실시간 침입탐지 시스템을 위한 새로운 특징랭킹과 특징선택 프레임워크에 대한 연구

A new feature ranking and feature selection framework for realtime IDS

이상재, 김세헌
대전광역시 유성구 구성동 373-1 KAIST 산업및시스템공학과
E-mail : sjlee@tmlab.kaist.ac.kr, shkim@kaist.ac.kr

Abstract

인터넷의 보급에 따라 네트워크를 통한 공격에 피해가 급증하고 있다. 이러한 네트워크 침해를 막기위해 여러 연구자들은 침입탐지 시스템(IDS)을 제안하였으나, 시스템의 탐지율에만 초점을 맞추고 있기 때문에 실시간(Realtime)으로 동작하지 못하고 있다. 실시간 IDS를 위하여 최근 다양한 특징선택(Feature selection)들이 제안되고 있다. 본[1] 논문에서는 특징들을 중요도의 순위를 정하는 새로운 랭킹 방법과 이 방법에 따라서 특징을 선택하는 특징 선택 알고리즘을 제안한다. 또한 제안된 알고리즘을 통하여 선택된 특징을 사용할 경우 탐지결과가 우수함을 실험으로 보여주고 있다.

1. Introduction

최근 인터넷의 보급에 따라서 많은 종류의 공격(Intrusion)들이 발생하게 되었고, 이에 따라서 네트워크를 통한 피해가 급증하고 있다. 다양한 네트워크를 통한 공격의 피해를 줄이기 위해서 침입 탐지 시스템 (Intrusion Detection System)이 제안 되었다. 침입 탐지 시스템은 크게 오사용탐지(Misuse Detection)과 이상탐지(Anomaly Detection)으로 분류된다. 오사용탐지 시스템은 알려진 공격에 대하여 빠르고 보다 정확하게 탐지하지만, 새로운 공격에 대하여 무방비한 특징을 가지고 있다. 이상탐지시스템은 알려진 공격에 대하여 탐지율을 다소 낮지만, 알려지지 않은 유해한 트래픽을 탐지하는 시스템이다.

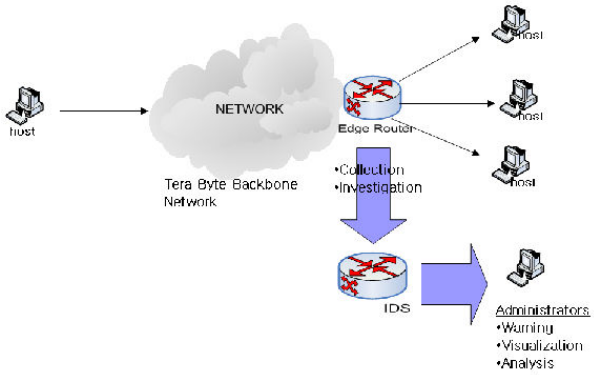

그림 1 침입탐지 시스템

최근 인터넷의 속도는 매우 증가하였으며, 공격 유형들 또한 매우 급속도로 증가하였다. 트래픽량과 공격유형의 증가에 따라서 침입탐지 시스템은 새로운 데이터를 형성하고 학습함으로써, 현재 네트워크에 적용 될 수 있다. 그러나 많은 침입탐지에 관한 연구들은 탐지율에

"본 연구는 지식경제부 및 정보통신연구진흥원의 대학IT연구센터 지원사업의 연구결과로 수행되었음" (IITA-2008-C1090-0801-0016)

초점을 맞추고 있기 때문에 변화한 환경에 대처할 수 없으며, 향상된 네트워크 속도에 대응하는 실시간(Realtime)시스템을 구축하기에는 힘들다.

침입탐지 시스템에 사용되는 모든 특징이 탐지에 큰 도움을 주는 것이 아니며, 일부 특징들은 매우 도움이 되고, 일부는 크게 영향을 미치지 않으며, 일부는 좋지 않은 영향을 미친다. 이와 같은 문제를 해결하기 위해서 몇몇 연구들은 특징추출(Feature Extraction)방법을 사용하였다[1][2]. 특징추출방법은 기존의 특징들로부터 변환을 통하여 탐지에 영향력이 큰 소수의 새로운 특징을 생성한다. 소수의 특징만을 사용하기 때문에 침입탐지 시스템의 연산량이 감소되고 좀더 빠르고 간단한 연산을 통하여 탐지가 가능하게 된다. 그러나 이러한 방법들은 새로운 특징을 생성하는 과정에서 매 순간마다 특징추출 연산을 필요로 하며, 많은 특징들을 사용하여 소수의 특징을 새롭게 생성해 내는 방법이므로 기본적으로 많은 수의 특징을 요구한다. 따라서 수집되고 처리해야 하는 특징의 수는 줄어들지 않게 되며, 매우 고속화된 네트워크에서 이와 같은 많은 수의 특징을 추출하는 것은 점점 힘든 일이 되어가고 있다.

