실시간 공격 탐지를 위한 Pearson 상관계수 기반 특징 집합 선택 방법★

강 승 호*, 정 인 선**, 임 형 석**

요 약

기계학습을 이용하는 침입 탐지 시스템의 성능은 특징 집합의 구성과 크기에 크게 좌우된다. 탐지율과 같은 시스템의 탐지 정확도는 특징 집합의 구성에, 학습 및 탐지 시간은 특징 집합의 크기에 의존한다. 따라서 즉각적인 대응이 필수인 침입 탐지 시스템의 실시간 탐지가 가능하도록 하려면, 특징 집합은 크기가 작으면서도 적절한 특징들로 구성하여야 한다. 본 논문은 실시간 탐지를 위한 특징 집합 선택 문제를 해결하기 위해 사용했던 기존의 다목적 유전자 알고리즘에 특징 간의 Pearson 상관계수를 함께 사용하면 탐지율을 거의 낮추지 않으면서도 특징 집합의 크기를 줄일 수 있음을 보인다. 제안한 방법의 성능평가를 위해 NSL_KDD 데이터를 사용하여 10가지 공격 유형과 정상적인 트래픽을 구별하도록 인공신경망을 설계, 구현하여 실험한다.

A Feature Set Selection Approach Based on Pearson Correlation Coefficient for Real Time Attack Detection

Kang Seung-Ho*, Jeong In-Seon**, Lim Hyeong-Seok**

ABSTRACT

The performance of a network intrusion detection system using the machine learning method depends heavily on the composition and the size of the feature set. The detection accuracy, such as the detection rate or the false positive rate, of the system relies on the feature composition. And the time it takes to train and detect depends on the size of the feature set. Therefore, in order to enable the system to detect intrusions in real-time, the feature set to be used should have a small size as well as an appropriate composition. In this paper, we show that the size of the feature set can be further reduced without decreasing the detection rate through using Pearson correlation coefficient between features along with the multi-objective genetic algorithm which was used to shorten the size of the feature set in previous work. For the evaluation of the proposed method, the experiments to classify 10 kinds of attacks and benign traffic are performed against NSL_KDD data set.

Key words : Intrusion Detection System, Artificial Neural Network, Multi-objective Genetic Algorithm, Feature selection, Pearson Correlation Coefficient, NSL_KDD data set

접수일(2018년 12월 10일), 게재확정일(2018년 12월 26일)

★ 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임. (No.NRF-2018R1D1A1B07043141)

* 동신대학교/정보보안전공

** 전남대학교/전자컴퓨터공학부

1. 서 론

정보통신 기술의 혁신적 발달에 힘입어 방대한 규모의 컴퓨터 응용프로그램 개발이 진행되고 이와 함께 네트워크의 기하급수적 확산을 가져 왔다. 이와 같은 응용프로그램의 개발 및 네트워크의 확산은 기업의 생산성 향상뿐 아니라 소비자의 소비행태의 변화 등 다양한 가치를 창출하는데 크게 이바지 하고 있다. 하지만 긍정적인 현상 이면에 개인 및 집단의 이익을 위해 불법적인 보안 침해 사고 또한 꾸준히 증가 하고 있다. 보안 침해 사고는 규모면에서의 증가뿐만 아니라 새로운 방법들이 끊임없이 등장하고 있다.

사이버 공격으로 발생하는 피해 규모는 이미 자연재해의 피해 규모를 넘어서고 있는 실정이다[1]. 따라서 사이버 공격을 조기에 탐지하고 적절한 방어 조치 및 사후 조치를 취할 수 있는 공격 탐지/방지 시스템(IDS/IPS)의 개발을 위해 많은 연구들이 진행되어 왔다. 하지만 APT(Advance Persistent Threat, 지능형 지속위협)공격과 같은 새로운 공격이 등장하면서 시그니처 기반의 기존 탐지 시스템들은 탐지 능력에 대해 끊임없이 의문이 제기되고 있다. 이런 상황을 극복하기 위해 기계학습 등 최근 급격히 발전하고 있는 인공지능 기반의 공격 탐지 시스템에 대한 연구에 많은 연구자들의 관심이 집중되고 있다[2-5].

