한국정보전자통신기술학회논문지(jkiiect)'17-02, Vol.10 No.1
논문 17-10-01-38 http://dx.doi.org/10.17661/jkiiect.2017.10.1.38

APEX 기반 침입 탐지 시스템 개발에 관한 연구 : (주)제이드 솔류션과 공동 연구

김병주*

A Study on Developing Intrusion Detection System Using APEX : A Collaborative Research Project with Jade Solution Company

Byung-Joo Kim*

요 약 정보 처리 기술의 컴퓨터 및 네트워크 의존도가 심화됨에 따라 컴퓨터 및 네트워크에 대한 침입 사례가 갈수록 증가하고 있다. 시스템 및 네트워크의 침입을 방지하기 위하여 호스트와 네트워크 기반 침입차단시스템(방화벽 등)이 개발되었지만 기존의 규칙 기반의 침입차단시스템만으로는 보안 관리에 많은 어려움이 있다. 이러한 이유로 인해 시스템 및 네트워크 자원에 대한 침입을 실시간으로 탐지하고 이에 대처하는 침입탐지시스템 개발에 대한 요구가 증가하고 있다. 본 논문에서는 비선형 자료에도 적용 가능하며 수렴성이 보장된 실시간 특징 추출 방법으로 APEX 알고리즘과 점증적 LS-SVM 분류기를 결합한 실시간 침입탐지 시스템을 개발하였다. 일반적으로 실시간 처리 방식은 메모리의 효율성이 좋고 학습 자료의 추가를 허용하는 장점이 있지만 일괄처리 방식에 비해 정확도가 떨어지는 단점이 있다. 따라서 제안한 시스템은 정확도 면에서도 일괄 처리 방식과 비슷한 성능을 나타내고 있어 상용화가 가능한 시스템이다.

Abstract Attacking of computer and network is increasing as information processing technology heavily depends on computer and network. To prevent the attack of system and network, host and network based intrusion detection system has developed. But previous rule based system has a lot of difficulties. For this reason demand for developing a intrusion detection system which detects and cope with the attack of system and network resource in real time. In this paper we develop a real time intrusion detection system which is combination of APEX and LS-SVM classifier. Proposed system is for nonlinear data and guarantees convergence. While real time processing system has its advantages, such as memory efficiency and allowing a new training data, it also has its disadvantages of inaccuracy compared to batch way. Therefore proposed real time intrusion detection system shows similar performance in accuracy compared to batch way intrusion detection system, it can be deployed on a commercial scale.

Key Words : APEX, KDD CUP 99, real time intrusion detection system, incremental LS-SVM, commercial scale

1. 서론

정보 처리 기술의 컴퓨터 및 네트워크 의존도가 심화됨에 따라 컴퓨터 및 네트워크에 대한 침입 가능성 및 침입 사례가 갈수록 증가하고 있다.

시스템 및 네트워크의 침입을 방지하기 위하여 호스트와 네트워크 기반 침입차단시스템(방화벽 등)이 개발되었지만 침입차단시스템만으로는 개별적인 시스템이나 대규모 시스템에 대한 보안 관리에 많은 어려움이 있다. 이러한 배경에서 시스템 및 네트워크 자원에 대한 침입을 실시간으로 탐지

This work was supported by a 2016 research grant from Youngsan University, Republic of Korea.
*Corresponding Author : Department of Computer Engineering, Youngsan University (bjkim@ysu.ac.kr)
Received January 12, 2017 Revised January 24, 2017 Accepted February 08, 2017

하고 이에 대처하는 수단이 요구됨에 따라 침입탐지시스템 개발의 요구가 증가하고 있는 추세이다. 침입탐지 시스템(IDS, Intrusion Detection System)은 컴퓨터 또는 네트워크의 감시를 통해 침입 발생 시 이를 적시에 탐지하고 대응하는 기능을 제공한다. 초기의 침입탐지 시스템들은 이미 알려진 공격 패턴들을 기반으로 전문가 시스템에 인코딩하여 침입 여부를 판단하였다. 이러한 방식은 새로 추가되는 공격에 대한 확장은 매우 어려운 실정이다. 이와 같은 문제점을 해결하기 위해 인공지능 기반의 다양한 침입탐지 시스템이 개발되고 있다. 그러나 현재까지의 연구 결과는 공격 패턴의 분류를 위한 분류기(classifier)의 학습 알고리즘 성능 향상에 많은 연구가 있다. 또한 개발된 시스템은 대부분 일괄처리 방식으로 동작하여 실시간 침입탐지 시스템의 적용에는 적합하지 못하다는 문제점이 있다.

현실세계에 적용 가능한 침입탐지 시스템은 실시간으로 발생하는 공격패턴에 대하여 침입탐지를 수행하여야 하므로 다음과 같은 요소들을 갖추어야 한다.

