연관 마이닝 기법을 이용한 침입 시나리오 자동생성 알고리즘

정경훈 · 주정은 · 황현숙 · 김창수

부경대학교

Automated Generation Algorithm of the Penetration Scenarios using Association Mining Technique

Kyung-hoon Jung · Jung-eun Ju · Hyun-suk Hwang · Chang-soo Kim

Pukyong National University

요 약

본 논문에서는 연관 마이닝 기법을 이용한 침입 시나리오 자동생성 알고리즘을 제안한다. 현재 알려진 침입 탐지는 크게 비정상 탐지(Anomaly Detection)와 오용 탐지(Misuse Detection)로 분류되는데, 침입 판정을 위해 전자는 통계적 방법, 특징 추출, 신경망 기법 등을 사용하며, 후자는 조건부 확률, 전문가 시스템, 상태 전이 분석, 패턴 매칭 등을 사용한다. 기존에 제안된 침입 탐지 알고리즘들의 경우 알려지지 않은 침입은 보안 전문가에 의해 수동적으로 시나리오를 생성·갱신한다. 본 알고리즘은 기존의 데이터 내에 있는 알려지지 않은 유효하고 잠재적으로 유용한 정보를 발견하는데 사용되는 연관 마이닝 알고리즘을 상태전이 기법에 적용하여 침입 시나리오를 자동으로 생성한다.

본 논문에서 제안한 알고리즘은 보안 전문가에 의해 수동적으로 생성되던 침입 시나리오를 자동적으로 생성할 수 있으며, 기존 알고리즘에 비해서 새로운 침입에 대응하는 것이 용이하고 시스템 유지 보수비용이 적다는 이점이 있다.

ABSTRACT

In this paper we propose the automated generation algorithm of penetration scenario using association mining technique. Until now known intrusion detections are classified into anomaly detection and misuse detection. The former uses statistical method, features selection, neural network method in order to decide intrusion, the latter uses conditional probability, expert system, state transition analysis, pattern matching for deciding intrusion. In proposed many intrusion detection algorithms unknown penetrations are created and updated by security experts. Our algorithm automatically generates penetration scenarios applying association mining technique to state transition technique. Association mining technique discovers efficient and useful unknown information in existing data. In this paper the algorithm we propose can automatically generate penetration scenarios to have been produced by security experts and is easy to cope with intrusions when it is compared to existing intrusion algorithms. Also It has advantage that maintenance cost is not high.

Ⅰ. 서 론

전세계에 물리적으로 떨어진 사용자들 사이의 정보 공유를 목적으로 하는 인터넷은 개방성을 중요시한다. 그러나 이러한 개방성으로 인해 시스템에 불법 침입하여 기밀 자료를 유출하거나 시스템 자체를 파괴시켜 시스템 사용을 못하게 하는 해킹이 문제시되고 있다. 따라서 이러한 불법적인 침입을 탐지하고 방어하는 침입탐지 시스템의 필요성이 절실하다.

침입탐지는 크게 비정상 탐지와 오용 탐지로 분류될 수 있는데, 전자는 컴퓨터 자원의 비정상적인 행위나 사용에 근거한 침입을 탐지하는 방법으로써 통계적 방법, 특징 추출, 신경망 등으로

나눌 수 있다. 후자의 경우, 시스템이나 응용 소프트웨어의 약점을 이용한 침입을 탐지하는 방법으로써 전문가 시스템, 상태 전이 분석, 패턴 매칭 등으로 나눌 수 있다. 지금까지 개발된 많은 침입 탐지 시스템들은 위에서 기술된 침입탐지 기법들을 조합하거나 응용하여 구현되었으며, 새로운 공격방법들이 발견될때마다 지속적으로 시스템 갱신을 하고 있다. 일반적으로 침입탐지 시스템의 갱신은 보안 전문가에 의해 수행되는데, 이는 시간적으로 상당한 지연을 야기시킬 뿐만 아니라 실시간 시스템 업그레이드를 할 수 없다는 한계를 지니고 있다. 따라서 알려지지 않은 새로운 침입에 대한 즉각적인 반응을 하는 침입탐지 시스템의 개발이 요구되는데, 이를 위한 방법 중의 하나가 데이터 마이닝(Data Mining) 기법이다.