공격 탐지 시스템이란 정상적이지 않은 네트워크 또는 호스트의 사용을 실시간이나 사후에 탐지하는 시스템으로, 방어하고자 하는 공격 방법에 따라 크게 두 가지로 분류된다[6]. 우선 기존에 알려진 공격들을 탐지하기 위한 오사용(misuse detection) 탐지 방식이 있다. 이 접근 방법의 특징은 사전에 정의된 규칙들을 이용해 잘 알려진 공격들을 탐지하는 방식으로 현장에서 가장 많이 사용되는 방법이다. IP 또는 포트 블랙 리스트를 이용해 특정 연결 시도를 제어하는 통상의 방화벽 등에서 사용하고 있다. 하지만 기존에 알려진 공격 방법들에 제한되어 있고, 변종 및 새로운 공격 방법에 적합한 규칙 생성이 쉽지 않아 이러한 공격에 대처하기 어려운 단점이 있다. 다른 탐지 방법은 이상 징후(anomaly detection) 탐지 방식이다. 이상 징후 탐지 방식은 정상인 사용 패턴을 근거로 이를 벗어난 행위를 이상 행위로 간주하는 방식으로 오사용 탐지 방식과 다르게 이전에 알려지지 않은 공격 방법에 대해서도 탐지할 수 있는 장점이 있다. 하지만 이상 징후 탐지 방식은 정상적인 사용 형태를 사이버 침해로 잘못 판단하는 문제점이 지적되고 있다. 최근 데이터 마이닝이나 인공지능, 기계학습 방법을 침입탐지 시스템에 도입하는 방법들이 연구자들의 관심을 모으고 있다. 특히 기존에 알려지지 않은 공격패턴을 찾고자하는 이상 징후 탐지 방식에 인공지능기술을 도입하는 시도가 많이 이루어지고 있으며 상당히 높은 정확성을 보여주고 있다. 본 연구에서도 이상 징후 탐지 시스템을 대상으로 한다.

한편 데이터 수집 및 방어 위치에 따라 호스트 기반 침입탐지 시스템(Host-based IDS)과 네트워크 기반 침입탐지 시스템(Network-based IDS)으로 구분한다[6]. 호스트 기반 시스템은 특정 서버나 장비를 대상으로 로그 파일, 시스템 콜, 파일 접근, 메모리 내용 등의 데이터를 이용해 대상 서버나 장비에 대한 공격을 사전에 차단하거나 사후에 확인하는 시스템이다. 반면, 네트워크 기반 시스템은 네트워크 장비를 통해 전송되는 패킷이나 패킷 트래픽을 대상으로 데이터 링크, 네트워크, 전송 계층의 패킷 정보를 주로 사용해 공격 여부를 판단하는 시스템을 말한다.

본 논문은 TCP 연결을 대상으로 네트워크 트래픽을 분석해 공격 여부를 탐지하는 네트워크 기반의 이상 징후 탐지 시스템을 대상으로 한다. 네트워크 기반의 이상 징후 침입 탐지 시스템은 호스트 기반의 탐지 시스템과 달리 대용량의 네트워크 트래픽을 대상으로 컴퓨팅 자원의 제약 아래서 침입 여부를 탐지해야 하는 어려움이 있다. 특히 기계학습에 기반한 시스템을 이용해 이상 징후를 탐지하는 경우 특징 추출을 위해 많은 전처리 시간을 요구하는 경우 실시간으로 탐지하기 어려운 문제에 직면할 수 있다[7]. 따라서 공격 탐지가 실시간으로 가능하려면 무엇보다 네트워크 트래픽의 성질을 대표하는 특징 집합의 크기가 작아야 한다.

특징 선택을 위한 접근법은 개별 특징의 성질을 기준을 하는 필터(filter) 접근법과 사용하는 기계학습 방법의 성능을 기준으로 하는 래퍼(wrapper) 접근법이 있다[8-10]. 일반적으로 필터 접근법에 비해 래퍼 접근법의 인식률이 높은 것으로 알려져 있지만 많은 시

간을 요한다는 단점이 있다. 최근 [11]은 기존 래퍼 방식의 이러한 단점을 해결하기 위해 특징 집합의 선택 기준을 특정 기계학습 방법의 정확도 대신 클러스터링의 정확도를 선택 기준으로 이용하고 특징 집합의 크기까지 동시에 고려하는 다목적 유전자 알고리즘을 제안하였다.

