망침입검출에서 속성선택에 의한 성능개선

박성호, 황철진

망보안을 위한 침입검출기술에서 현재 중요한 방향은 미지의 공격을 검출하기 위한 이상검출에 대한 연구[1]이다.

이상검출의 성능을 높이기 위하여 인공신경망, 유전알고리듬, 면역원리, 자료발굴을 비롯한 지능적인 방법들이 많이 연구되고 적용되고있다. 특히 베이스분류에 기초한 이상검출기술[2]은 분류알고리듬이 간단하고 정상모형구축이 비교적 쉬우며 검출률이 높은 우점을 가지고있으나 오경보률이 높고 계산량이 많다는 약점을 가지고있다.

론문에서는 베이스분류에 기초한 침입검출에서 거친모임리론에 의한 속성축소방법을 리용하여 검출률과 성능을 높이는 방법을 제안하였다.

1. 베이스분류에 기초한 망침입검출

베이스분류방법은 확률통계학에 기초하고있으며 간단히 보면 확률의 크기를 리용하여 사건이 소속되는 클라스를 판단하는것이다.

분류과정은 다음과 같다.

$X = \{a_1, a_2, \cdots, a_m\}$이 분류하려는 대상이고 매 a_i는 X의 하나의 특징속성이라고 하고 클라스들의 모임 $C = \{y_1, y_2, \cdots, y_n\}$이 존재한다고 하자.

① $P(y_1|X),\ P(y_2|X),\ \cdots,\ P(y_n|X)$를 계산한다.

② 만일 $P(y_k|X) = \max\{P(y_1|X), P(y_2|X), \cdots, P(y_n|X)\}$라면 X를 y_k에 소속시킨다.

①의 조건부확률은 다음과 같이 계산한다.

ㄱ) 이미 분류된 훈련표본모임을 리용한다.

ㄴ) 각 클라스별로 각 특징속성의 조건부확률추정을 얻는다.

$$P(a_1|y_1),\ P(a_2|y_1),\ \cdots,\ P(a_m|y_1)$$
$$P(a_1|y_2),\ P(a_2|y_2),\ \cdots,\ P(a_m|y_2)$$
$$\vdots$$
$$P(a_1|y_n),\ P(a_2|y_n),\ \cdots,\ P(a_m|y_n)$$

ㄷ) 만일 각 특징속성이 조건부독립이라면 베이스정리에 근거하여 다음과 같이 계산할수 있다.

$$P(y_i \mid X) = \frac{P(X \mid y_i)P(y_i)}{P(X)}$$

분모가 모든 클라스에 대하여 상수이므로 분자를 최대화하면 된다.

또한 각 특징속성들이 조건부독립이므로

$$P(X \mid y_i)P(y_i) = P(a_1 \mid y_i)P(a_2 \mid y_i)\cdots P(a_m \mid y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j \mid y_i)$$

이다.

만일 클라스들이 모두 등확률이라고 가정하면 즉 $P(y_1) = P(y_2) = \cdots = P(y_n)$이면 우의 계산에서 $P(X|y_i)$만을 최대화하면 된다.

$$P(X \mid y_i) = \prod_{j=1}^{m} P(a_j \mid y_i)$$

베이스분류에 기초한 침입검출에서는 속성들사이에 반드시 서로 독립이여야 한다는 가정을 준수해야 하는데 이것은 실천에서 항상 만족되는것은 아니다. 또한 미지의 공격을 검출할수 있다 해도 비교적 많은 시간이 요구되며 이로 하여 실시간성요구를 만족시키기 어렵다.

론문에서는 이러한 약점을 극복하고 베이스침입검출의 성능을 높이기 위하여 속성선택, 속성축소방법을 적용하기로 한다.

2. 속성축소에 의한 침입검출의 성능개선

거친모임리론에 기초한 속성선택방법에서 기본방법은 식별행렬방법이다.

정보체계 $T = (U, A, C, D)$에서 하나의 $n \times n$ 행렬 $M(T) = \{m_{ij} : i, j = 1, 2, 3, \cdots, n\}$의 원소들이 $m_{ij} = \{(a \in C,\ a(x_i) \neq a(x_j)) \wedge (d \in D,\ d(x_i) \neq d(x_j)),\ i,\ j = 1, 2, \cdots, n\}$ 으로 결정된다면 행렬 $M(T)$를 정보체계 T의 식별행렬이라고 부른다. 여기서 U는 대상령역모임, A는 속성모임, C는 조건속성모임, D는 결정속성모임이며 $A = C \cup D$ 이다.

