네트워크 침입 탐지를 위한 변형된 통계적 학습 모형

전 성 해†

요 약

최근 대부분의 정보 교류가 네트워크 환경 기반에서 이루어지고 있다. 때문에 외부의 침입으로부터 시스템을 보호해 주는 네트워크 침입 탐지 기술에 대한 연구가 매우 중요한 문제로 대두되고 있다. 하지만 시스템에 대한 침입 기술은 날로 새로워지고 더욱 정교화 되고 있어 이에 대한 대비가 절실한 실정이다. 현재 대부분의 침입 탐지 시스템은 이미 알려진 외부의 침입으로부터의 경험 데이터를 이용하여 침입 규칙을 모형화 하고, 이를 바탕으로 외부의 침입에 대비하는 전략을 취하고 있다. 이러한 방법으로는 새롭게 나타나는 침입 유형에 효과적으로 대처하지 못하게 된다. 따라서 본 논문에서는 통계적 학습 이론과 우도비검정 통계량을 이용하여 새로운 침입 유형까지 탐지해 낼 수 있는 변형된 통계적 학습 모형을 제안하였다. 즉, 기존의 정상적인 네트워크 사용에서 벗어나는 형태들에 대한 모형화를 통하여 시스템에 대한 침입 탐지를 수행하였다. KDD Cup-99 Task 데이터를 이용하여 정상적인 네트워크 사용을 벗어나는 새로운 침입을 제안 모형이 효과적으로 탐지함을 확인하였다.

Hybrid Statistical Learning Model for Intrusion Detection of Networks

Sung-Hae Jun†

ABSTRACT

Recently, most interchanges of information have been performed in the internet environments. So, the technique, which is used as intrusion detecting tool for system protecting against attack, is very important. But, the skills of intrusion detection are newer and more delicate, we need preparations for defending from these attacks. Currently, lots of intrusion detection systems make the model of intrusion detection rule using experienced data, based on this model they have the strategy of defence against attacks. This is not efficient for defense from new attack. In this paper, a new model of intrusion detection is proposed. This is hybrid statistical learning model using likelihood ratio test and statistical learning theory, then this model can detect a new attack as well as experienced attacks. This strategy performs intrusion detection according to make a model by finding abnormal attacks. Using KDD Cup-99 task data, we can know that the proposed model has a good result of intrusion detection.

키워드 : 침입 탐지(Intrusion Detection), 통계적 학습 모형(Statistical Learning Model), 우도비검정(Likelihood Ratio Testing)

1. 서 론

군사 목적 등 제한된 목적으로만 사용되었던 인터넷이 지금은 대부분의 기업들과 일반인들에게도 사용되어지고 있다. 이러한 환경 속에서 대부분의 정보 교류가 네트워크 기반에서 이루어지고 있기 때문에 외부의 침입으로부터 시스템을 보호하기 위한 보안의 중요성은 매우 중요한 문제로 부각되었다. 시스템에 대한 침입 기술도 날로 새로워지고 더욱 정교화되고 있다. 특히, 네트워크 시스템에 대한 공격용 프로그램이 무분별하게 인터넷에 유포되고, 이로 인하여 네트워크 지식이 부족한 일반인들도 시스템에 대한 침입이 가능하게 되었다. 최근 몇 년 동안 컴퓨터 보안 관련 범죄가 급증한 원인 중 하나도 이러한 문제 때문이다. 인터넷 사용이 급속히 증가하고 누구나 쉽게 네트워크 시스템에 연결이 가능하게 되었고, 인터넷에서 쉽게 구할 수 있는 침입 프로그램을 이용해 크래킹(cracking), 서비스 거부 공격(Denial of Service)등을 통해 시스템의 작동을 불가능하게 만들 수 있게 되었다. 네트워크 공격 유형에 대한 패러다임의 변화는 이미 시작되었다. 최근 야후(Yahoo), 아마존(Amazon) 등, 유명 인터넷 사이트에 대한 분산 서비스 거부공격(DDOS, Distributed Denial Of Service)에서 사용된 공격 도구에서 문제의 심각성이 잘 나타나 있다. 즉, 현재 대부분의 침입 탐지 시스템은 이미 알려진 외부의 침입으로부터의 경험 데이터를 이용하여 침입 규칙을 구축하여 외부의 침입에 대비하는 전략을 취하고 있다[5, 8, 9, 11, 14]. 하지만 이러한 방법으로는 새롭게 나타나는 침입 유형에는 대비하지 못하는 문제가 발생한다. 이러한 문제점을 해결하기 위하여 연구되고 있는 분야중 하나가 침입 탐지 시스템(IDS : Intrusion Detection System)이다[16].