본 연구에서는 [11]이 제시한 다목적 최적화 문제 해결 방법에 특징들 사이의 상관관계를 추가로 고려하면 이상 징후 탐지 시스템의 정확성을 거의 낮추지 않으면서도 학습 및 탐지 시간을 줄일 수 있음을 보이고자 한다.

2. NSL_KDD 데이터

NSL_KDD 데이터[12,13]는 MIT Lincoln Lab.에서 DARPA 침입탐지 평가를 위해 생성한 1999 KDD CUP 데이터[14]를 수정, 보완한 것으로 보안 데이터 과학자들에게 가장 널리 사용되고 있는 데이터 집합이다. 총 38가지의 공격 유형으로 구성되어 있으나, 학습 데이터에는 24가지 공격 유형만이 포함되어 있다.

대체적으로 기존의 연구들은 공격 유무만을 판별하는 2진 문제 혹은 공격의 4가지 범주만을 판별하는 문제를 다루어 왔다. 이처럼 적은 수의 클래스를 분류하는 경우 대부분의 방법들이 지나치게 높은 인식률을 보여 주어 방법 간의 성능 차를 비교하는데 무리가 따른다. 이러한 결가는 NSL_KDD 데이터의 한계에 기인한 측면도 있다고 생각된다. 따라서 제안한 특징 추출 알고리즘은 6가지 DoS 공격 이외에 훈련 집합의 레코드 수가 1000개 이상인 satan, ipsweep, portsweep, nmap 공격을 추가하여 총 10가지 공격을 대상으로 실험하였다. 표 1은 실험에 사용한 정상 레코드와 10가지 공격 유형의 레코드 수를 보여준다.

전처리 과정의 하나인 특징들에 적용한 정규화 방법도 [11]에서 사용한 방법과 거의 동일한 방법을 사용하였다. 이 방법은 M. Sabhanani[15]이 제시한 방법으로 우선 특징의 유형에 따라 4가지로 나누고 각 유형에 따라 정규화를 다르게 한다. 자세한 전처리 절차는 [11]을 참고하기 바란다. 다만, 본 논문에서는 위의 방법을 적용한 결과 특징들 중 src_bytes와 dst_bytes는 정규화 후의 값의 범위가 각각 [0, 9.14], [0, 9.12]로 다른 특징들에 비해 여전히 높은 값을 갖는 것을 확인하였다. 따라서 두 특징이 [0, 1] 사의 값을 갖도록 추가로 최소-최대 정규화를 적용하여 모든 특징들이 [0, 1]사이의 값을 갖도록 하였다.

<표 1> NSL_KDD 데이터에 있는 정상 레코드 수와 10가지 공격 유형 레코드 수 구성

공격유형	Training Data	Test Data
normal	67344	9711
neptune	41214	4657
teardrop	892	12
smurf	2646	665
pod	201	41
back	956	359
land	18	7
satan	3633	735
ipsweep	3599	141
portsweep	2931	157
nmap	1493	73

3. 특징 선택 알고리즘

3.1 다목적 유전자 알고리즘

본 논문이 제안한 특징 선택 방법은 [11]이 이전에 제시한 특징 선택 알고리즘인 다목적 유전자 알고리즘에 특징 간의 Pearson 상관계수를 이용하는 방법을 추가한 것이다. 따라서 이 절에서 우선 간략하게 다목적 유전자 알고리즘을 설명하겠다. 자세한 내용은 [11]을 참조하기 바란다.

