시스템 콜 서열과 유전자 알고리즘을 이용한 침입탐지 기법

김신재° 위규범
아주대학교 정보통신전문대학원
{venddol8°, kbwee}@ajou.ac.kr

Intrusion Detection using System Call Sequences and Genetic Algorithms

Sinjae Kim° Kyubum Wee
Graduate School of Information and Communication, Ajou University

요 약

시스템 콜 서열(system call sequence)을 기반으로 한 침입 탐지 기법에는 다양한 알고리즘들이 사용되어 왔으며,시스템 콜 서열을 인식하는 오토마타를 생성하는 기법도 많이 연구되었다. 그러나 효율적인 오토마타를 생성하는 것은 계산복잡도가 높은 어려운 작업이다. 본 논문에서는 유전자 알고리즘을 사용하여 자동적으로 오토마타가 생성되는 과정을 설명하며, 생성된 오토마타가 침입탐지에 효과적으로 이용될 수 있음을 실험을 통하여 보인다.

1. 서 론

침입은 시스템 자원에 대하여 비인가된 사용자가 무결성, 기밀성 또는 가용성을 훼손하는 일련의 행위를 말한다. 침입 탐지 시스템(IDS : intrusion detection system)은 대상 시스템에 대한 비인가 된, 비정상적인 행동을 탐지, 구별하고 이에 대응하는 기능을 가진 시스템이다. 침입 탐지 시스템은 탐지 기법에 따라 오용 탐지 기법(misuse detection)과 비정상 행위 탐지 기법(anomaly detection)으로 나뉜다. 비정상 행위 탐지 기법은 사용자의 정상적인 행위를 모델링 한 후에 그 모델을 가지고 사용자의 행위를 분석하여 정상적인 행위에서 얼마나 벗어 났는지를 측정하여 주어진 임계값(threshold)을 넘으면 그 행위를 침입으로 판단한다. 이런 비정상 행위 탐지 기법의 장점은 새로운 형태의 침입 행위를 탐지 할 수 있으며 단점으로는 구현이 어렵고 오용 탐지 가법에 비해 오탐지율(false positive)이 높다는 것이다. 비정상 행위 탐지 기법에 있어서 중요한 부분은 정상 행위에 대하여 모델링하는 방법이다.

침입 탐지 시스템은 주로 특권이 있는 프로세스의 시스템 콜을 감사 자료로 모아 모델링한것을 사용한다. 시스템 콜을 사용하는 방법에는 데이터 마이닝(data mining), 시스템 콜의 빈도수(frequencey-based methods), 유한상태기계(finite state automata)를 이용하는 방법들이 있다[1,2].

유한상태기계는 프로그램의 시퀀스를 인식할 수 있는 오토마타를 만드는 방법이다. 유한상태기계의 장점은 입력되는 데이터에 대한 각 시스템 콜이 일정 시간을 유지하며 검사되고 유한한 저장 공간에 비해 많은 양의 데이터를 처리할 수 있다. 또한 상태들간의 변화가 루프나 분기를 이룰 수 있기 때문에 새로운 패턴에 대해서도 처리가 가능하다.

본 논문에서는 유한상태기계를 자동적으로 생성 할 수 있는 유전자 알고리즘을 제안하고 생성된 오토마타가 침입탐지에 효율적으로 이용될 수 있음을 실험을 통하여 보인다.

2. 관련 연구

2.1. 유한상태기계 (FSA : finite state automata)

유한상태기계는 5개의 표기형식으로 구성된다.

(S,I,f,x_0,F)

S는 유한한 상태의 집합, I는 입력장치로부터 들어오는 데이터를 표현 할 수 있는 심볼의 집합, f는 상태간의 변화를 시켜주는 함수, x_0는 초기의 상태이며 S의 원소이고, F는 최종 상태 집합이며 S의 부분집합이다[3]. 유한상태기계의 처음 시작 상태가 x_0로 시작되며 입력 장치로부터 차례대로 데이터가 입력 되면 오토마타의 상태가 천이 되고 더 이상의 입력할 데이터가 없을때 마지막 상태가 집합 F에 속하면 이 데이터를 승인(accept)하고 그렇지 않으면 기각(reject)한다

표1. 입력 데이터에 대한 상태 배열.