본 논문에서는 이러한 문제점을 해결하기 위하여 우도비검정(likelihood ratio test : LRT) 통계량(power)[7]을 침입 요소를 나타내는 입력변수에 대한 가중치로 사용한 변형된 통계적 학습 이론(Hybrid Statistical Learning Theory : HSLT)을 제안하였다. 즉 기존의 정상적인 네트워크 사용에서 벗어나는 것들에 대한 모형화를 통하여 시스템의 침입을 찾아내었다. 이러한 침입 탐지 구축 전략은 새롭게 출현되는

† 정 회 원 : 청주대학교 통계학과 교수
논문접수 : 2003년 7월 22일, 심사완료 : 2003년 8월 19일

침입 유형에 대한 탐지도 가능하게 하였다. 제안 모형에서는 침입에 대한 행위를 수량화하고 LRT를 통한 검정통계량을 계산하여, 이를 입력변수들에 대한 가중치에 적용하여 이론적 기반과 분류의 성능이 우수한 통계적 학습 모형(Statistical Learning Theory : SLT)[6]인 SVM(Support Vector Machine)에 적용하였다. 제안한 기법을 이용하여 기존의 방법으로는 탐지가 불가능했던 새로운 유형에 대한 비교적 정확한 침입 탐지가 가능하게 됨을 KDD Cup-99 task 데이터를 이용한 실험을 통하여 확인되었다.

2. 통계적 학습 이론과 우도비 검정

기존의 네트워크 공격방법은 이미 알려져 있기 때문에 현재 구축되어 있는 침입 탐지 시스템의 이용을 통하여 침입에 대한 탐지가 가능하였다. 하지만 이러한 보안기술의 발전과 더불어 이를 침입하려는 새로운 공격 기법이 속속 등장하였다. 이러한 공격은 보다 정교한 침입 탐지 기법을 필요로 하며 대응방법에 대한 적응적인 변화를 요구한다. 앞으로 다가올 공격들은 감당하기 어려운 것들이 많아질 것이기 때문에 이에 대한 신중한 대응이 필요하다. 실질적이고 빠른 해결책이 필요하며, 이러한 전략은 현존하는 공격 문제에 대한 대응을 포함하여 앞으로 발생할 잠재적인 위협을 막을 수 있어야만 한다. 이러한 대응 전략을 위해서는 기존의 공격 패턴에 의한 침입 탐지가 아니라 정상적인 행동에서 벗어나는 모든 행동들에 대해서 탐지할 수 있는 새로운 기술이 요구된다. 기존의 침입 탐지 시스템은 알려진 공격으로부터 침입 탐지 규칙을 추출하기 때문에 새로운 침입에 대한 대응에는 한계가 있지만 LRT 검정통계량 가중치 기반의 SVM 기법을 적용한 HSLT 침입 탐지 모형은 기존과 같은 유형의 침입은 물론이고 새로운 침입에 대한 탐지도 가능하였다. 이 장에서는 제안 모형인 HSLT의 두가지 이론적 배경인 LRT와 SLT에 대하여 설명한다.

2.1 우도비 검정 통계량

주어진 학습 데이터의 각 변수는 하나의 확률 변수(random variable)가 된다. 2.2절의 학습 데이터 개체 $z=(x, y)$에서 x와 y도 각각 확률 변수이다. 확률 변수들 간의 연관성 유무, 또는 연관성의 정도를 확인하는 통계적 기법이 우도비 검정(Likelihood Ratio Testing : LRT)이다[2]. 변수 x와 y에 대한 LRT를 위한 가설(hypothesis)은 다음 식과 같다.

$$H_o : \mu_x - \mu_y = 0 \quad vs \quad H_1 : \mu_x - \mu_y \neq 0 \tag{1}$$

식 (1)의 가설에서 귀무가설(null hypothesis) H_0는 두 변수 x와 y 사이에 관련성이 없다는 것이고 대립가설(alternative hypothesis) H_1는 두 변수 간에 연간성이 존재한다는 것을 나타낸다. LRT는 H_0와 H_1중의 하나를 결정하는 통계적 가설 검정 기법 중의 하나이다. 본 논문의 침입 탐지 모형의 각 입력변수에 대한 가중치 정보로 사용할 LRT 검정 통계량 T는 다음식과 같이 표현된다.

$$T(x_1, x_2, \cdots, x_l) = \frac{\sqrt{l} \cdot \overline{x}}{\sqrt{\sum_{i=1}^{l} (x_i - \overline{x})^2/(l-1)}} \tag{2}$$
$$where, \quad \overline{x} = \frac{1}{l} \sum_{i=1}^{l} x_i$$