특징 집합은 각 특징이 포함되는지에 따라 0, 1이 할당된 41차원의 벡터 S로 표현되면 알고리즘의 해로 사용된다[11].

$$S = <s_1, s_2, s_3, ..., s_i, ..., s_{41}>, \ s_i \in \{0, 1\}$$

전통적인 래퍼 방식이 가지고 있는 속도 문제를 해결하기 위해 특정 기계학습의 정확성 대신 k-평균 클러스터링을 특징 집합의 적합성 평가를 위해 사용한다. 주어진 특징 벡터 S를 이용해 k-평균 클러스터링을 실행한 후 데이터 x의 소속 클래스와 원 소속 클래스의 동일성을 확인해 0 혹은 1값을 $\omega(x)$로 부여한다. 각 데이터를 대상으로 $\omega(x)$를 구하고 이를 모두 더한

후 전체 데이터 수 N으로 나눈 비율을 정확성을 위한 목적함수로 정의한다[11].

$$Fit_{detect}(S) = \sum_{i=1}^{N} \omega(x_i)/N \quad (1)$$

특징 집합의 시간과 관련이 있는 목적 함수는 식(2)와 같이 특징 벡터의 크기, $\phi(S)$에 반비례하도록 정의한다[11].

$$Fit_{\leq n}(S) = (41 - \phi(S))/41 \quad (2)$$

최종 목적함수는 식(3)과 같이 두 가지 목적 함수의 중요도에 따라 매개변수 λ를 이용한 가중치 합으로 정의한다[11].

$$Fit(S) = \lambda Fit_{detect}(S) + (1-\lambda) Fit_{time}(S),\ 0 \leq \lambda \leq 1 \quad (3)$$

3.2 Pearson 상관계수를 이용한 특징 선택

대용량의 네트워크 트래픽을 대상으로 실시간 탐지를 위해서는 적절한 기계학습 알고리즘을 선택하는 것 외에 정확성을 유지하면서도 크기가 작은 특징 벡터를 선택해야 한다. [11]에서는 최적 특징 집합을 추출하기 위해 유전자 알고리즘을 사용하였다. 하지만 선택된 특징 간의 Pearson 상관계수를 활용하면 탐지 시스템의 정확성을 해치지 않으면서도 특징 집합의 크기를 더욱 줄일 수 있다. Pearson 상관계수는 식(4)로 정의되며 [-1, 1] 사이의 값을 갖는다.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (4)$$

여기서, cov는 공분산, σ_X는 X의 표준 편차, 그리고 σ_Y는 Y의 표준편차를 각각 나타낸다. 두 특징 간의 상관계수가 -1 값에 가까우면 음의 상관관계에, +1 값에 근접하면 양의 상관관계에 있다고 할 수 있고, 0에 가까운 값을 갖는 경우엔 유의미한 상관관계가 없다고 판단할 수 있다.

<표 2> Pearson 상관계수가 0.7 이상인 특징들

특징	Pearson 상관계수가 0.7 이상인 특징들
duration	
protocol_type	
service	
flag	
src_bytes	dst_bytes, logged_in, same_srv_rate
dst_bytes	src_bytes, logged_in, dst_host_srv_count, dst_host_same_srv_rate
land	
wrong_fragment	
urgent	
hot	is_guest_login
num_failed_logins	
logged_in	src_bytes, dst_bytes
num_compromised	num_root
root_shell	
su_attempted	
num_root	num_compromised
num_file_creations	
num_shells	
num_access_files	
num_outbound_cmds	
is_host_login	
is_guest_login	hot
count	
srv_count	
serror_rate	srv_serror_rate, dst_host_serror_rate, dst_host_srv_serror_rate
srv_serror_rate	serror_rate, dst_host_serror_rate, dst_host_srvserror_rate
rerror_rate	srv_rerror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate
srv_rerror_rate	rerror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate
same_srv_rate	src_bytes, dst_host_srv_count, dst_host_same_srv_rate
diff_srv_rate	
srv_diff_host_rate	
dst_host_count	
dst_host_srv_count	dst_bytes, same_srv_rate, dst_host_same_srv_rate
dst_host_same_srv_rate	dst_bytes, same_srv_rate, dst_host_srv_count
dst_host_diff_srv_rate	
dst_host_same_src_port_rate	
dst_host_srv_diff_host_rate	
dst_host_serror_rate	serror_rate, srv_serror_rate, dst_host_srv_serror_rate
dst_host_rerror_rate	serror_rate, srv_serror_rate, dst_host_serror_rate
dst_host_rerror_rate	rerror_rate, srv_rerror_rate, dst_host_srv_rerror_rate
dst_host_srv_rerror_rate	rerror_rate, srv_rerror_rate, dst_host_rerror_rate

