고리즘에 대해 설명하고, 제3절에서는 System Call Traces에 대한 내용과 System Call Traces를 통해 생성된 예와 이를 유전자 알고리즘을 통해 정상적인 행위 패턴을 능동적으로 생성하고 선택하는 방법을 제시한다. 제4절에서는 본 연구를 통해 얻은 효과와 침입탐지시스템을 구축하는데 있어서 유전자 알고리즘의 효율적인 응용방법 그리고 유전자 알고리즘을 다른 유사 분야에 응용하기 위한 방안에 대해 논한다.

2. 기존의 침입 탐지 기법

침입 탐지 시스템에 있어 핵심기술은 행위 판별(Behavior Classification)과 자료축소(Data Reduction) 기술이다[1]. 행위판별은 주어진 일련의 행위들에 대해 이것이 침입인지 아닌지를 판단하는 문제이고, 자료축소는 수 메가바이트에 이르는 방대한 양의 데이터를 의미있는 소규모의 자료집단으로 줄여나가는 과정이다.
대체적으로 침입탐지에 있어서는 규칙기반시스템(Rule-based System)과 신경망 또는 통계적 분류 시스템의 방법을 도입하여 사용하고 있다[2]. 기존에 사용된 규칙기반 시스템이나 신경망, 통계적인 분류시스템은 많은 양의 초기 학습을 필요로 하며, 시스템의 유지보수를 위해 많은 노력이 필요하다는 문제점 외에도 새로운 공격 유형에는 약하다는 단점이 있다. 행위판별은 패턴 매칭에 의존하고 있는데, 이는 패턴 자체가 완전하지 못할 경우 시스템의 방어에 커다란 허점을 나타나게 된다. 데이터의 축소 문제는 시스템 상에서 이루어지는

모든 행동들에 대해 감사 정보를 만들어야 하는데, 그 양이 엄청나며, 분석작업에는 시스템의 디스크 용량이나 CPU에 엄청난 부하를 가하게 된다. 이 문제를 해결하기 위해 영국의 University College London(UCL)에서는 기존의 신경망 기법과 유전자 알고리즘(GA), 그리고 전문가 시스템 기술을 소규모 시제품 시스템에 적용하고 있다[3].

[표 1] 침입탐지 시스템의 방법론의 비교 분석

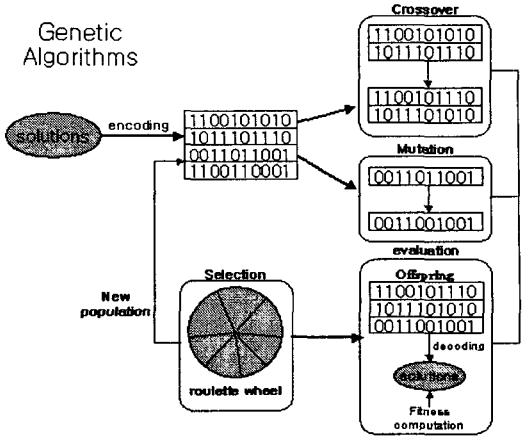
기존의 침입탐지 시스템의 방법	유전자 알고리즘을 이용한 침입탐지 시스템의 방법
• 시스템 콜의 집합을 사용	• DFA로 정의된 시스템 콜 패턴의 시작점으로 이루어진 유전자 집합을 이용
• 시스템 콜의 범위(거리 : Distance)를 지정해 주어야 한다.	• 시스템 콜의 범위를 지정하지 않아도 된다.
• 시스템 콜의 분포에 대한 정보를 정량적인 방법을 사용해 얻는다.	• 적합도를 계산한 후 유전자 조작을 통해 얻어진 패턴과 비정상적인 행위와의 일치성을 찾는다.
• 패턴의 일치성을 검사하는데 빠른 반면 정확성이 떨어진다.	• 기존의 방법에 비해 느리지만 침입패턴의 일치성에 대한 탐색이 정확하다.