식 (2)에서 T-값이 클수록 두 확률 변수 사이의 연관성은 크게 된다. T-값은 t-확률분포(probability distribution)상의 기각역 정보를 제공하는 값이며 입력변수의 척도(scale)에 따라 값의 범위가 달라진다. 때문에 서로 다른 척도를 가진 입력변수(input variable)들에 대한 목표변수(target variable)에 대한 유의성을 비교하기 위해서는 이 값을 유의확률(p-value)로 바꾸어 주어야 한다. 이 값은 $-\infty$에서 $+\infty$까지의 범위를 같게 되는 T-값을 0에서 1사의 확률 값으로 척도에 관련 없이 표준화시켰기 때문에 서로 다른 척도를 가진 입력변수들에 대해서도 동등한 비교가 가능하게 된다. T 값과 유의확률은 반비례한다. 즉 유의 확률 값이 작을수록 두 변수간에 서로 유의한 차이가 없다는 귀무가설을 기각하게 되어 해당 입력변수가 공격과 정상사용의 레이블을 가진 목표 변수에 상관성이 높다고 할 수 있다. 따라서 본 논문에서는 유의 확률 값이 작은 입력 변수에 대하여 큰 가중치를 부여하는 전략을 취하였다. 즉, 본 논문의 실험에서는 각 입력 변수들에 대하여 공격과 정상사용의 두 레이블을 갖는 목표 변수와의 T 통계량 값을 구하여 해당 입력변수에 곱함으로써 가중치를 결정하였다.

2.2 통계적 학습 모형

본 논문의 침입 탐지 모형 구축을 위하여 사용하는 변형 이전의 통계적 학습 모형으로 SVM을 이용하였다[1, 13]. 클래스 레이블들을 가진 목표변수 y와 입력벡터(input vector) x로 구성된 데이터 집합 S는 다음 식과 같은 데이터 구조로 표현된다[3, 4].

$$(y_1, x_1), (y_2, x_2), \cdots, (y_l, x_l), \quad x_i \in R^N, \; y_i \in \{-1, 1\} \tag{3}$$

대부분의 분류모형 구축의 경우에 입력공간(input space)에서 서로 다른 클래스 레이블을 분류하는 정확한 초평면(hyperplane)을 찾는 것은 매우 제한적이기 때문에 바로 분류 모형을 사용하기가 어렵다[12]. 이러한 상황에서 해결 방안은 입력공간을 더 높은 차원의 특징 공간(feature space)으로 사상(mapping)시키고, 이 특징 공간에서 최적의 초평면을 찾는 것이다. $z = \phi(x)$를 입력공간 R^N에서 특징 공간 Z로의 사상 ϕ를 갖는 특징 공간 벡터로 표현하면, (w, b)의 쌍으로 이루어진 다음의 초평면을 구해야 한다.

$$w \cdot z + b = 0 \tag{4}$$

식 (4)의 초평면 식을 구하게 되면 다음의 식 (5)의 함수에 의해 개개의 x_i들을 분류할 수 있게 된다.

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases} \tag{5}$$

여기서 $w \in Z$이고 $b \in R$이다. 특히, 집합 S는 (w, b)의 쌍이 존재하면 선형분류 가능(linearly separable)이라고 하고 다음의 부등식이 S의 모든 원소들에 대해 성립한다.

$$\begin{cases} (w \cdot z_i + b) \geq 1, & \text{if } y_i = 1 \\ (w \cdot z_i + b) \leq -1, & \text{if } y_i = -1 \end{cases} \quad i = 1, 2, \cdots, l \qquad (6)$$

선형분류 가능한 집합 S는 두 개의 서로 다른 클래스 레이블들의 학습 데이터의 사영(projection)들 사이의 마진(margin)을 최대화 하는 유일한 최적 초평면을 구할 수 있다. 만약 집합 S가 선형 분류 가능이 아니면 분류규칙 위반(classification violations)이 SVM 형식에서 허용되어져야 한다[10]. 선형분류 가능이 아닌 데이터를 다루기 위하여 음이 아닌 변수 ξ_i를 도입하여 아래 식과 같이 식 (6)을 일반화한다.

$$y_i(w \cdot Z_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, l \qquad (7)$$

식 (7)에서 ξ_i는 식 (6)을 만족하지 않는 x_i들이다. 그러므로 $\sum_{i=1}^{l} \xi_i$는 오분류(misclassification)의 양을 나타내는 측도로서 고려된다. 따라서 최적 초평면을 구하는 문제는 아래의 문제에 대한 해(solution)가 된다.

$$\begin{aligned} &minimize \ \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i \\ &subject \ to \ y_i(w \cdot z_i + b) \geq 1 - \xi_i \end{aligned} \qquad (8)$$

여기서 $\xi_i \geq 0$이고 $i = 1, 2, \cdots, l$이다. C는 상수(constant)이며 조정 모수(regularization parameter)이다. 이 모수의 조정으로 마진 최대화와 분류 규칙 위반 사이의 균형을 맞출 수 있게 된다[10, 13]. 식 (8)에서 최적 초평면을 찾는 것은 다음의 라그랑지 변환(Lagrangian transformation)을 통하여 풀 수 있는 문제가 된다.