모든 특징들 간에 Pearson 상관계수를 계산하면 총 41×(41-1)/2 개의 상관계수를 구하게 된다. 이 중 높은 상관관계에 있는 특징들은 서로 중첩되므로 하나의 특징으로 대표해도 특징 집합을 입력으로 하는 기계학습 기반 탐지 시스템의 정확성을 떨어뜨리지 않을 것이라고 생각할 수 있다. 사전에 높은 상관관계를 정의할 임계치를 정해야 하는데, 0.8 이상의 상관관계를 보여 주는 특징 쌍은 존재하지 않았고 여러 번의 실험을 통해 0.7을 높은 상관관계를 나타내는 상관계수의 임계치로 사용하게 되었다. 표2는 각 특징들에 대해 Pearson 상관계수가 0.7 이상인 특징들을 보여준다. 예를 들어 특징 src_bytes는 dst_bytes, logged_in, same_srv_rate과 0.7 이상의 Pearson 상관계수를 갖는다.

상관계수를 특징 선택에 사용하는 방법은 간단하다. 단지 특징 간의 상관계수가 사전에 정의된 특정 임계치 이상(여기서는 ρ >= 0.7)인 특징 집단에 대해선 이 중 하나를 대표 특징으로 사용하고 다른 특징을 제외시키면 된다. 따라서 각 특징 집합을 대표할 수 있는 최소수의 특징들을 선택하면 특징 벡터의 크기를 최소화 할 수 있다.

한편, 특징 집합을 노드 집합 N으로 표현하고 주어진 임계치 이상의 상관계수를 갖는 특징들 사이를 에지 집합 E로 연결하여 그래프 $G(N,E)$로 표현하면, 그래프로부터 모든 특징을 커버하는 최소수의 특징 집합 N'을 선택하는 최소 지배 집합(Minimum dominating set) 문제임을 알 수 있다. 최소 지배 집합 문제는 NP-hard에 속한다[16]. 특징 간의 상관계수를 이용해 다목적 유전자 알고리즘에 의해 구해진 해의 크기를 축소화하기 위한 방법을 의사코드로 표현하면 표3과 같다.

하지만, 상관계수를 이용해 대표하는 최소수의 특징들을 선택 문제는 복잡도 측면에서 쉬운 문제가 아니다. 최소 지배 집합 문제처럼 NP-hard에 속한 문제는 현재까지 최적해를 보장하는 효율적인 알고리즘이 존재하지 않는다. 따라서 본 논문에서는 Pearson 상관계수 리스트(Corr)만을 사용해 최종해 S'을 구하는 시간복잡도 $O(n^2)$을 갖는 간단한 휴리스틱 알고리즘을 설계하였다. 알고리즘은 유전자 알고리즘에 의해 구해진 특징 벡터를 대상으로 특정 특징이 벡터에 존재하면 해당 특징과 상관계수가 임계치 이상인 특징들을 Pearson 상관계수 리스트(Corr)로부터 찾아 이를 특징 벡터에서 제거하는 과정으로 구성된다. 표4에 알고리즘의 의사코드를 제시한다.

<표 3> Pearson 상관계수를 이용한 특징 집합 최소화 알고리즘

Algorithm: Feature Selection Using Pearson Correlation Coefficient

Input: 다목적 유전자 알고리즘에 의해 구해진 해(S), 임계치(*Threshold*)

Output: Pearson 상관계수 리스트(*Corr*), 최종해(S')

1. *Corr* ← Pearson 상관계수를 계산
2. G ← *Corr*와 *Threshold*를 이용해 그래프 생성
3. N' ← 그래프 G로부터 최소 지배 집합 선택
4. P ← 모든 특징들을 지배 특징 N'들에 배타적으로 분할
5. S' ← 주어진 해 S에 있는 특징들 중 같은 분할에 속하는 특징들은 분할을 대표하는 지배 특징으로 대체