정보체계를 표현하는 결정표에 의하여 정보체계의 식별행렬을 구성하고 m_{ij}가 하나의 속성에 의하여 구성된다면 이 속성은 그 정보체계의 필요속성을 구성하며 모든 필요속성을 찾으면 곧 정보체계의 핵모임을 구성하는것으로 된다.

정보체계의 결정표에 의하여 식별행렬을 구성하고 그로부터 축소와 핵모임을 구하는 과정에는 다량의 중간자료가 산생되고 그것에 대한 기억요구가 제기되며 특히 속성모임의 차원수증가에 따라 계산량이 매우 커지게 된다. 이로부터 큰 차원수의 속성모임에 대하여 속성축소의 시공간적성능을 높이는것은 현실적으로 중요하게 제기된다.

론문에서는 거친모임리론의 한가지 중요개념인 속성의존도에 기초한 속성선택방법을 제안하였다.

다음의 정의에 의하여 속성의존도에 대한 개념을 확장하기로 한다.

정의 정보체계의 모든 속성모임 A에서 2개의 속성으로 구성된 모임을 2속성모임이라고 부르며 n개 속성으로 구성된 모임을 n속성모임이라고 부른다. 모든 조건속성으로 구성된 속성모임을 전체 조건속성모임(C)이라고 부르며 모든 결정속성으로 구성된 속성모임을 전체 결정속성모임(D)이라고 부른다. 이 속성들을 다속성모임이라고 하며 특히 속성모임이 하나의 속성만을 포함하면 단속성모임이라고 한다.

원래 속성의존도개념에서는 결정속성의 단일조건속성에 대한 의존정도만을 론의하는데 망침입자료에 대한 처리를 비롯한 실제응용에서 단일속성의존도는 의의가 없다.

그러므로 이러한 자료들을 전제로 하여 다속성모임에 대한 의존도를 고찰함으로써 결정표의 축소와 핵을 구하는 과정의 성능을 개선하였다.

다음의 결정표로 표현되는 정보체계를 고찰하자.(표 1)

표 1. 정보체계 T의 결정표

Ex	a	b	c	d	D	Ex	a	b	c	d	D
E1	2	0	1	0	0	E5	1	2	0	1	1
E2	2	1	1	0	0	E6	0	2	2	1	1
E3	0	1	1	0	0	E7	0	2	1	1	2
E4	1	2	2	0	1	E8	2	2	1	1	2

결정속성과 조건속성이 분류방법을 구성하는데서는 같으므로 여기서는 여러개의 조건속성으로 구성된 조건속성모임을 론의하며 결정속성은 단속성으로만 구성한다.

표 1에서 a, b, c, d는 조건속성이며 D는 결정속성이다.

표 1의 결정표로부터 거친모임리론의 의존도정의에 따라 결정속성 D의 단일조건속성 a, b, c, d에 대한 의존도를 각각 구할수 있다.

정의에서 다속성모임으로 확장한데 따라 우선 이 결정표의 모든 2속성모임의 의존도를 구한다. 2속성의존도를 구하는 방법과 단계는 결정속성의 단일속성의존도를 구하는 방법[3]과 류사하다.

먼저 모든 2속성모임을 찾는다. 그것들은 각각 $\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}$이며 대응하는 조건속성모임들의 등가클라스분할은 각각 다음과 같다.

$$U/\{a, b\} = \{\{E1\}, \{E2\}, \{E3\}, \{E4, E5\}, \{E6, E7\}, \{E8\}\}$$
$$U/\{a, c\} = \{\{E1, E2, E8\}, \{E3, E7\}, \{E4\}, \{E5\}, \{E6\}\}$$
$$U/\{a, d\} = \{\{E1, E2\}, \{E3\}, \{E4\}, \{E5\}, \{E6, E7\}, \{E8\}\}$$
$$U/\{b, c\} = \{\{E1\}, \{E2, E3\}, \{E4, E6\}, \{E5\}, \{E7, E8\}\}$$
$$U/\{b, d\} = \{\{E1\}, \{E2, E3\}, \{E4\}, \{E5, E6, E7, E8\}\}$$
$$U/\{c, d\} = \{\{E1, E2, E3\}, \{E4\}, \{E5\}, \{E6\}, \{E7, E8\}\}$$

계속하여 결정속성의 등가클라스분할을 구한다.

$$U/D = \{\{E1, E2, E3\}, \{E4, E5, E6\}, \{E7, E8\}\}$$

다음 조건속성등가클라스와 결정속성등가클라스의 사귐모임을 얻어야 하며 얻어진 사귐모임의 원소수를 구한다.