	S0	S1	S2	S3	S4	S5	S6	S7	F
a	S1	-1	-1	-1	S5	-1	S7	-1	-1
b	-1	S2	-1	-1	S6	S6	-1	F	-1
c	S7	S3	-1	-1	-1	-1	S1	-1	-1
d	-1	-1	S4	S4	-1	-1	-1	-1	-1

표1은 심볼 a,b,c,d에 대한 각 상태에서의 변화 될 수 있는 상태들로 구성되며 상태 변화가 생길 수 없는 경우에는 -1로 구성된다. 이 상태 배열를 통해 그림1처럼 유한상태기계를 구성하여 입력된 데이터의 마지막 상태가 F가 된다면 입력 데이터를 승인하고 그렇지 않으면 기각한다. 논문에서 사용될 상태 배열은 상태들간의 싸이클 형성을 허용하며 유효한 상태로의 전이 비율과 유효하지 않은 상태로의 전이 비율을 동등하게 하여 실험한다. 즉, '-1'로의 전이 확률과 S의 원소에 대한 전이 확률이 같다[4].

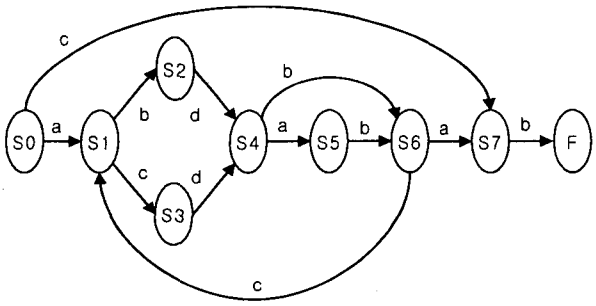

그림1. 표1에 대해 생성된 유한상태기계

2.2. 유전자 알고리즘(GA:Genetic Algorithms)

유전자 알고리즘은 생물진화의 원리로부터 착안된 알고리즘으로서, 확률적 탐색이나 학습 및 최적화를 위한 한 가지 기법이다. 일반적인 유전자 연산(genetic operations)을 살펴 보면 가능한 개체들을 여러 개 생성하여 군집(population)을 형성하고, 개체들을 선택(selection)하여 교배(crossover)하고 돌연변이(mutation)을 적용하여 더 좋은 후보해들의 군집으로 진화해 나아간다[5].

2.2.1. 군집(population)

초기 군집의 개체들은 임의적인 길이를 갖고 있는 염색체로 구성된다. 개체들은 그림2에서와 같이 격자 모양으로 배치된다. 본 논문의 실험에서 초기군집의 개체들을 size(개체)-n<...<size(개체)+n의 크기로 초기화 한다. n의 크기는 5이하의 임의적인 값이며 개체의 크기는 20이다.

2.2.2. 적합도 평가(fitness evolution)

진화 과정 속에서 보다 좋은 성질의 개체들로 군집을 형성하기 위해서는 각각의 개체에 대한 적합도가 평가되야 한다. 적합도를 평가하는 방법에 있어서 크게 세가지 사항을 고려한다. 첫째, 견고성(consistency)은 개체의 염색체 즉, 상태들이 실험 데이터 집합을 어느 정도 수용할 수 있는지를 평가한다. 둘째, 최소화(smallness)는 개체의 크기 즉, 상태의 수를 나타내며 상태의 수가 적을수록 오토마타가 최적화 되어 졌다고 볼 수 있다. 마지막으로 일반화(generalization)는 실험외의 데이터 집합에 대하여 정상의 데이터와 비정상의 데이터를 어느정도 구분 할수 있는지를 평가한다.

2.2.3. 선택(selection)

다음 세대의 군집을 이루는 개체들을 만들기 위하여 현재 세대의 군집에서 개체를 뽑아내는 과정이다. 좋은 성질의 부모 개체들이 많이 선택 되어질 수록 자손 개체들의 성질이 우수하다. 본 논문에서 사용하는 방식은 하나의 부모 개체를 임의적으로 선정하며 선정된 개체의 8근방의 이웃 중 적합도가 가장 높은 개체를 또다른 부모 개체로 선정하는 방식을 사용했다.