데이터 마이닝은 데이터 내에 있는 알려지지 않은 유효하고 잠재적으로 유용한 정보를 발견하기 위한 작업으로써, 크게 추출될 지식의 형태와 적용 기술의 종류에 따라 분류할 수 있다. 전자의 경우 특성화(characterization), 분류화(classification), 군집화(clustering), 연관(association), 순차(sequential)로 나눌 수 있으며, 후자의 경우 통계, 데이터베이스의 쿼리, 의사결정 트리, 신경망 등으로 나눌 수 있다.

본 논문에서는 데이터 마이닝 기법 중에서 연관기법을 침입 탐지 알고리즘에 적용하여 침입 시나리오를 자동적으로 생성하는 알고리즘을 제안하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 침입탐지 및 연관기법에 대해서 기술하고, 3장에서는 본 연구에서 제안한 침입 시나리오 자동생성 알고리즘에 대해 기술하며, 마지막 4장에서는 결론 및 향후 연구과제를 기술한다.

II. 관련연구

2.1 침입탐지 기법

2.1.1 통계학적 방법

침입탐지 시스템 개발에 있어서 가장 초기 방법 중의 하나인 통계학적 침입 탐지 방법[1]은 시스템 상에서 생성된 감사 자료(audit data)의 양과 형태의 변화를 측정하여 침입을 탐지한다. 통계학적 비정상 탐지는 임계값 탐지(threshold detec- tion)과 프로파일 기반 비정상 탐지(profile-based anomaly detection)으로 분류할 수 있다. 임계값 탐지의 경우 시스템의 정상 동작 동안 일어날 이벤트 예상치가 지정한 양을 능가할 경우 이를 검출하며 이러한 특정 이벤트 발생의 기록을 목적으로 한다. MIDAS(Multics Intrusion Detection and Alerting System)[2], NADIR (Network Anomaly Detection and Intrusion Reporter)[3] 등이 이 범주에 속한다. 프로파일 기반 비정상 탐지는 시스템 내에 있는 감사 로그들의 감시를 통해 침입을 탐지하는 방법으로써 이미 구축된 사용패턴의 변화량을 검사한다. 여기에는 SRI의 IDES(Intrusion Detection Expert System)[4], MIDAS[2], NADIR[3], Haystack[5] 등이 있다.

2.1.2 규칙기반(Rule-Based) 방법

감사 데이터 내의 사용 패턴을 식별하기 위해 통계학 공식을 사용하는 통계학적 방법과는 달리 규칙기반 비정상 탐지는 사용 패턴을 표현하고 저장하기 위해 규칙의 집합들을 사용한다. W&S(Wisdom and Sense)[6], TIM(Time-based Inductive Machine)[7] 등이 규칙기반 방법에 속한다.

2.1.3 상태전이(State Transition) 방법

상태 전이 분석 방법은 공격 패턴을 특정 시스템의 상태 전이 순서로 표현한다.

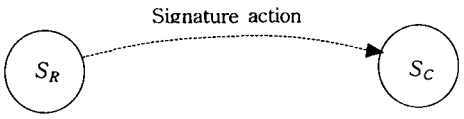


그림 1 STAT의 상태 전이도

(그림 1)과 같이 상태 S_R에서 상태 S_C로 이동하기 위해서는 이벤트들이 요구되며, 하나의 상태가 주어진 조건을 만족하면 다음의 상태는 어떤 것인지를 알 수 있게 된다. STAT(State Transition Analysis Tool)[8], NetStat[9] 방법 등이 있다.