<표 4> Pearson 리스트를 이용해 특징 추출을 위한 휴리스틱 알고리즘

Algorithm: Heuristic for Feature Selection Using PCC List

Input: 다목적 유전자 알고리즘에 의해 구해진 해 특징 벡터 S, Pearson 상관계수 리스트 *Corr*

Output : 최종해 S'

```
1. Begin
2.   for each feature in S
3.     if feature(i) is not 0 # i 번째 특징이 특징
                                벡터에 있는 경우
4.       rm_list ← Corr[i]; # i 번째 특징과
                상관계수가 threshold 이상인 특징 리스트
5.       S[rm_list] ← 0; # 특징 벡터에서 리스트 제거
6. End
```

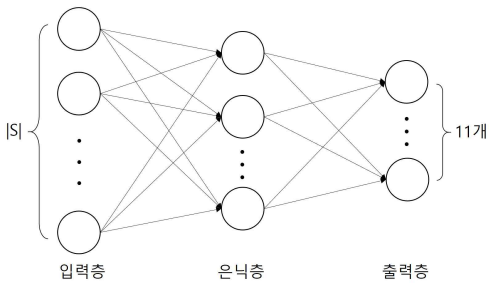
특징선택을 위한 알고리즘은 크게 두 단계의 절차로 구성된다<표 5>. 우선 다목적 유전자 알고리즘을 이용한 최적 특징 집합 선택을 실행한다. 다음엔 첫 번째 절차를 거쳐 선택된 특징 집합을 대상으로 하여 특정 임계치 이상의 Pearson 상관계수를 갖는 특징들은 그 중 하나로 대표한다.

<표 5> Pearson 상관계수를 추가한 전체 특징 선택 알고리즘

Feature Selection Algorithm
1. $S \leftarrow$ Multi-objective genetic algorithm
2. $S' \leftarrow$ Heuristic Algorithm(S)

4. 다중 퍼셉트론

모든 네트워크 트래픽은 앞에서 제시한 특징 선택 알고리즘에 의해 선택된 특징만을 사용해 입력 벡터로 변경된 후 기계학습 모델의 입력으로 사용된다. 본 논문에서는 기계학습 방법 중 대표적인 지도학습 방법의 하나인 인공 신경망을 사용하였다. 인공 신경망의 네트워크 구조는 입력층, 은닉층, 출력층으로 구성된 3계층 다층 퍼셉트론이며 아래 (그림 1)과 같다.



(그림 1) 3계층 다층 퍼셉트론 그림

입력 계층의 노드 수는 사용하는 입력 벡터의 크기와 동일하고 출력 계층은 공격 10가지와 정상 1가지를 대표하기 위해 11개의 노드로 구성하였다. 중간에 위치한 은닉 계층의 노드 수는 다양한 노드 수를 사용해 실험한 후 가장 성능이 좋았던 노드 수, 여기서는 20을 사용하였다. 계층 간 연결은 완전 연결 방식을 사용하였으며 Xavier[17]가 제안한 초기화 방법으로 초기화하였다.

Xavier 초기화는 [$-r$, $+r$] 사이에서 임의의 값을 추출해 각 가중치 값을 초기화하는 방법이다. 여기서 $r = \sqrt{\frac{6}{n_{iinput} + n_{output}}}$ 이고, n_{input}과 n_{output}은 가중치를 초기화해야하는 계층의 입출력 연결 수를 말한다. 은닉 계층과 출력 계층의 노드에서 사용한 활성 함수는 0과 1사의 값을 출력하는 시그모이드 함수 $S(x) = \frac{1}{1+e^{-x}}$를 사용하였다. 비용함수는 식(5)와 같은 교차 엔트로피를 사용하였으며, 경사 하강법 기반의 역전파 알고리즘을 사용하여 신경망을 훈련시켰다. 훈련에 미니 배치 방식을 적용했고 훈련에 의해 개선되는 에러율의 폭이 일정 수준 이하이면 훈련을 중단하였다.

$$C(\Theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_j^i \log(\hat{p_j^i}) \tag{5}$$

여기서, Θ: 파라미터(에지 가중치), n: 배치 크기, k: 클래스 개수, y: i번째 사례가 j 클래스이면 1값을 아니면 0값을 부여, p: 사례의 j 번째 소프트맥스 함수 값을 각각 나타낸다. 식(6)은 소프트맥스 함수이다.