여기서는 결정속성 D의 2속성모임 $\{a, c\}$에 대한 의존도를 실례로 계산한다.

$$(U/\{a, c\}) \cap (U/D) = \{\{E4\}, \{E5\}, \{E6\}\}$$
$$\text{원소수} = 3$$

여기서 결정속성 D의 2속성모임 $\{a, c\}$에 대한 의존도는 $\gamma_c(D) = 3/8 = 0.375$이다.

같은 방법으로 모든 2속성모임들에 대하여 의존도를 구하면 표 2와 같다.

표 2. 2속성모임들의 의존도

속성모임	$\{a, b\}$	$\{a, c\}$	$\{a, d\}$	$\{b, c\}$	$\{b, d\}$	$\{c, d\}$
의존도	0.75	0.375	0.75	1	0.5	1

류사한 방법으로 계산한 결정속성의 3속성모임들에 대한 의존도는 표 3과 같다.

표 3. 3속성모임들의 의존도

속성모임	$\{a, b, c\}$	$\{a, b, d\}$	$\{a, c, d\}$	$\{b, c, d\}$
의존도	1	0.75	1	1

표 2와 3으로부터 2속성모임의존도에서 결정속성의 조건속성모임 $\{b, c\}$와 $\{c, d\}$에 대한 의존도는 1이며 3속성모임의존도에서 값이 1인것은 $\{a, b, c\}, \{a, c, d\}, \{b, c, d\}$이다.

이로부터 2속성모임중 의존도가 1인 모든 속성모임은 3속성모임에 포함되며 실례에서 의존도가 1인 2속성모임 $\{b, c\}$와 $\{c, d\}$는 이미 그 결정표의 2개의 축소이며 핵모임은 그것들의 사귐인 c임을 알수 있다.

식별행렬방법에 의하여 결정표의 축소와 핵을 구하는 방법과 비교하면 속성의존도에 기초한 방법은 명백히 시공간성능에서 우월하다는것을 알수 있다.

속성의존도에 의한 속성축소와 핵을 구하는 과정은 다음과 같다.

이 과정은 하나의 순환과정이며 순환회수의 상한값은 이 정보체계의 조건속성의 총개수로서 N으로 표시한다.

n이 1이면 모든 단속성의 의존도를 계산하고 n이 2이면 모든 2속성모임의 의존도를 계산한다.

이렇게 계속하여 n이 어떤 값일 때 속성의존도가 1인 하나 또는 여러개의 속성모임이 처음으로 나타나면 전체 계산과정을 끝낸다. 이때 얻어진 속성모임이 바로 전체 조건속성모임의 축소이며 축소들의 사귐에 의하여 그 정보체계의 핵을 얻을수 있다.

특수한 경우로서 n이 N과 같은데 여전히 속성의존도가 1인 속성모임이 없으면 이 결정표는 일치성결정표가 아니라고 판단한다.

3. 베이스기반침입검출체계에서의 실험과 분석

베이스분류기에 기초한 침입검출체계의 구성을 다음의 그림에 보여주었다.