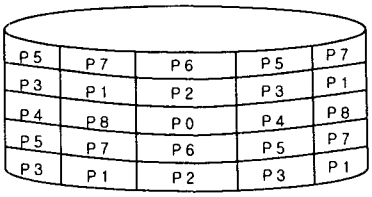

그림2. 군집의 개체들의 배치 형태

그림2에서 P0를 하나의 부모 개체로 선정하였다면 8근방의 이웃{ p1,p2,p3,p4,p5,p6,p7,p8 } 중 적합도가 가장 높은 개체를 다른 부모 개체로 선정한다.

2.2.4. 교배(crossover)

선택된 부모 개체들로부터 새로운 개체를 생성한다. 교배 지점에 따라 단순, 복수점, 일정(uniform)교배로 구분되어 지며, 일정교배는 선택되어진 두 부모에서 우수한 염색체들만 뽑아 자손에게 전달한다. 본 논문에서는 일정교배를 사용하여 각 상태의 적합도가 높은 것만 자손에게 전달한다.

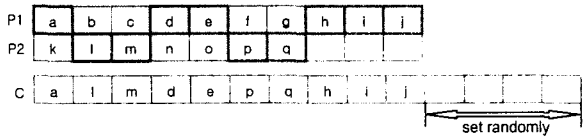

그림3. 일정교배(uniform crossover)

일정교배가 이루어지기 전에 선행되어야 할 것은 자손의 염색체 길이를 결정하는 문제이다. 자손의 염색체의 길이의 범위는 [size(Pmax)-n ... size(Pmax)+n] 사이의 값을 가진다. Pmax는 선택되어진 부모 개체중 상태의 크기가 큰 개체를 말하며 n의 값은 2이다. 그림3과 같이 부모의 길이 보다 생성되 자손의 길이가 크다면 임의적인 염색체 정보를 담는다.

2.2.5. 돌연변이(mutation)

새로운 자손 개체의 일부 염색체의 정보를 변화 시킨다. 돌연변이가 없는 경우에는 초기 개체들의 조합 이외의 기대치를 가질 수 없기 때문에 최적화된 값을 얻기 위해서는 작은 변이확률을 가져야 한다. 실험에서는 교환 변이(swap mutation)를 사용하여 임의적으로 선택된 두 부분의 유전자를 교환한다.

2.2.6. 세대모델(generation model)

일반적으로 유전자 알고리즘에서는 모든 개체가 일제히 자손을 만들고, 다음 세대의 집합을 만드는 이산세대 모델(discrete generation)과 연속세대 모델(contiguous generation model)을 사용한다. 본 논문에서는 연속세대 모델 방법 중 다음 세대로 넘어갈 때 2개체만을 선택하고 두개 자손의 개체를 만든어 적합도가 낮은 개체를 두개 제거하여 새로운 군집을 형성하는 정상상태(steady-state model)을 사용하여 실험한다.

3. 데이터 집합 (Data sets)

침입 탐지를 위한 유한상태기계를 자동으로 생성하기 위해서는 일관성있는 시스템 콜 데이터를 사용해야 한다

[6,7]. Forrest등의 연구에서 사용한 데이터를 이용하였다. 사용한 데이터의 집합은 표2와 같다.

표2. 실험에 사용된 데이터 집합

Program	Normal Data Sets			Intrusion Data Sets		
	Number of System call	Number of unique System calls	Number of Sequences	Number of System call	Number of unique System calls	Number of Sequences
syn lpr	2,399	37	9	164,233	37	1001
ps	6144	22	19	6970	22	26
inetd	541	35	3	6372	35	31
login	8906	46	24	4857	46	13
named	1800	41	5	1800	41	5
sendmail	19527	47	147	4878	47	14

4. 실 험

3장에서 제시한 데이터 집합을 기반으로 자동화된 유한상태기계를 생성할 수 있는 유전자 알고리즘을 설계하여 실험 하였다. 첫 번째 실험은 적합도의 견고성(consistency)과 최소화(smallness)의 값들이 최적의 값을 얻도록 세대 교체를 실험 하였다. 실험 결과는 아래의 그림과 같다.