$$p_j = \frac{e^{s(j)}}{\sum_{m=1}^{k} e^{S(m)}} \tag{6}$$

5. 실험 및 결과 분석

설계한 시스템의 성능 분석을 위해 우선 NSL_KDD 학습 데이터와 테스트 데이터를 대상으로 선택된 특징만을 사용해 데이터를 다시 가공하였다. 새로 생성한 학습 데이터를 이용해 인공신경망 기반의 이상 징후 탐지 시스템을 학습시킨 다음 학습된 시스템을 대상으로 테스트 데이터를 이용해 정확도를 측정하였다. 아래 표 6은 이와 같을 실험을 20회 한 후 평균을 구한 것이다.

<표 6> 실험 결과

	No PCC	0.7 PCC
평균 길이	13.5	10.3
훈련 정확도	0.877	0.875
테스트 정확도	0.866	0.861
훈련 시간(s)	39.2	37.66
테스트 시간(s)	0.0047	0.0041

Pearson 상관계수를 사용하지 않고 다목적 유전자 알고리즘만을 사용한 경우(표 6에서 No PCC), 훈련 데이터를 다시 사용해 정확도를 측정한 결과 87.7%의 정확도를 보여 주었고 테스트 데이터를 사용했을 때는 약 86.6%의 정확도를 보여 주었다. 한편 특징 집합의 크기는 평균 13.5로 원래 특징 집합의 크기 41의 약 33%에 불과하였다. 훈련 시간은 39초, 테스트에 걸리는 시간은 0.0047초였다.

다음으로 임계치를 0.7로 한 Pearson 상관계수 방법을 추가해서 실험한 결과, 훈련 정확도와 테스트 정확도가 87.5%와 86.1%로 첫 번째 방법 다목적 유전자 알고리즘만을 사용한 경우와 큰 차이가 없음을 확인하였다. 평균 길이는 10.3으로 약 30% 이상의 개선이 있었고 훈련 시간과 테스트 시간에서도 개선이 있음을 확인 할 수 있다. 물론 훈련 데이터와 테스트 데이터의 규모가 작아 현저한 차이를 확인할 수는 없지만, 실제 현장에서 대규모의 네트워크 트래픽을 대상으로 데이터의 수집 및 전송, 전처리 절차를 비롯해 지속적인 학습 및 탐지 등에 걸리는 시간은 사용하는 특징 집합의 크기에 매우 큰 영향을 받는다.

6. 결론

기계학습을 이용한 이상 징후 탐지 시스템은 다른 기계학습 기반의 예측 시스템과 마찬가지로 특징 집합 구성과 크기에 시스템의 정확성과 훈련 및 탐지 시간이 크게 의존한다.

본 논문은 실시간 이상 징후 탐지를 위해 기존에 제시된 다목적 유전자 알고리즘을 이용한 특징 집합 선택 방법에 특징 간의 Pearson 상관계수를 이용하는 방법을 추가하면 특징 집합의 크기를 크게 줄일 수 있음을 보였다. 다만, Pearson 상관계수를 사용해 주어진 특징 벡터 크기를 최소화하기 위해서는 최소 지배 집합 문제를 해결해야 한다. 본 논문에서는 최소 지배 집합 문제를 해결하는 방법을 직접 사용하지 않고 상관계수 리스트만을 이용해 $O(n^2)$ 시간에 근사해를 구하는 알고리즘을 설계하였다.

NSL_KDD 데이터를 대상으로 10개의 공격과 정상 트래픽을 판별하는 탐지 시스템을 인공신경망을 이용해 구현하였다. 제안한 특징 선택 알고리즘에 의해 추출된 특징을 사용한 경우 다목적 유전자 알고리즘만을 사용하여 얻은 특징 집합을 사용했을 때보다 약 30%의 특징 집합의 크기를 줄일 수 있었다. 이는 정확성에서 거의 손실을 가져오지 않고 달성한 것으로 일일 평균 테라바이트 이상의 대량 트래픽이 발생하는 네트워크에서는 실시간으로 이상 징후를 탐지하기 위한 의미있는 연구결과라 할 수 있다.