3. 유전자 알고리즘을 이용한 새로운 침입 패턴 생성

3.1 유전자 알고리즘

유전자 알고리즘은 자연도태와 자연적 유전현상의 원리와 기술에 기초한 적응적탐색(Adaptive-Search) 기법으로서 1975년 Holland에 의해 처음 제안되었다[9]. 유전자 알고리즘은 복잡한 수식을 요구하지 않으며 탐색 공간과 같은 한정적인 요소에 의해 제한받지 않는다. 유전자 알고리즘은 후보 해(Candidate solution)들의 고정된 크기 모집단(Population)을 반복적으로 처리하며 각 후보 해는 생물학적 체계에서 유추한 염색체(Chromosome)라 불리는 고정된 크기의 문자열에 의해 표현된다. 각 염색체는 목적 함수(objective function, 적합도 함수 : fitness function)에 의해 평가된 적합도 값(fitness value)을 가지고 있으며, 한 염색체의 적합도는 그들이 생존하고 자식을 생산할 수 있는 능력을 결정한다. 즉, 적합도 이상의 값을 갖는 개체는 자손을 만들기 위해 살아남지만 그렇지 못한 개체는 도태된다. 각 염색체는 유전자(gene)의 일련으로 구성되어 있으며, 일반적으로 유전자 알고리즘에서는 비트(bit)들의 일련으로 표현되지만 정수나 실수들도 사용할 수 있다.

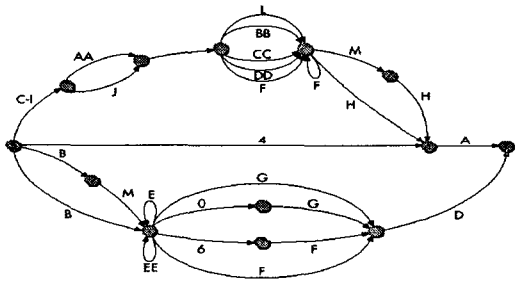
유전자 알고리즘에 대한 전체적인 작업순서를 도식화하면 그림 1과 같이 표현할 수 있다[4,5].



[그림 1] 유전자 알고리즘의 전체적인 구조

3.2 DFA로 정의한 사용자 정상행위 패턴 설계

그림 2는 정상적인 행위를 전문가의 실험을 토대로 하기 위해 New Mexico 대학에서 System Call Traces에 대한 내용을 유한오토마타(이하 DFA)로 나타내었다[6].



[그림 2] SendMail를 DFA로 추론한 정상적인 System Call Traces에 대한 매크로 패턴

이를 다시 패턴 정규화를 통해 ($2^{2^{37}}$)개의 패턴을 생성하였다. 패턴의 규칙을 표현하면 그림 3과 같다.

[표 2] Traces된 내용을 의미를 부여해서 정의한 Macro

A : 8 8 8 8 8 1

B : 108 107 107 109 108 107 107 109 108 107 107 109 8 7 8 8 43 43 44 108 107 107 109 64 8 88 53 30 21 53 30

C : 7 5 69 69 7 141 69 8 26 48 7 30 69 8 7 5 69 69 8 7 5 69 69 8 8 88

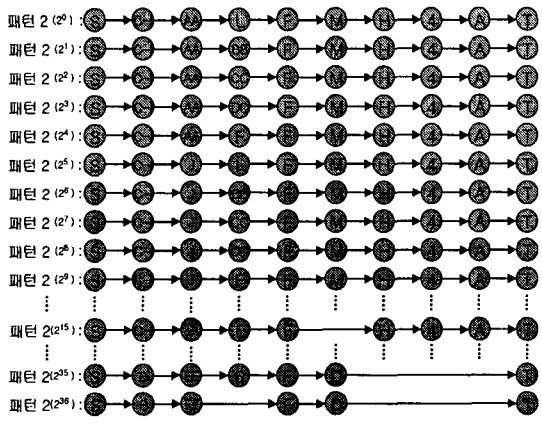
⋮

X : 7 53 30 0 5 8 5 8 R 0

Y : 8 9 115 115 22 131 44 108 107 107 109 64 853 30 21

3.3 새로운 정상 패턴에 대한 유전자 알고리즘의 적용 과정

유전자 알고리즘을 적용하기 위해 다음과 같이 초기 개체집단, 즉 초기 유전자 코드들의 집합을 생성할 수 있다. 염색체를 다음과 같이 DFA에 대한 패턴을 분류하면 그림3과 같다. 이를 유전자의 형태인 이진화를 통해 본 본문에서 요구하는 새로운 패턴을 생성하고자 한다. 이진화로 표현된 각각의 Gene은 시스템 콜에 대한 특성을 가지고 있다. 이를 Locus라 한다. 단지 패턴의 일정한 규칙을 Gene들의 집합들로 정의하였다.



[S : Start , T: Terminal]

[그림 3] 정상 패턴 분류

위의 염색체를 유전자 알고리즘에 적용하기 위해 이진수로 표현하면 다음과 같다.