첫째 전처리 과정에서 실시간 특징추출이 가능해야 한다. 이는 본 연구의 목적이 상용화 제품을 개발하는데 주안점을 두고 있기에 기하급수적으로 늘어나는 침입 패턴에 대해서도 확장성(scalability)이 있으려면 실기간 특징 추출이 가능해야 한다. 또한 효과적인 특징추출은 학습시간의 감소, 분류기의 성능 향상을 기대할 수 있다. 많은 분류 연구 결과에서 많은 수의 특징을 사용하는 것은 계산상의 비효율성 뿐만 아니라 분류 성능도 좋지 않은 결과를 나타낸다. 둘째 분류기에서는 실시간으로 공격패턴을 분류해 내기 위해 일괄처리 방식이 아닌 실시간 분류가 가능해야 한다. 일괄처리 방식의 경우 새로운 학습 자료가 추가 되면 전체 학습 자료에 대해 다시 학습 하여야 하는 단점이 있어 실시간 침입탐지 시스템에는 적절한 분류기가 될 수 없다. 따라서 본 연구에서는 아직 국내에서 상용화 되어 있지 않은 현실 세계에 적용 가능한 침입탐지 시스템을 개발하기 위해 (주) 제이드솔류션과 공동 개발하고 있는 실시간 공격패턴에 대해 점증적으로 특징 추출 및 분류를 할 수 있는 침입탐지 시스템을 개발하는데 연구의 목표를 두고자 한다.

2. 본 론

2.1 기존 연구

침입탐지 시스템에 대한 기존의 연구 및 제품은 대부분 기계학습 및 데이터마이닝 기법을 사용하는 것으로 학습된 자료를 바탕으로 일괄처리 방식의 분류기의 성능을 개선하는데 많은 발전이 이루어졌는데 반해 효율적인 특징 추출 및 실시간 분류기에 관련된 제품 연구는 많이 이루어지지 않았다. 이러한 최근의 동향 중 본 연구에서 중점을 두는 특징 추출 및 분류에 관련된 연구들을 살펴보고 이를 통해 개선방향을 본 연구에 적용하고자 한다.

먼저 특징 추출을 위한 연구를 살펴보면 신경망과 지지벡터기계(support vector machine:SVM)를 결합한 침입탐지 시스템이 있다[1][8][9]. 이는 전체 학습 및 테스트 패턴에서 하나의 특징을 제거한 후 나머지 특징들을 바탕으로 학습을 한 후 이를 테스트 데이터에 적용하여 그 결과를 성능평가 행렬에 기록하는 방식으로 모든 특징들에 대해 이 과정을 반복한다. 기록된 성능평가 행렬을 바탕으로 각 특징의 순위를 매긴 후 중요한 특징들을 추출한 후 이를 SVM의 학습 자료로 사용하여 학습을 한 후 새로운 침입에 대해 공격 유무를 판단하는 침입탐지 시스템을 제안하였다. 하지만 이러한 특징 추출 기법은 특징 추출을 위해 많은 시간을 필요로 하며 실시간 방식이 아닌 일괄처리 방식의 특징 추출 기법이다. 또한 SVM이 최근 여러 분야에서 우수한 분류기로 각광을 받고 있지만 일괄처리 방식의 분류기 이므로 실시간 침입탐지 시스템에는 적합하지 못하다.

분류기에 관련된 연구는 크게 지도학습(supervised learning)을 기반으로 하는 분류와 군집화 기법(clustering technique)에 의한 분류로

나눌 수 있다. 먼저 지도학습에 관한 최근의 연구를 살펴보면 CCA-S 기법이 있다[2]. 이 기법은 기존의 일괄처리 학습에 비해 실시간 학습이 가능하다는 장점을 가지고 있으나 지도학습 방법에 의해 군집화를 수행하므로 학습에 많은 자료가 필요하며 만일 이용 가능한 학습 자료가 없는 경우에는 군집화 작업을 수행할 수 없어 침입탐지 시스템이 동작 하지 못하는 단점이 있다. 이외에 대부분의 지도학습 기반의 분류 연구에서는 SVM을 사용한 방법이 신경망을 사용한 방법에 비해 성능이 우수하다는 연구 결과를 내놓고 있으나 앞에서 언급한 것처럼 일괄처리 방식의 지도학습 분류기는 실시간 침입탐지 시스템에는 적절한 방법이 되지 못한다. 또한 지도학습 방법에서는 반드시 학습과정이 필요한데 이를 위해서는 많은 양의 분류된 데이터(labeled data)를 필요로 한다. 이러한 방대한 양의 학습데이터의 수집 및 분류는 많은 비용이 들며 분류기의 성능은 학습데이터에 의해 좌우되는 단점을 가지고 있다. 이 밖에 기계학습 및 통계학 분야에서 사용하는 기존의 군집화 기법은 대규모 데이터에 적용하기가 어려우며 실시간 군집화가 어렵다. 또한 학습 전에 미리 군집의 개수를 지정해야 하는 단점도 있다. 따라서 현실적으로 새로운 침입 패턴이 계속 발생하는 상황에서 기존의 군집화 기법을 침입탐지 시스템에 적용하기에는 적당하지 않다.