앞으로 정확성을 해치지 않으면서도 특징 벡터의 크기를 더욱 줄이기 위한 효율적인 알고리즘을 개발하고자 한다.

참고문헌

[1] 미래창조과학부, “정보보호가 기본이 되고 창조경제 먹거리 산업화를 위한 K-ICT 시큐리티 발전 전략”, 2015.

[2] 이광호, 김종화, 김지원, 윤석준, 김완주, 정찬기, “퍼지추론을 이용한 정량적 사이버 위협 수준 평가방안 연구”, 융합보안논문지, 제18권, 제2호, pp.19-24, 2018.

[3] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu and V. C. M. Leung, “A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View”, IEEE Access. vol.6, pp. 12103-12117, 2018.

[4] Y. Xin, etc. “Machine Learning and Deep Learning Methods for Cybersecurity”, IEEE Access. vol. 6, pp. 35365-35381, 2018.

[5] H. Jiang, J. Nagra and P. Ahammad, “SoK: Applying Machine Learning in Security - A Survey”, arXiv:1611.03186, 2016.

[6] A. A. Ghorbani, W. Lu and M. Tavallaee, ‘Network Intrusion Detection and Prevention’, Springer 2010.

[7] 한명묵, “침입탐지시스템에서의 특징 선택에 대한 연구”, 융합보안논문지, 제6권, 제3호, pp.19-24, 2018.

[8] G. Chandrashekar, F. Sahin, “A survey on feature selection methods”, Computers & Electrical Engineering, Vol. 40, Issue 1, pp. 16-28, 2014.

[9] J. Suto, S. Oniga and P. P. Sitar, “Comparison of wrapper and filter feature selection algorithms on human activity recognition”, Proc. 6th International Conference on Computers Communications and Control (ICCCC), DOI: 10.1109/ICCCC.2016.7496749, 2016.

[10] H. S. Huang, “Supervised feature selection: A tutorial”, Artificial Intelligence Research, Vol. 4, No. 2, 2015.

[11] 김태희, 강승호, 실시간 탐지를 위한 인공신경망 기반의 네트워크 침입탐지 시스템, 융합보안논문지, 제17권, 1호, pp. 31-38, 2017.

[12] NSL_KDD data set. Avalilable on: http://nsl.cs.unb.ca /NSL-KDD/

[13] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Proc. 2009 IEEE Int. Conf. Comput. Intell. Security Defense Appl. CISDA, pp. 53-58, 2009.

[14] KDD Cup 1999. Available on:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 2007.

[15] M. Sabhnani and G. Serpen, “Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context,” Proc. of International Conference on Machine Learning: Models, Technologies, and Applications, pp. 209-215, 2013.

[16] M. R. Garey and D. S. Johnson, ‘Computers and Intractability: A Guide to the Theory of NP-Completeness’, W.H. FREEMAN AND COMPANY, 1979.

[17] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks”, Proc. the 13th International Conference on Artificial Intelligence and Statistics 2010.

〔저자소개〕

강 승 호 (Seung-Ho Kang)
1994년 8월 : 전남대학교 전산학과 학사
2003년 2월 : 전남대학교 전산학과 석사
2009년 8월 : 전남대학교 전산학과 박사
2013년 9월 ~ 현재 : 동신대학교 정보보안전공 교수
email : drminor@dsu.ac.kr

정 인 선 (In-Seon Jeong)
2001년 2월 : 여수대학교 전자계산학과 학사
2006년 2월 : 전남대학교 전산학과 석사
2011년 2월 : 전남대학교 전산학과 박사
2017년 8월 ~ 현재 : 전남대학교 박사후연구원
email : jis0755@gmail.com

임 형 석 (Hyeong-Seok Lim)
1983년 2월 : 서울대학교 컴퓨터공학과 학사
1985년 2월 : 한국과학기술원 전산학과 석사
1993년 8월 : 한국과학기술원 전산학과 박사
1987년 9월 ~ 현재 : 전남대학교 전자컴퓨터공학부 교수
email : hslim@jnu.ac.kr