2.2 연관 마이닝 기법

고객의 트랜잭션에 있는 항목간의 규칙을 표현하는 기법인 연관규칙은 어떤 사건이 발생될 때 다음 사건이 발생하는 관련성을 의미한다. 연관규칙에는 지지도와 신뢰도라는 두 가지 척도가 있는데, 지지도는 생성된 연관규칙에 있는 항목집합이 전체 항목에서 차지하는 비율을 의미하고, 신뢰도는 조건부를 만족하는 트랜잭션이 결과부까지 만족하는 비율을 의미한다. 이를 수식으로 나타내면 다음과 같이 나타낼 수 있다. 수식에서 R은 연과규칙을 나타내며, X, Y는 각각 트랜잭션을 나타낸다.

$$support(R) = support(X \cup Y)$$

$$confidence(R) = support(X \cup Y) \ / \ support(X)$$

연관 알고리즘은 크게 두 단계로 나타낼 수 있는데, 첫 단계는 빈발 항목집합을 발견하는 단계이고, 두 번째 단계는 설정한 빈발 항목집합에 대해 모든 부분집합을 생성하여 최소한의 신뢰도 이상인 규칙을 찾아낸다. 이러한 연관 알고리즘에는 AIS[10], Apriori[11] 등이 있다.

III. 침입 시나리오 자동생성 알고리즘

3.1 상태전이 분석 도구(STAT)

STAT(State Transition Analysis Tool)는 침입을 수집된 감사 레코드를 (그림 1)과 같은 상태 전이도로 표현하며, 이러한 상태 전이도는 State Description Table(SDT)과 rule chain으로 변환된다. SDT는 상태 전이도의 각 상태에 대한 기술이며, rule chain은 SDT 내에서의 상태 위치와 상태 변화을 야기시키는 Signature action, 그리고 상태와 상태 사이의 의존성 등을 나타낸다. (그림 2)는 감사레코드와 SDT, rule chain의 형식을 보여주고 있다.

Subject ID	Subject Permissions	Action	Object ID	Object Owner	Object Permissions

(a) 감사 레코드 형식

상태 / 침입	S_1	S_2	S_3	...	S_n
P_1	상태 기술	상태 기술	상태 기술	...	상태 기술

(b) State Description Table

State description	Signature action	Rule dependence

(c) Rule Base의 Rule chain

그림 2 STAT에서 사용되는 자료구조들

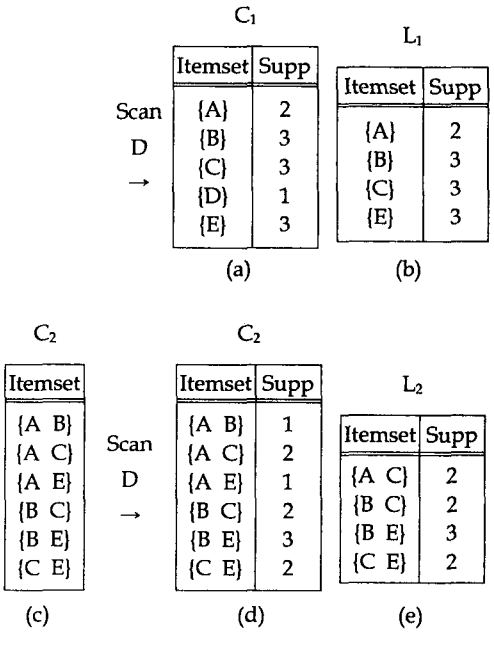
본 논문은 침입 시나리오의 자동 생성에 관한 연구이므로, STAT에서의 침입 판정에 관한 알고리즘의 자세한 설명은 [8]을 참조하기 바란다.