많은 특징들로부터 새로운 특징을 생성하는 특징추출 방법과는 다른 특징선택(Feature Selection)방법은 탐지에 도움이 되는지 아닌지의 여부에 따라서 특징 자체를 선택한다. 이러한 특징선택 방법을 사용하게 되면, 어떠한 성질을 가진 특징이 탐지에 도움이 되는지 분석할 수 있다. 또한, 고속의 네트워크에서 수집해야하는 특징의 수가 줄어들고, 침입탐지 시스템에서의 연산량 또한 줄어들게 된다. H.Gunes Kayacik et al.[3]은 Information gain을 계산하고 이를 통하여 침입탐지 특징간의 연관성을 분석하였다. Yang et al.[4]은 Information Gain과 Chi-square 통계량을 동시에 사용하여 특징선택을 수행 하였고, 이를 통하여 매우 가벼운 침입탐지시스템 개발을 시도하였다. Zorana Bankovie et al.[5]은 PCA를 구하면서 사용되는 고유값을 사용하여 특징선택을 수행하였다. Gary Stein et al.[6]은 유전알고리즘을 사용하여 특징을 선택한 후 의사결정트리를 사용하여 공격을 분류하였다.

본 논문에서는 침입탐지를 위한 특징을 선택하기 위하여 새로운 랭킹방법을 제안한다. 제안된 랭킹방법은 매우 단순하여 사용하기가 쉽다. 특징들의 중요성에대한 랭킹을 이용하여 특징선택을 하고, 선택된 소수의 특징만을 사용하여 실시간 침입탐지이 가능한 시스템을 제안한다.

2. Ranking criteria

제안하는 특징선택 방법은 특징랭킹을 통해서 수행이 된다. 새로운 특징랭킹은 아래의 그림2와 같은 기준을 가지는 측정치(measurement)를 만족하도록 하였다.

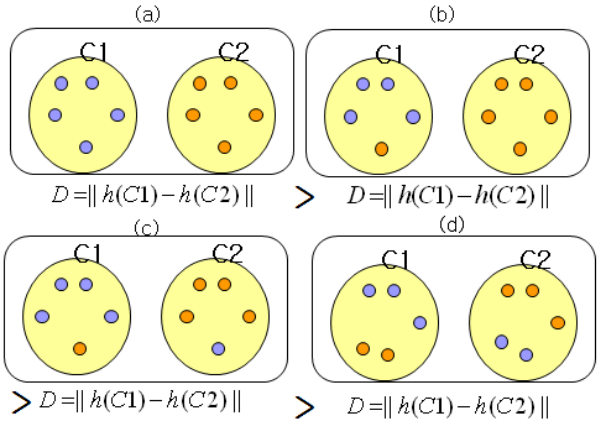

그림 2 Ranking criteria

그림2와 같이 두 개의 그룹 C_1과 C_2에서 뽑은 자료가 있다. 그림에서 파란색원과 주황색원은 하나의 상수 a, b라고 하자. a와 b는 다음과 같이 전처리 되었음을 가정한다.

$$a \neq b, \quad a \geq 0, \quad b \geq 0 \tag{1}$$

그림(2)(a)에서는 C_1에서는 상수a만 뽑혀 나오고 C_2에서 뽑은 수는 상수 b뿐인임을 보여주고 있다. 함수 h는 각 군집을 설명하는 함수이다. 이렇게 서로 다른 군집에서 중첩되지 않고 서로 다른 값만을 발생할 때 랭킹측정치(Ranking measurement) D는 높은 값을 보여야 한다. 그림(2)의 (b),(c) 그리고 (d)로 이동하면서 순차적으로 각 군집에서 획득한 데이터가 서로 중첩함을 보이고 있다. 이렇게 중첩한 데이터를 발생할 경우 랭킹측정치는 많은 중첩을 발생 할수록 작은 값을 보인다.