$$\begin{aligned} &maximize \ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ &subject \ to \ \sum_{i=1}^{l} y_i \alpha_i = 0 \qquad 0 \leq \alpha_i \leq C, \ i = 1, 2, \cdots, l \end{aligned} \qquad (9)$$

여기서 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)$는 식 (7)의 제한 조건과 관련된 음이 아닌 라그랑지 승수(multiplier)들의 벡터이다. Kuhn-Tucker 정리는 SVM 이론에서 중요한 역할을 한다. 이 정리에 의하여 식 (9)의 해 $\bar{\alpha}_i$는 다음을 만족한다.

$$\begin{aligned} &\bar{\alpha}_i(y_i(\bar{w} \cdot z_i + \bar{b}) - 1 + \bar{\xi}_i) = 0 \\ &(C - \bar{\alpha}_i)\bar{\xi}_i = 0, \ i = 1, 2, \cdots, l \end{aligned} \qquad (10)$$

식 (10)의 첫 번째 식으로부터 구한 해 $\bar{\alpha}_i$는 식 (7)의 등호를 만족시킨다. $\bar{\alpha}_i > 0$인 x_i를 support vector라고 부른다. 분류가 가능하지 않은(nonseparable) 경우에는 support vector는 두 가지의 형태로 존재한다. $0 < \bar{\alpha}_i < C$인 경우의 support vector x_i는 $y_i(\bar{w} \cdot z_i + \bar{b}) = 1$과 $\xi_i = 0$을 만족하고, $\bar{\alpha}_i = C$인 경우의 ξ_i는 널(null)이 아니고 대응되는 support vector x_i는 식 (6)을 만족하지 않는다. 이 support vector들은 오차(error)로서 간주된다. $\bar{\alpha}_i = 0$에 대응되는 x_i는 결정 마진(decision margin)과 떨어져서 정확하게 분류된다. 최적 초평면 $\bar{w} \cdot z + \bar{b}$를 구축하기 위하여 다음의 식과 스칼라 $\bar{b}$가 필요하다.

$$\bar{w} = \sum_{i=1}^{l} \bar{\alpha}_i y_i z_i \qquad (11)$$

이것은 식 (10)의 첫 번째 식의 Kuhn-Tucker 조건에 의해 결정된다. 결정 함수(decision function)는 식 (5)와 식 (11)에 의해 다음식과 같이 일반화된다.

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i z_i \cdot z + b \right) \qquad (12)$$

ϕ에 대한 어떠한 지식(knowledge)도 없기 때문에 식 (9)와 식 (12)의 계산은 불가능하다. 하지만 SVM은 ϕ에 대해서 알 필요가 없다. 단지 커널(kernel)이라 불리우는 $K(\cdot, \cdot)$가 다음과 같은 식에 의해 특징 공간 Z에 데이터의 내적(dot product)을 계산한다.

$$z_i \cdot z_j = \phi(x_i) \cdot \phi(x_j) = K(x_i, x_j) \qquad (13)$$

Mercer의 정리를 만족하는 함수들은 내적 계산이 가능하고 따라서 커널로써 사용이 가능하다[13]. SVM 분류기(classifier)를 구축하기 위하여 아래와 같은 차수(degree) d의 다항(polynimial) 커널을 사용한다.

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^d \qquad (14)$$

따라서 비선형 분류 가능 초평면은 다음 식의 해로서 구해진다.

$$\begin{aligned} &maximize \ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ &subject \ to \ \sum_{i=1}^{l} y_i \alpha_i = 0 \quad 0 \leq \alpha_i \leq C, \ i = 1, 2, \cdots, l \end{aligned} \qquad (15)$$

그리고 최종적인 결정 함수는 다음과 같다.

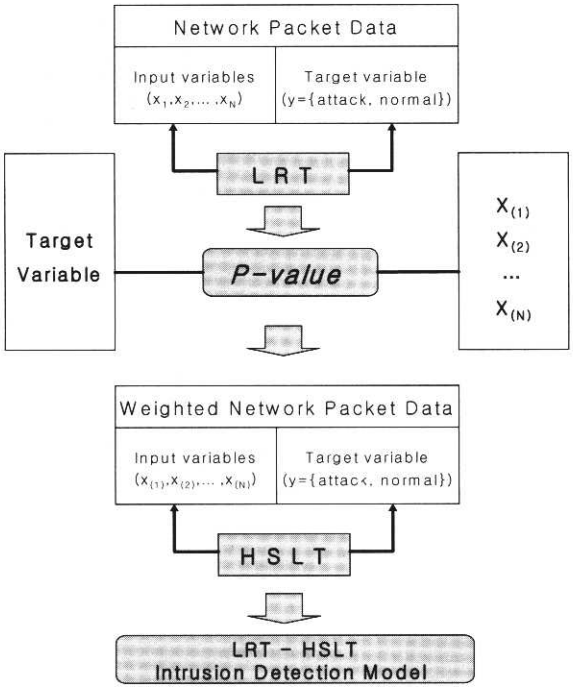
$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \qquad (16)$$

제안 모형은 LRT에 의해 가중치가 적용된 입력 변수들을 이용하여 공격과 정상 사용의 두 개의 레이블을 갖은 목표 변수의 클래스를 분류하게 된다.