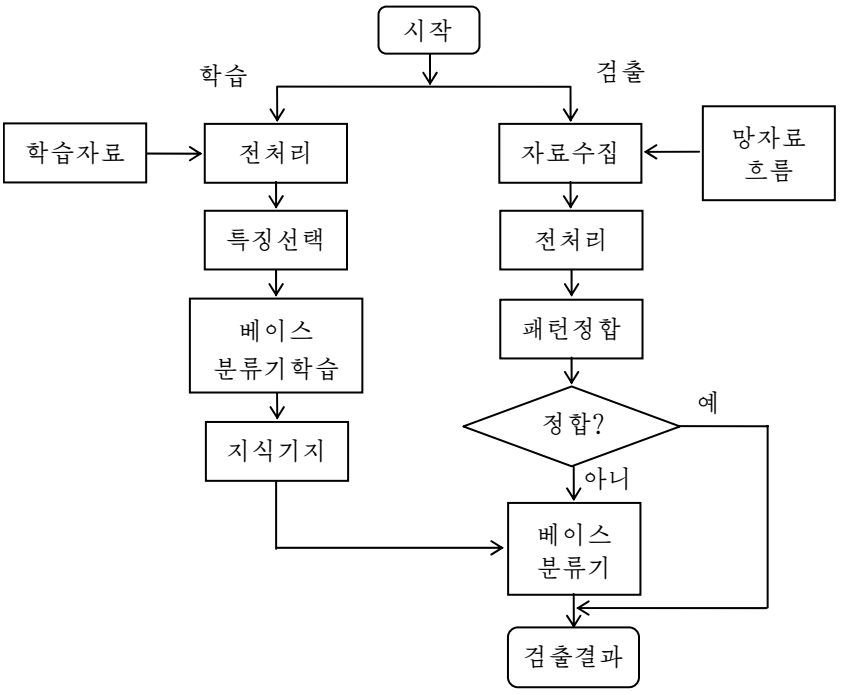

그림. 베이스분류기에 기초한 침입검출체계구성

학습단계에서는 주어진 학습자료에 기초하여 베이스분류기의 학습을 진행한다. 전처리된 학습자료에 대하여 앞에서 고찰한 속성축소방법을 리용하여 속성선택을 진행하며 얻어진 속성모임에 대하여 베이스분류기를 학습시킨다.

검출단계에서는 수집된 망자료흐름에 대하여 분석을 진행하고 같은 방법으로 속성선택을 진행한다. 오용검출방식의 패턴정합법을 적용하여 침입을 검출하며 정합이 성공되면 그것에 의하여 검출결과가 얻어진다. 정합이 성공하지 못하면 이상검출을 위한 베이스분류과정을 진행한다.

실험에서는 학습자료와 검사자료로서 망침입검출검사자료인 KDD－99자료모임을 사용하였다. KDD－99자료모임의 매 망자료기록은 41개의 속성들로 구성된다.

실험에서는 KDD－99로부터 2개의 자료모임(T1, T2)을 선택하는데 매 조에는 4가지 류형의 대표적인 자료류형에 대하여 각각 1 000개의 자료가 포함된다. T1자료모임은 학습자료모임으로 사용하고 T2는 검사자료모임으로 사용한다.

첫 실험에서는 원래의 베이스분류침입검출체계의 검출률과 의존도법에 의한 속성선택을 적용한 자료를 사용한 경우의 검출률 그리고 두 경우의 검출시간을 비교하였다.

표 4에서 속성선택을 사용한 경우의 검출률을 비교하였다.

표 4. 속성선택을 사용한 경우의 검출률비교

검출률/%	정상	1형	2형	3형	4형
선행방법	85.6	98.8	97.4	90.1	86.8
제안방법	85.3	98.8	97.3	90.1	86.6

실험결과로부터 속성선택을 한 후의 검출률은 약간 작아졌는데 이것은 실천에서 접수될수 있는것이라고 본다. 한편 속성선택을 한 후에 전체 체계의 검출시간은 원래시간의 0.36으로 감소되였는데 이것은 전체 체계의 성능이 크게 높아졌음을 보여준다.

다음의 실험에서는 거친모임리론의 대표적인 속성선택방법인 식별행렬에 의한 방법과 의존도에 기초한 방법의 검출률과 검출시간에 대한 비교를 진행하였다.(표 5) 사용한 자료는 우와 같다.

표 5. 속성선택방법들의 비교

방 법	식별행렬법	의존도법
검출률/%	94.8	94.8
검출시간	1	0.57

표 5로부터 두 방법의 검출률은 류사하며 속도에서는 의존도법에 의한 속성선택이 거의 2배 높다는것을 알수 있다. 두 방법에서 속성축소의 정지조건을 같이 주었으므로 검출률은 류사하나 제안된 방법의 성능개선이 뚜렷하다는것을 보여준다.

맺 는 말

베이스분류에 기초한 망침입검출체계에서 속성선택을 진행하여 성능을 높이였으며 거친모임리론의 의존도에 의한 속성선택방법을 제안하고 베이스망침입검출에 적용하여 검출률과 검출시간을 개선하였다.

참 고 문 헌

[1] S. Ricardo et al.; IEICE Transactions on Information and Systems, 3, 357, 2009.

[2] G. Gowrison et al.; Applied Soft Computing, 2, 921, 2013.

[3] James F. Peters et al.; Transactions on Rough Sets I, Springer-Verlag, 12～43, 2005.

주체107(2018)년 2월 5일 원고접수

Performance Improvement by Attribute Selection in the Network Intrusion Detection System

Pak Song Ho, Hwang Chol Jin

This paper describes a performance improvement by attribute reduction of rough set theory in the network intrusion system based on Bayesian classification.

Key words : intrusion detection system, rough set theory, attribute reduction