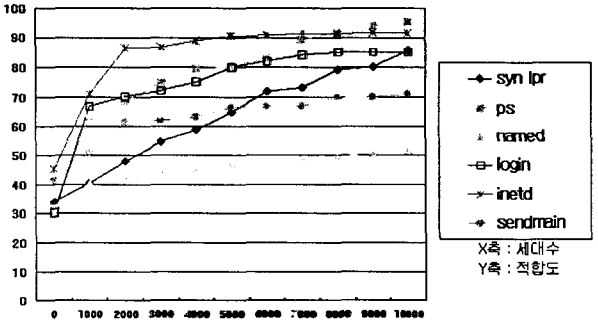

그림4. 실험 결과의 적합도

그림4에서 알수 있는 것은 데이터 집합을 구성하고 있는 시퀀스의 개수와 각각의 시퀀스들을 이루고 있는 시스템 콜의 길이에 따라서 그 결과가 다르다는 것이다. named의 경우 시스템 콜의 개수에 비해 시퀀스의 길이가 길었기 때문에 보다 좋은 적합도의 값을 얻기 위해서는 다른 데이터 집합 보다 더 많은 진화 과정이 필요하다. ps의 경우는 전체 시스템 콜 개수에 비해 많은 시퀀스가 존재하여 다른 데이터 집합 보다 더욱 빠른 시간에 적합도가 좋은 값을 얻을 수 있었다. 다음 실험은 10,000번의 세대를 통해 얻은 적합도 값을 이용한다.

표3. 성능평가 계측량

TP(True Positive) : 악성을 악성으로 판단
TN(True Negative) : 정상을 정상으로 판단
FP(False Positive) : 정상을 악성으로 판단
FN(False Negative) : 악성을 정상으로 판단
Detection rate = TP / (TP+FN)
False positive rate = FP / (TN+FP)

표3에서 제시하는 성능평가에 따라 유전자 알고리즘을 사용하여 생성된 유한상태기계의 성능평가를 실시한 결과는 표4와 같다.

표4. 각 프로그램에 대한 실험 결과

	True Positive	True Negative	False Positive	False Negative	Detection Rate	False Positive rate
Syn lpr	999/1001	8/9	1/9	2/1001	99.80%	11.11%
Ps	20/26	17/19	2/19	6/26	76.92%	10.52%
Inetd	19/21	3/3	0/3	2/21	90.47%	0%
Login	10/13	19/24	5/24	3/13	76.92%	20.83%
Named	3/5	3/5	2/5	2/5	60%	40%
Sendmail (CERT)	10/14	121/147	26/147	4/14	71.42%	17.68%

5. 결 론

시스템 콜을 기반으로한 침입 탐지 시스템이 많은 장점을 가지고 있음에도 불구하고 시스템 콜 데이터 집합으로 유한상태기계를 효율적으로 만들지 못하기 때문에 잘 사용하지 못했다. 본 논문에서는 이러한 문제의 해결책으로 유전자 알고리즘을 통해 자동적으로 오토마타를 생성하는 기법을 제안 하였다.

본 연구에서 사용한 데이터 집합의 특징에 따라 유전자 알고리즘의 개선이 필요한 부분도 있었다. 본 연구에서 제안한 유전자 알고리즘의 여러 연산(operation)과 인수(parameter)들을 변화시켜 오토마타의 성능을 향상 시키기 위한 연구를 진행 중이다.

참 고 문 헌

[1] 임영환,위규범, "침입탐지를 위한 유한상태기계의 생성 기법", 한국정보처리학회, 제10-C권, 제2호, pp.119-124, 2003.

[2] K. Wee and B. Moon, "Automatic Generation of Finite State Automata for Detecting Intrusion using System Call Sequences", Proceedings of MMM-ACNS, LNCS 2776, pp.206-216, St. Petersburg, Russia, Sept. 2003.

[3] R. Sekar and M. Bendre, "A Fast Automaton-Based Method for Detecting Anomalous Program Behaviors", Proceedings of the IEEE Symposium on Security and Privacy, pp.144-155, 2001.

[4] A. Belz and B. Eskikaya, "A Genetic Algorithm for Finite State Automata Induction with an Application to Phonotactics", Proceedings of the European Summer School in Logic, Language and Information Workshop, 1998.

[5] Melanie Mitchell, "An Introduction to Genetic Algorithms", The MIT Press, 1996.

[6] C. Warrender, S. Forrest and B. Pearlmutter, "Detecting Intrusions Using Systme Calls : Alternative Data Models", Proceedings of the IEEE Symposium on Security and Privacy, pp.133-145, 1999.

[7] http://www.cs.umm.edu/~immsec/systemcalls.htm