3. LRT-HSLT 침입 탐지 모형

3.1 모형의 구조

본 논문에서 제안하는 침입 탐지 모형은 LRT 과정과 HSLT 과정의 2단계 프로세스로 구성된다. 우선 원래의 네트워크 패킷 데이터에 대하여 목표 변수와 입력 변수들 사이의 유의한 연관성의 정도를 LRT를 통한 유의확률 값에 의해 연관성의 순위를 결정하고 이를 통하여 각 입력 변수에 대한 목표 변수의 가중치를 결정하였다.



(그림 1) LRT-HSLT 침입 탐지 모형

다음으로 HSLT 단계에서는 기존의 SLT에서는 모든 입력 변수들이 목표 변수에 대하여 동일한 중요도를 가지지만 본 논문에서 사용한 LRT 과정에 의해 각 입력 변수의 가중치가 새롭게 부여되어 침입 탐지 모형이 구축된다. (그림 1)에서 $x_{(i)}$는 유의 확률에 의하여 목표 변수에 i번째로 유의한 입력 변수를 의미한다. 즉, 전체 입력 변수들 중에서 i번째로 큰 가중치 값을 갖게 되는 변수이다.

3.2 제안 침입 탐지 모형의 알고리즘

제안 모형의 알고리즘은 다음과 같이 3단계로 이루어진다.

단계 1 : (LRT step)

① 목표변수(y)와 입력변수들 $(x_1, x_2, \cdots, x_N)$간의 LRT 시행 → T-검정통계량 계산

$$T(x_1, x_2, \cdots, x_l) = \frac{\sqrt{l} \cdot \overline{x}}{\sqrt{\sum_{i=1}^{l} (x_i - \overline{x})^2/(l-1)}}$$

$$where, \quad \overline{x} = \frac{1}{l} \sum_{i=1}^{l} x_i$$

② 입력변수들간의 척도 영향을 없애기 위하여 T-값을 유의 확률(p-value)로 변환

$$p-value = P(t > T_i)$$

→ p-value의 크기 비교를 통한 입력 변수들의 서열화

$$(\{x_1, x_2, \cdots, x_N\} \rightarrow \{x_{(1)}, x_{(2)}, \cdot, x_{(N)}\})$$

$x_{(i)}$: i번째 가중치 크기를 갖는 입력 변수

③ p-value에 반비례하여 가중치 부여

단계 2 : (HSLT step)

침입 탐지 모형의 구축

$$f(x) = sign(w \cdot x + b), \quad z = \phi(x)$$

$x = (x_{(1)}, x_{(2)}, \cdots, x_{(N)})$

(w, b) : parameters

$\phi(\cdot)$: 입력 공간에서 특징 공간으로의 함수(mapping)

단계 3 : (Application step)

새로운 공격 데이터의 $f(x)$ 계산

→ $f(x)$의 값에 의해 공격과 정상을 분류

공격 : $f(x) = 1$ if $w \cdot z + b > 0$

정상 : $f(x) = -1$ if $w \cdot z + b \leq 0$

4. 실험 및 결과

4.1 실험 데이터

본 논문에서 제안하는 LRT-HSLT 침입 탐지 모형의 성능 평가를 위한 실험은 침입 탐지 분야에서 널리 사용되고 있는 KDD Cup(Knowledge Discovery & Data mining Cup) 99의 실험 데이터인 Network Packet Data를 이용하였다. 이 데이터는 1998년 미국 국방과학연구소에서 DARPA 침입 탐지 평가(Intrusion Detection Evaluation) 프로그램에 의해 추출되어졌다. 그런데 이 데이터는 공격의 여러 유형 중에서도 서비스 공격 거부(DOS)와 같은 특정 유형에 극단적으로 편중되어 있다. 이처럼 원 데이터가 DOS 공격에 편중된 구조를 띠고 있는 이유는 DOS 공격의 경우, 공격의 특성상 많은 패킷의 전송을 통하여 시스템의 처리 능력 이상의 요구를 보냄으로써 정상적인 서비스를 불가능하게 하기 때문이다. 즉, 네트워크 패킷 수가 과도하게 많게 된 것이다. 제안된 침입 탐지 학습 모형은 이처럼 한쪽으로 편중된 데이터로부터의 학습은 효율적이지 못하고, 또한 대부분의 데이터를 DOS로 판단하게 되는 경우가 발생할 수 있기 때문에 원 데이터 전체를 모형 구축에 사용하지 않고, 전처리 과정을 통하여 정확한 침입 탐지 모형이 구축될 수 있도록 하였다.