3. Proposed Algorithm

이미 설명한 바와 같이 제안하는 알고리즘은 각 그룹에서 획득한 데이터가 얼만큼 순수하게 분류 되고, 이에 따라서 함수 h가 군집을 얼마나 잘 설명하는지에 따라서 특징의 랭킹값이 결정이 되어진다. 본 논문에서 위에서 설명한 기준을 만족하는 가장 단순한 함수로써 h를 다음과 같이 정의하였다.

$$h(C_k) = \sum_{i \in C_k}^{N_k} x_i, \tag{2}$$

C_k, N_k, $\mathbf{x}_i$는 각각 k번째 군집, k번째 군집의 데이터 개수, 데이터의 값을 표현하고 있다. 식(2)과 같이 함수 h를 정의하였을 경우 그림(2)의 (a)-(d)는 다음과 같이 만족됨을 알 수 있다. 식(1)의 가정을 사용하여 식(3)의 (a)>(b)>(c)>(d) 됨을 간단히 계산할 수 있다. 따라서 제안하는 함수 h를 사용하는 랭킹측정치는 식(4)와 같이 표현된다.

$$\begin{aligned}
&\text{(a)} \quad D = \| \sum_{i \in C_1}^{N_1} x_i - \sum_{i \in C_2}^{N_2} x_i \| \\
&\text{(b)} \quad D = \| (\sum_{i \in C_1}^{N_1 - 1} x_i - \sum_{i \in C_2}^{N_2 - 1} x_i) + (b - b) \| \\
&\text{(c)} \quad D = \| (\sum_{i \in C_1}^{N_1 - 2} x_i - \sum_{i \in C_2}^{N_2 - 2} x_i) + (a - a) + (b - b) \| \\
&\text{(d)} \quad D = \| (a - b) + 0 + 0 + 0 + 0 \|
\end{aligned} \tag{3}$$

$$\begin{aligned}
D &= \| h(C_1) - h(C_2) \| \\
&= \| \sum_{i \in C_1}^{N_1} w^T x_i - \sum_{i \in C_2}^{N_2} w^T x_i \|
\end{aligned} \tag{4}$$

w는 여러 특징을 사용할지 아닐지를 정하는 indicator 변수이다($w \in \{0,1\}$). 식(4)에서 L_2 Norm을 사용하면, 식(4)는 다음과 같이 쓸 수 있다.

$$(\sum_{i \in C_1}^{N_1} w^T x_i - \sum_{i \in C_2}^{N_2} w^T x_i)^2 \tag{5}$$

식(5)를 각각의 변수로 나누어 표현하면,

$$\begin{bmatrix} (\sum_{i \in C_1}^{N_1} w_1 \delta_{1,c1}{}^T - \sum_{i \in C_2}^{N_2} w_1 \delta_{1,c2}{}^T)^2 \\ \vdots \\ (\sum_{i \in C_1}^{N_1} w_d \delta_{d,c1}{}^T - \sum_{i \in C_2}^{N_2} w_d \delta_{d,c2}{}^T)^2 \end{bmatrix} \tag{6}$$

$$= \begin{bmatrix} w_1 (\sum_{i \in C_1}^{N_1} \delta_{1,c1}{}^T - \sum_{i \in C_2}^{N_2} \delta_{1,c2}{}^T)^2 \\ \vdots \\ w_d (\sum_{i \in C_1}^{N_1} \delta_{d,c1}{}^T - \sum_{i \in C_2}^{N_2} \delta_{d,c2}{}^T)^2 \end{bmatrix}, \tag{7}$$

δ_d는 d번째 특징에 대한 데이터들로 이루어진 행벡터이다. 만약 $N_1=N_2=...N_k$ 이라면,

$$\frac{D}{N^2} = w^T \begin{bmatrix} (\sum_{i\in C_1}^{N_1} \delta_{1,c1} - \sum_{i\in C_2}^{N_2} \delta_{1,c2})^2 / N^2 \\ \vdots \\ (\sum_{i\in C_1}^{N_1} \delta_{d,c1} - \sum_{i\in C_2}^{N_2} \delta_{d,c2})^2 / N^2 \end{bmatrix} \quad (8)$$

$$= w^T \begin{bmatrix} (\mu_1^1 - \mu_1^2)^2 \\ \vdots \\ (\mu_d^1 - \mu_d^2)^2 \end{bmatrix} \quad (9)$$

따라서 제안한 랭킹기준에 따르는 단순한 함수 h를 정의하였을 경우, 식(9)와 같이 단순한 결과를 보여주고 있다. 식(9)에의하면 각 군집의 평균의 거리차 이에 따라서 특징들의 중요도가 결정된다고 할 수 있다.

4. Experimental results

제안한 특징선택 방법이 침입탐지시스템에 미치는 결과를 실험하기 위하여 KDD 99[4] 침입탐지 데이터를 사용하여 실험하였다. KDD 99 데이터셋은 침입탐지시스템의 서로 다른 방법간의 밴치마킹을 제공하도록 설계되었다.