3.2 Apriori 연관 알고리즘

Apriori 연관 규칙 탐사는 두 단계로 구성된다. 첫 번째 단계에서는 미리 결정된 최소 지지도인 s_{min} 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들인 빈발 항목집합(large itemset)을 찾는다. 두 번째 단계에서는 빈발 항목집합을 사용하여 데이터베이스로부터 연관규칙을 생성한다. 이때 모든 빈발 항목집합 L에 대해 공집합을 제외한 L의 모든 부분집합을 찾는다. 그리고 Ldm 각 부분집합 A에 대해 만약 지지도 supp(A)에 대한 supp(L)의 비율이 적어도 최소 신뢰도 c_{min} 이상이면 (supp(L)/supp(A) $\geq c_{min}$), A→(L-A) 형태의 규칙을 출력한다. 잠재적인 빈발 항목집합들의 수는 모든 항목들의 멱집합(power set)의 크기와 같으며, 고려될 항목들의 크기에 대하여 기하급수적으로 증가한다. 따라서 모든 알고리즘들은 실제로 빈발한 항목들을 찾기 위해 후보(candidates)라 지칭하는 빈발 가능성이 있는 항목집합들을 생성한 후, 데이터베이스를 읽어나가면서 각 후보 항목집합들의 지지도를 계산한다. (그림 3)과 (그림 4)는 위에서 기술한 단계 1, 2의 수행과정을 나타내고 있으며, (그림 5)는 이러한 수행에 대한 Apriori 알고리즘을 보여주고 있다. (그림 3)은 예제 데이터베이스이고, (그림 4)는 Apriori 알고리즘에 의해 빈발 항목집합을 구하는 과정을 보여주고 있으며, C_k, L_k, D는 각각 후보 항목집합과 빈발 항목집합, 데이터베이스를 나타내며, 첨자 k는 항목집합내의 항목수를 나타낸다.

TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

그림 3 데이타베이스



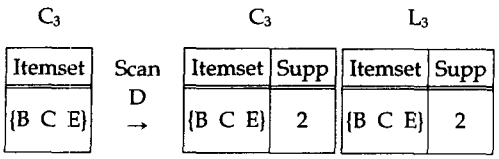


그림 4 후보 항목집합과 빈발 항목집합의 생성

1: L_1 = {large 1-itemsets}
2: **for** ($k=2; L_{k-1} \neq \emptyset; k++$) **do begin**
3: C_k=apriori-gen(L_{k-1}); //New candidates
4: **forall** transactions $t \in D$ **do begin**
5: C_t=subset(C_k, t);//Candidates contained in t
6: **forall** candidates $c \in C_t$ **do**
7: c.count++;
8: **end**
9: L_k = { $c \in C_k$ | c.count $\geq$ minsup}
10:**end**
11:Answer = $\bigcup_k L_k$;

그림 5 빈발 항목집합 생성 알고리즘

(그림 6)은 (그림 5)의 라인 3에 있는 apriori-gen에 대한 알고리즘으로써 후보 항목집합을 생성하며, join 단계와 prune 단계로 구성된다. (그림 6)의 라인 2~4는 join 단계이로써 k번째 후보 항목집합을 생성하며, 라인 6~9는 prune 단계로써 생성된 후보 k-항목집합에서 필요없는 후보 항목집합을 삭제한다.

1: **Algorithm** Apriori-gen
2: **insert into** C_k // Join step
3: **select** $a.item_1, \cdots, a.item_{k-1}, b.item_{k-1}$
4: **from** $L_{k-1}a, L_{k-1}b$
5: **where** $a.item_1 = b.item_1, \cdots, a.item_{k-2} = b.item_{k-2}, a.item_{k-1} < b.item_{k-1}$;
//Prune step: now prune rules with subsets missing in L_{k-1}
6: **forall** itemset $c \in C_k$ **do**
7: **forall** (k-1)-subsets s of c **do**
8: **if** ($s \notin L_{k-1}$) **then**
9: **delete** c from C_k ;

그림 6 후보 항목집합 생성 알고리즘

3.3 침입 시나리오 자동생성 알고리즘

본 논문에서 제안한 침입 시나리오 자동생성 알고리즘인 AGAPS(Automated Generation Algorithm of the Penetration Scenarios)는 State Description Table(SDT)을 (그림 8)의 수정된 Apriori 연관 알고리즘의 입력으로 사용하며, 그 출력인 후보 침입 시나리오는 후보 SDT인 CSDT (Candidate SDT)와 후보 rule chain에 저장된다. 후보 rule chain의 경우 STAT의 rule chain과 동일하며, CSDT의 형식은 (그림 2)와 같다. CSDT는 후보 침입 시나리오가 발생하면 값이 증가되는 Count 항목이 추가된 것을 제외하면 (그림 2(b))의 구조와 동일하다. Count 항목값 N이 설정된 임계값이상이면 후보 침입 시나리오인 CP_1은 SDT와 rule chain에 저장된다.