4.2 데이터 전처리 과정

<표 1>는 전처리 과정을 거치지 않은 편이된(biased) 원 데이터의 분포이다.

〈표 1〉 원 데이터의 분포

클래스의 레이블	데이터 수
DOS (attack)	3,883,370
R2L (attack)	1,126
U2R (attack)	42
Probing (attack)	41,102
Normal	972,780

<표 1>에서처럼 원 데이터의 분포는 DOS 공격에 지나치게 편중(bias) 되어진 특성을 보이고 있다. 공격(attack)의 경우 DOS는 3,883,370개 인데 비해 U2R은 불과 42개에 지

나지 않았다. 이처럼 학습 모형의 구축에 있어서 특정 패턴을 지나치게 많이 학습하게 되면 구축된 모형의 일반성이 보장될 수 없기 때문에 학습에 참여하는 데이터의 크기를 조절할 필요성이 있게 된다. 본 논문의 실험에서는 학습 데이터의 패턴 비율을 조정하기 위하여 DOS 데이터와 Normal 데이터에서는 표본추출을 하였다. 그러나 R2L, U2R, Probing은 원 데이터의 패턴 수가 상대적으로 매우 작기 때문에 표본 추출을 하지 않고 전체 데이터를 모형 학습에 사용하였다. 전처리 과정에서 중복되는 동일 데이터는 모두 개개의 결과로 고려함으로써 중복 데이터에 대한 가중처리를 하였다. 즉, 중복 데이터의 경우에 개체 수도 중요한 결정의 요소가 될 수 있다고 생각되었기 때문이다. 본 실험에서는 원 데이터의 비용 행렬(cost matrix)을 그대로 사용하였다. 오분류(misclassification)된 패킷들은 오분류가 어떻게 되었는지에 따라서 그 비용이 달라지기 때문이다. KDD 99cup의 데이터에 기반한 오분류 비용 행렬은 다음과 같다.

$$\begin{pmatrix} 0 & 2.3 \\ 1.9 & 0 \end{pmatrix} \tag{17}$$

위의 행렬에서 1행은 실제 공격(attack)이고 2행은 실제 정상 사용(normal)이다. 그리고 1열은 공격 예측이고 2열은 정상 사용 예측이다. 예를 들어 실제 공격인 경우를 정상 사용으로 잘못 예측했을 때의 비용은 2.3이 되고 공격으로 정확하게 예측하게 되면 비용은 0이 된다.

4.3 각 입력변수들에 대한 가중치 결정

본 실험에서 가중치는 공격과 정상의 두 클래스간의 차이에 대한 유의성을 많이 반영할 수 있는 입력 변수들의 서열화를 위하여 LRT의 T-검정통계량을 사용하였다. T-값은 두 클래스간의 정확한 분류에 많이 반영되는 입력 변수인 경우에는 증가하고, 그렇지 않은 경우에는 감소하게 된다. 따라서 이 값을 학습에 참여하는 전체 입력 변수들에 반영함으로써 학습 데이터에 대한 가중치 적용 효과를 얻었다. 총 34개의 입력 변수들[16]에 대하여 목표 변수에 대한 LRT를 통한 T-값을 계산하였다. 그런데 T-값은 입력 변수의 척도가 커짐에 따라 연관성과는 상관없이 증가하기 때문에 척도가 작은 입력변수는 목표 변수와의 연관성이 크다 할지라도 이 값이 상대적으로 작을 수 있다. 이러한 변수들 간의 척도에 대한 문제점을 해결하기 위하여 변수의 척도와 상관없이 입력변수들의 연관성을 비교할 수 있는 유의 확률 값(p-value)을 사용하였다. 이 값은 입력 변수와 목표 변수의 연관성이 없다는 귀무가설에서 검정통계량(T-값)의 계산 결과가 귀무가설을 기각하는 정도를 0에서 1사이의 값으로 나타내는 확률을 의미한다. <표 2>는 본 실험에 사용된 입력 변수들과 목표 변수간의 유의 확률의 일부를 나타내고 있다.