Label	Network Data Features	Label	Network Data Features	Label	Network Data Features
A	duration	O	su_attempted	AC	same_srv_rate
B	protocol_type	P	num_root	AD	diff_srv_rate
C	service	Q	num_file_creations	AE	srv_diff_host_rate
D	flag	R	num_shells	AF	dst_host_count
E	src_byte	S	num_access_files	AG	dst_host_srv_count
F	dst_byte	T	num_outbound_cmds	AH	dst_host_same_srv_rate
G	land	U	is_host_login	AI	dst_host_diff_srv_rate
H	wrong_fragment	V	is_guest_login	AJ	dst_host_same_src_port_rate
I	urgent	W	count	AK	dst_host_srv_diff_host_rate
J	hot	X	srv_count	AL	dst_host_serror_rate
K	num_failed_login	Y	serror_rate	AM	dst_host_srv_serror_rate
L	logged_in	Z	srv_serror_rate	AN	dst_host_rerror_rate
M	num_compromised	AA	rerror_rate	AO	dst_host_srv_rerror_rate
N	root_shell	AB	srv_rerror_rate		

[표1] KDD 데이터셋의 특징(feature)

표1은 KDD 데이터셋이 포함하고 있는 특징들을 보여주고 있다. 위와 같이 KDD는 총 41개의 특징을 포함하고 있다. 제안한 알고리즘을 통하여 41개의 특징들간의 중요도를 계산한 결과가 그림3에 보여지고 있다; 가로축은 각 특징들이며 세로축은 중요도를 나타낸다. 또한 이 β값으로 특징들을 랭킹을 매길 수 있으며, 랭킹 순위가 높은 소수의 특징들을 사용함으로써, 특징선택을 수행하게 된다.

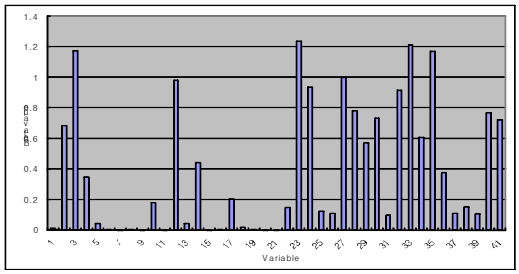

그림 3 특징들의 중요도 (β)

5. Conclusion

인터넷의 사용이 보편화 되어감에 따라서 네트워크를 통한 침해문제는 매우 시급하고도 중요한 문제가 되었다. 네트워크를 통한 침입은 개인정보에서 상업적인 문제에 이르기까지 많은 피해의 발생을 야기한다. 이러한 침입을 방지하기 위해 침입탐지시스템이 제안되고 있으나, 고속화 되어가는 네트워크와 공격의 다양성 때문에 실시간으로 적용되지 못하고있다.

본 논문에서는 침입탐지율은 높은 수준으로 유지시키며, 실시간으로 적용되는 침입탐지시스템을 개발하기 위하여 특징랭킹과 선택 방법을 제안하였다. 제안된 방법은 매우 간결하여 쉽게 사용할 수 있다. 특징간의 중요도의 순위가 밝힘으로써, 특정 공격은 어떠한 특징에 의존함을 분석할 수 있다. 또한, 소수의 특징만을 사용하기 때문에 네트워크 장비에서 수집하고 분석하는 양이 줄어들게 된다.

6. References

[1]Dong Seong Kim,HaNam Nguyen, Thandar Thein, Hong Sou Park,"An Optimized Intrusion Detection System Using PCA and BNN",APPSITT2005.

[2] Xin Xu, Xuening Wang, "An adaptive network intrusion detection method based on PCA and support vector machines", International conference on advanced data mining and application 2005.

[3]H.Gunes Kayacik, A. Nur Zincir-Heywood, Malcolm I. Heywood,"Selecting Features for intrusion detection : Afeature relevance analysis on KDD 99 Intrusion Detection Datasets",www.cs.dal.ca/projectx.

[4]Yang Li, Bin-Xing Fang, You Chen, Li Guo, "A Lightweight Intrusion Detection Model Based On Feature Selection and Maximum Entropy Model", ICCT06.

[5] Zorana bankovie, Dusan Stepanovie, Slobodan Bojanic, Octavio Nieto-Taladriz,"Improving network security using genetic algorithm approach",Volume 33, Issues 5-6, September-November 2007, Pages 438-451, Security of Computers & Networks.

[6]Gary Stein, Bing Chen, Annie S. 쪄, Kien A. Hua,"Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection", ACM Proceedings of the 43rd annual Southeast regional conference,vol2,2005.