패턴 2 (2^0) : 000000000000000 00000000000000000
패턴 2 (2^1) : 000000000000000 00000000000000010
패턴 2 (2^2) : 000000000000000 00000000000000100
패턴 2 (2^3) : 000000000000000 00000000000001000
패턴 2 (2^4) : 000000000000000 00000000000010000
패턴 2 (2^5) : 000000000000000 00000000000100000
패턴 2 (2^6) : 000000000000000 00000000001000000
패턴 2 (2^7) : 000000000000000 00000000010000000
패턴 2 (2^8) : 000000000000000 00000000100000000
패턴 2 (2^9) : 000000000000000 00000001000000000
⋮
패턴 2(2^15) : 000000000000000 00100000000000000
⋮
패턴 2(2^35) : 010000000000000 00000000000000000
패턴 2(2^36) : 100000000000000 00000000000000000

[그림 4] 정상 패턴 염색체를 이진수로 표현

3.3.1 패턴 구성을 위한 클러스터링 검사 방법

유전자 조작을 통해 새롭게 생성된 집합이 본 논문에서 정의한 패턴과 일치하는지를 검사하기 전에 매크로로 표현된 Gene들이 클러스터링을 통해 서로 관계를 잘 맺었는지 그렇지 않은지를 검사를 해야한다. 다음은 그림 2의 DFA에 대해 새로운 침입 행위와 적합도를 검증하기 위해 클러스터링을 위한 적합도 함수를 다음과 같이 정의하였다[7]. 이를 정의함으로써 각 유전자 집합이 로그(log)로부터 하나의 패턴을 생성하여 그림 3에서 분류한 패턴들과의 비교를 통해 정상패턴과 비정상패턴을 판별할 수 있다

$$Energy(g) = \alpha \times \frac{dist(g)}{avedist} + \beta \times length(g)$$

$$fitness(g) = 1 - \frac{1}{1 + energy(g)}$$

※ **dist(g)** : 클러스터의 중심들의 리스트를 g라고 하고 이 중심들을 가지고 각각의 입력 값이 어느 클러스터에 속하는지를 결정했을 때 각각의 입력 값이 속하는 중심과 그 입력 값과의 거리를 평균한다.

※ **length(g)** : 유전자 g의 길이를 나타낸다.

※ **avedist** : 입력 값들 사이의 평균거리를 나타내는 값

위의 식에서 α와 β값이 너무 크거나 작게 주어진다면 패턴을 너무 많이 발생하거나 적게 발생 될 수가 있기 때문에 α와 β값을 유동적으로 조정을 하여야한다. 본 논문에서는 α와 β값을 각각 0.5와 0.3으로 하였다.

최적의 패턴을 구성하기 위해 개체들을 진화시키는 절차는 다음과 같다.

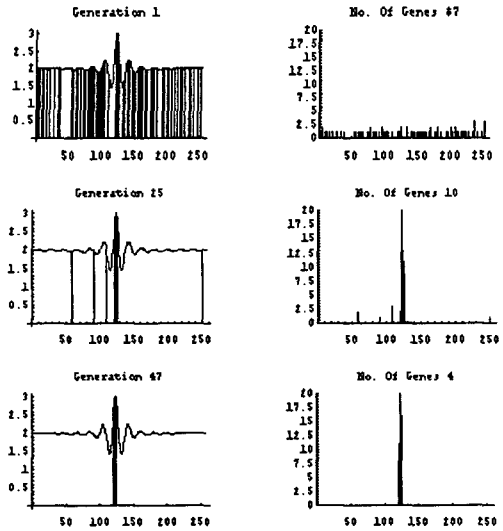
Step1 : 임의의 초기 개체군, 즉 후보 개체(가중치 변종)들의 패턴을 생성하고 각 패턴에 대해 적합도 값을 계산한다

Step2 : 기존의 패턴에서 적합도값에 따라 적합한 개체들을 선택해서 다음 세대의 패턴을 형성한다.

Step3 : 돌연변이와 교배를 통해 새로 선택된 패턴들의 염색체를 변화시킨다.

Step4 : 각 패턴의 적합도 값을 계산하여 최적합(best-fit) 패턴을 해로서 기록한다. **Step2**로 돌아가서 반복한다.

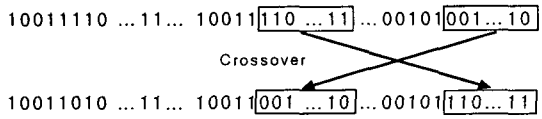
위의 절차를 통해 다음과 같이 세대 변이를 거쳐 패턴의 적합성을 판정할 수 있다. 개체의 수는 100개로 하였으며, 각 개체의 크기는 2^{37}bit, 교차율은 8회, 돌연변이 율은 0.01로, 세대수는 50회로 하였다. 다음은 유전자들이 세대 변이 과정을 통해 패턴을 수렴하는 과정을 표현한 것이다.