<표 2>에서 첫 번째인 duration은 연결 시간(number of seconds of the connection)을 나타내는 입력변수이며 0.0143의 유의 확률 값을 갖기 때문에 목표변수와의 연관성은 높은 것으로 볼 수 있다. 즉, 유의 확률값이 작을수록 공격과 정상 사용의 분류에 더 많은 영향을 줄 수 있다고 통계적으로 판단되어 본 논문에서는 해당 입력 변수의 유의 확률 값에 반비례하여 더 많은 가중치를 부여하였다. 만약 입력 변수들의 척도가 모두 동일하면 유의 확률을 사용하지 않고 T-값을 사용해도 무방하다.

〈표 2〉 각 입력변수의 유의 확률

입력 변수	p-value
duration	0.0143
src_bytes	0.1543
dst_bytes	0.0961
...	...
dst_host_srv_serror_rate	0.0165
dst_host_rerror_rate	0.0360
dst_host_srv_rerror_rate	0.0547

4.4 제안 모형의 성능 평가

제안된 LRT-HSLT 모형과의 성능 평가를 위하여 데이터 마이닝 분류 기법들 중에서 모형의 성능이 입증되고 있는 SVM(support vector machine), 분류나무 모형(classification tree), 로지스틱 회귀모형(logistic regression), 그리고 다층 신경망(multi-laver perceptron : MLP)과 비교하였다. 분류 성능의 비교는 기계학습 분류 모형에서 주로 사용되고 있는 평가 측도인 오분류율과 데이터 마이닝 분류모형에서 주로 사용하고 있는 지지도(lift value)를 사용하였다. <표 3>은 제안 모형과 비교 모형들에 대한 오분류율 결과이다.

〈표 3〉 비교 모형들의 오분류율

모 형	오분류율
LRT-SLT	0.0514
SVM	0.0619
Classification tree	0.2143
Logistic regression	0.1017
MLP	0.0931

위의 오분류율에 의한 비교 결과에서는 제안 모형이 로지스틱 회귀모형에 비해서는 2배이상 정확히 분류하고 있고, 현재 이진 분류기로서 좋은 성능을 보이고 있는 SVM에 비해서도 오분류율이 작게 나타나고 있음을 알 수 있다. 그런데 위의 결과를 보면 분류나무 모형의 오분류율이 다른 모형들에 비해 특히 좋지 않게 나왔다. 이는 연속형 입력 변수를 분류 나무 모형에 적용하기 위하여 범주화 하는 과정에서 정보의 손실이 발생하였기 때문으로 볼 수 있다. 따라서 분류나무 모형의 사용은 침입 탐지 데이터에는 적절치 않다고 판단되어 리프트 값(lift value)에 대한 모형 비교에서는 이 모형은 배제하였다. 분류 모형에서는 리프트 값보다도 오분류율이 모형 성능 평가에서 우선시되기 때문이기도 하다. 일반적으로 분류에 대한 리프트 값은 다음과 같이 구할 수 있다.

$$LV = \frac{\%Resp}{BL - LV} \tag{18}$$

위 식에서 %Resp는 해당 클래스의 전체 수에 대한 해당 클

래스에서 목표 변수의 특정 레이블의 빈도에 대한 백분율을 나타낸다. 또한 BL-LV(base line lift value)은 분류 모형이 적용되지 않은 원래의 학습 데이터에 대한 리프트 값이다. <표 4>는 이들 비교 모형들의 리프트 값의 결과를 나타내고 있다.

〈표 4〉 비교 모형들의 Lift Value

모 형	lift value
LRT-SLT	2.98
SVM	2.41
Logistic regression	2.09
MLP	1.98

위의 결과를 보면 LRT-HSLT 모형의 리프트 값이 2.98로 나왔다. 이는 이 모형에 의해 공격의 가능성에 대한 분류 결과 상위 10%의 예측이 모형을 구축하기 전에 비해 2.98배 성능 향상을 보이고 있는 것이다. 이에 비해 다른 모형들의 리프트 값은 제안 모형에 비해 작음을 알 수 있다.

5. 결론 및 향후 과제

본 논문에서는 시스템의 보안을 위해 사용되고 있는 침입 탐지 시스템에 대한 전략으로서 LRT-HSLT 침입 탐지 모형을 제안하였다. 기존의 대부분의 침입 탐지 시스템의 모형과 달리 제안 모형은 새로운 침입에 대한 적응성을 높이기 위하여 공격과 정상의 두 클래스만을 분류하는 침입 탐지 전략에 대하여 연구하였다. 즉, 기존의 연구들이 알려진 침입에 대한 패턴을 추출하고 모형화 하여, 또다시 같은 침입이 발생하는 경우에 이를 탐지하는 것에 중점을 둔 것에 비해, 본 연구는 알려지지 않은 새로운 침입에 대해서도 적응성을 가지고 탐지해 낼 수 있는 모형을 제안하였다.