	Count	S_1	···	S_3
CP_1	N	상태 기술	···	상태 기술

그림 7 후보 SDT(CSDT)

1: L_1 = {SDT}
2: **for** ($k=2; L_{k-1} \neq \emptyset; k++$) **do begin**
3: $CSDT_k$ = apriori-gen(L_{k-1});
4: **forall** transactions $t \in SDT$ **do begin**
5: $CSDT_t$ = subset($CSDT_k$, t);
6: **forall** candidates $c \in C_t$ **do**
7: c.count++;
8: **end**
9: L_k={ $c \in C_k$ | c.count $\geq$ minsup}
10: **end**
11: Answer = $\bigcup_k L_k$;

그림 8 수정된 Apriori 알고리즘

1: **Algorithm** Modified Apriori-gen
2: **insert into** $CSDT_k$
3: **select** $a.item_1, \cdots, a.item_{k-1}, b.item_{k-1}$
4: **from** SDT
5: **where** $a.item_1 = b.item_1, \cdots, a.item_{k-2} = b.item_{k-2}, a.item_{k-1} \neq b.item_{k-1}$;
6: **forall** itemset $c \in CSDT_k$ **do**
7: **forall** (k-1)-subsets s of c **do**
8: **if** ($s \notin L_{k-1}$) **then**
9: **delete** c from $CSDT_k$

그림 9 수정된 Apriori-gen 알고리즘

IV. 결 론

본 논문에서는 저장된 감사 레코드로부터 알려지지 않은 침입 시나리오를 자동생성하는 알고리즘을 제안하였다. 침입 판정을 위해 STAT 상태전이 기법을 사용하였으며, STAT의 감사 레코드를 Apriori 연관 알고리즘의 입력으로 사용하여 알려지지 않은 침입 시나리오를 생성하였다.

그러나 연관기법의 경우 생성된 규칙의 적합여부는 임계값에 의해 결정되는데, 본 연구에서 제안한 연관 알고리즘의 경우 임의로 정해서 사용하였다. 따라서 연관 알고리즘에서 사용되는 임계값을 통계학적 모델에 근거하여 결정할 필요가 있다.

참고문헌

[1] 한국정보보호센터, 정보시스템 침해사고방지기술('97), pp.163-168, Jan. 1998

[2] Sebring, M. M., Shellhouse, E., Hanna, M.E. and Whitehurst, R.A., "Expert System in Intrusion Detection: A Case Study", Proceedings of the 11th National Computer Security Conference, Baltimore, MD, pp.74-81, Oct. 1988

[3] B. Hubbard, T. Haley, N. McAuliffe, L. Schaefer, N. Kelem, D. Wolcott, R. Feiertag and M. Schaefer, "Computer System Intrusion Detection", Trusted Information Systems, Inc., RADC-TR-90-413 Final Technical Report, Dec, 1990

[4] T.F. Lunt, "Real-Time Intrusion Detection", Proceedings COMPCON, San Francisco, CA, pp.348-353, Feb. 1989

[5] S.E. Smaha, "Haystack: An Intrusion Detection System", Proceedings of the IEEE Fourth Aerospace Computer Security Applications COnference, pp.37-44, Dec. 1988

[6] H.S. Vaccaro and G.E. Liepins, "Detection of Anomalous Computer Session Activity", Proceedings of the IEEE Symposium on Research in Security and Privacy, Oakland, CA, pp.280-289, May 1989

[7] K. Chen, S.C. Lu and H.S. Teng, "Adaptive Real-Time Anomaly Detection Using Inductively Generated Sequential patterns", presented at the Fifth Intrusion Detection Workshop, SRI Internation, Menlo Park, CA, May 1990

[8] P. Porras, "STAT - A State Transition Analysis Tool for Intrusion Detection", Master's thesis, Computer Science Defpartment, University of California, Santa Barbara, June 1992

[9] G. Vigna and R. Kemmerer, "NetSTAT: A network-based intrusion detection approach", Proceedings of the 14th Annual Computer Security Applications Conference, Scottsdale, Arizona, Dec. 1998