초기의 모 집합들이 아래의 그림처럼 흩어져 있던 유전자들이 세대변이를 거치면서 한곳에 수렴하는 것을 볼 수 있다. 이를 통해 패턴의 적합성을 검증할 수 있다.

(2) Crossover를 통한 새로운 침입 패턴에 대한 탐지 방법 제안

알려지지 않은 새로운 침입 패턴에 대한 탐지 방법으로 Crossover를 이용하여 정상 패턴에서 많은 양의 패턴을 생성한다. 돌연변이는 염색체 비트열 중에서 임의의 위치의 비트를 변경시킴으로써 이루어지고, 돌연변이가 일어나는지의 여부는 주어진 확률에 따라 결정된다. 교배는 임의의 교배점 다음에 위치하는 비트 열을 다른 염색체의 비트 열과 교환함으로써 이루어지며, 교배하는지의 여부도 주어진 확률에 따라 결정된다. 이를 통하여 알려지지 않은 새로운 패턴이 생성되어 진다.



[그림 5] Crossover를 통해 새로운 정상 패턴이 생성되어지는 과정

그림 4에서 $2^{2^{36}}$의 패턴에서 정의되어지지 않은 행위에 대한 패턴을 그림 5의 Crossover를 통해 $(2^{2^{36}})'$을 생성한다. Crossover로 생성되어진 유전자 집합 즉, 패턴들도 정상 행위에서 발생되어진 패턴이기 때문에 정상 패턴으로 간주할 수 있다. 이는 유전자적 성질을 통해서 증명할 수 있다[1]. 다시 말하면, Crossover를 하더라도 정상 집합의 범위를 벗어나지 않고, 정상 패턴에 대한 성질을 가지고 있으므로 이를 정상 패턴으로 간주하게 된다.

4. 결론 및 향후과제

본 논문에서는 침입에 대한 행위 즉, 시스템 콜에 대한 패턴을 정의하였다. 정의된 패턴을 유전자 알고리즘을 통해 검증된 Crossover를 이용하여 새로운 형태의 침입 판정을 위한 패턴을 생성하였다. 새로이 생성된 정상패턴을 $2^{2^{37}}$개 생성하여 이를 벗어나면 침입으로 간주하고, 이를 관리자에게 통보함으로써 침입의 유무를 판정하게 된다. 이는 전형적인 침입탐지시스템에서 학습되어지는 패턴들은 전문가나 다른 환경으로부터 주어짐에 반해 본 논문에서 제안하는 방법은 인공지능의 관점에서 접근하였다. 유전자 알고리즘을 이용한 시스템은 능동적으로 패턴 예(Pattern examples)를 탐색함으로써 긍정적인 결함(False Positive) 및 부정적 결함(False Negative)을 줄이고자 하였다. 이는 유전 연산자에 의한 전역적 탐색을 통해 수행된다.

많은 논문들에서 침입 탐지에 대한 효율성 향상을 위해 많은 노력을 해왔다[2,3,6]. 본 논문을 통해 유전자 조작을 통해 정의된 패턴들이 침입탐지시스템에 적용시켰을 때 침입에 대한 판정을 사용자의 요구에 맞게 얼마나 효율적으로 탐지해줄 것인가에 대한 검증에 관한 연구가 앞으로 수행되어야 할 것이다.

참고문헌

[1] D. Goldberg "Genetic Algorithm in Search, Optimization and Machine Learning" Addision Wesley 1989

[2] Kumar S. and Spafford G. "A Pattern Matching model for Misuse Intrusion Detection" In Proceedings of the 17th National Computer Security Conference. October. pp11~21. 1994

[3] 최종욱 "면역학 기반의 외부 침입 탐지 시스템" 정보처리학회지 제7권 제2호 2000. 3

[4] M. Gen and R. W. Cheng 'Genetic Algorithms and Engineering Design' John Wiley and Sons New York 1997.

[5] Z.Michalewicz 'Genetic Algorithms + Data Structures = Evolution Programs' second edition SpringerVerlag New York 1994.

[6] Andrew P. Kosoresow and Steven A. Hofmeyr "Intrusion Detection via System Call Traces" IEEE Computer Society 1997

[7] 김남훈, 문성광, 박세진, 이인 '유전자 알고리즘의 이해와 구현' 프로그램세계 1997. 7