공격의 기법이 점점 지능화됨에 따라 기존의 방법으로는 침입에 대한 탐지에 한계를 보이고 있고 자동화, 분산화 공격에 의해 하나의 공격 목표에 대해서 한 곳에서 공격이 일어나지 않고 동시에 여러 곳에서 이루어지는 특징을 보이고 있는 야후, 알타비스타 등의 분산서비스 거부와 같은 공격 사례에 대응하기 위하여 제안 모형은 일반적인 감시데이터를 사용하지 않고 네트워크 패킷 데이터를 사용함으로서 외부에서 공격하는 침입을 좀더 효과적으로 탐지할 수 있도록 하였다. 네트워크 패킷 데이터를 분석하여 모형화 하기 위하여 통계적 가설검정 기법인 LRT에 의한 T-값, 혹은 유의 확률 값을 이용하여 각 입력변수의 가중치를 구하고 이를 SVM의 입력변수로 적용한 변형된 통계 학습 모형을 개발하였다. 실험을 통하여 제안된 모형이 기존의 분류 모형들보다 새로운 packet에 대한 적응성이 뛰어남을 확인하였다.

본 논문에서는 공격과 정상 사용의 이진 분류에 의한 침입 탐지 모형을 구축하였으나 공격을 다시 여러 유형으로 나누게 되면 이진 분류를 넘어가게 된다. 이런 경우 현재 Zhu 등이 활발하게 연구하고 있는 통계적 학습 모형인 IVM(Import Vector Machine)[15]을 본 논문의 제안 모형에 적용하여 다중 분류 모형에 의한 침입 탐지 모형을 구축할 수 있을 것이다. 이는 향후 연구과제로 남긴다.

참 고 문 헌

[1] A. Benhur, D. Horn, H. T. Siegelmann, V. Vapnik, "A support vector clustering method," Proceedings. 15th International Conference on Pattern Recognition, Vol.2, pp.724-727, 2000.

[2] G. Casella, R. L. Berger, "Statistical Inference," Duxburt Press, 1990.

[3] N. Cristianini, J. S. Taylor, "An Introduction to Support Vector Machine," Cambridge University Press, 2000.

[4] C. Cortes and V. N. Vapnik, "Support vector networks," Machine Learning, Vol.20, pp.273-297, 1995.

[5] S. M. Emran, M. Xu, N. Ye, Q. Chen, X. Li, "Probabilistic techniques for intrusion detection based on computer audit data," IEEE Transactions on Systems, Man and Cybernetics, Part A, Vol.31, pp.266-274, 2001.

[6] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning," Springer, 2001.

[7] 전명식, "수리통계학," 자유아카데미, 1996.

[8] W. Lee, S. J. Stolfo, K. W. Mok, "A data mining framework for building intrusion detection models," Proceedings of the 1999 IEEE Symposium on Security and Privacy, pp.120-132, 1999.

[9] 이한성, 임영희, 박주영, 박대희, "SVM과 클러스터링 기반 적응형 침입 탐지 시스템", 퍼지및지능시스템학회논문지, 2003.

[10] M. Pontil and A. Verri, "Properties of support vector machine," M. I. T. AI Memo, No.1612, 1997.

[11] 유신근, 이남훈, 신영철, "침입탐지시스템 평가 방법론", 정보처리학회논문집, Vol.7, No.11, pp.3445-3461, 2000.

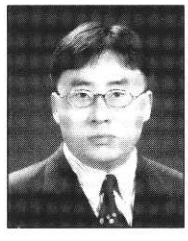
[12] V. N. Vapnik. "The Nature of Statistical Learning Theory," New York, Springer-Verlag, 1995.

[13] V. N. Vapnik, "Statistical Learning Theory," New York : Wiley, 1998.

[14] N. Ye, X. Li. "scalable clustering technique for intrusion signature recognition," 2001 IEEE Man Systems and Cybernetics Information Assurance Workshop, West Point, NY, June, 2001.

[15] J. Zhu, T. Hastie, "Kernel Logistic Regression and the Import Vector Machine," NIPS2001 conference, Vancouver, November, 2001.

[16] Lincoln Laboratory, Massachusetts Institute of Technology, http://www.ll.mit.edu/IST/ideval/data.



전 성 해

e-mail : shjun@chongju.ac.kr

1993년 인하대학교 통계학과(학사)

1996년 인하대학교 대학원 통계학과 (이학석사)

2001년 인하대학교 대학원 통계학과 (이학박사)

2001년~현재 서강대학교 대학원 컴퓨터학과 공학박사수료

1996년~1997년 효성그룹 전자통신연구소 연구원

2000년~2001년 NCR Korea 데이터마이닝 컨설턴트

2002년~2003년 경기대학교 정보통신대학원 겸임교수

2003년~현재 청주대학교 통계학과 전임강사

관심분야 : 유비쿼터스, 데이터마이닝, 지능형 에이전트, 기계학습