2.2 기존 연구 개발의 문제점

앞에서 열거한 선행 연구 개발의 문제점을 정리하면 아래와 같다.

① 전처리 과정을 고려하지 않음.

침입탐지 시스템에서 분류기의 성능을 향상시키기 위해서는 침입탐지 패킷에서 중요한 특징을 잘 추출하여 이를 분류기의 학습 자료로 사용하는 전처리(preprocessing) 과정이 반드시 필요하나 이제까지 대부분의 연구에서는 이 과정이 없는 경우가 많다. 이러한 전처리를 통해 얻을 수 있는 이점은 다음과 같다.

- 학습 자료의 차원이 줄어 빠른 학습이 이루어진다. 이는 실시간으로 동작되는 침입탐지 시스템에서는 중요한 요소가 된다.

- 분류기의 학습시 성능을 저하시킬 수 있는 중복되는 특징들을 제거하고 중요한 특징들을 추출함으로 인해 분류기의 성능을 향상 시킬 수 있다.

② 점증적인 특징 추출이 이루어지지 않음.

실시간으로 발생하는 침입패턴들에 대해 특징을 추출하려면 점증적 갱신이 가능하여야 한다. 그러나 현재까지의 연구에서는 점증적 특징 추출 기법을 사용한 시스템이 없었다. 따라서 점증적 특징 추출을 할 수 있는 알고리즘의 개발이 필요하다.

③ 점증적 분류를 할 수 있는 분류기 개발이 필요.

최근의 침입탐지 분류기에 대한 연구에서 SVM이 신경망, 유전자알고리즘, 클러스트링 기법에 비해 우수한 성능을 나타내었다는 연구 결과를 많이 볼 수 있다. 하지만 SVM을 학습시키는데 이용되는 기법인 QP(quadratic programming)는 계산과정에서 복잡한 계산이 요구되며 시스템 구현에도 어려움이 따른다. 또한 학습에 요구되는 메모리는 데이터 수의 제곱에 해당하며 학습 속도 또한 느린 단점이 있다. 더구나 SVM은 점증적인 방식이 아닌 일괄처리방식(batch way)에 의해 작동하므로 실시간 처리가 요구되는 침입탐지 시스템에는 적합한 분류기가 될 수 없다. 따라서 점증적으로 분류를 할 수 있는 분류기에 대한 연구가 필요하다.

2.3 개발 시스템

2.3.1 실시간 특징 추출

Hall에 의해 제안된 실시간 고유공간 갱신 방법을 간략히 설명하기 전에 먼저 수식에서 사용되는 벡터 및 행렬은 다음과 같이 정의한다. 벡터는 열벡터(column vector)를 의미하며 소문자로 나타내고 행렬은 대문자로 표시한다. 그리고 행렬의 크기는 아랫 첨자로 나타낸다. 예를 들어 A_{mn}

는 $m \times n$ 크기의 행렬을 의미한다. 행렬에서 열의 확장(column extension)은 대괄호(square brackets)로 나타낸다. 따라서 $[A_{mn} b]$ 는 $m \times (n+1)$ 크기의 행렬을 나타내며 벡터 b는 행렬 A의 마지막 열에 추가 된다. 동적인 고유공간 갱신 기법을 설명하기 위해 다음을 정의한다.

$U_{nk} = [u_j], j = 1 \cdots k$ 는 현재 까지 학습 자료 $x_i, i = 1 \cdots N$ 에서 구한 고유벡터 집합을 나타내며, $\lambda = diag(\Lambda_{nn})$ 는 고유치 행렬(eigenvalue matrix) Λ_{nn} 의 대각요소를 내림차순으로 정렬한 것을 나타내며 $\bar{x}$ 는 x 의 평균을 의미한다. 온라인 PCA 방법은 학습자료 x_{N+1} 이 추가 되었을 때 이전 학습 자료의 저장 없이 갱신하는 것이다. 먼저 새로운 학습 자료가 추가 되었을 때 갱신된 평균은 식 (1)과 같이 구한다.

$$\bar{x}' = \frac{1}{N+1}(N\bar{x} + x_{N+1}) \qquad (1)$$

식 (1)에 의해 새로운 평균이 구해지면 추가된 학습 자료에 의해 갱신된 고유 벡터 집합을 구할 수 있다.

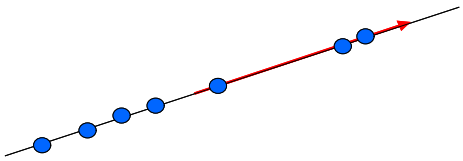

그림 1. 이차원 공간에서 하나의 고유벡터에 의해 모든 자료가 표현된 상태
Fig. 1. A set of points in two dimensional space, all data points can be described by a single base vector.

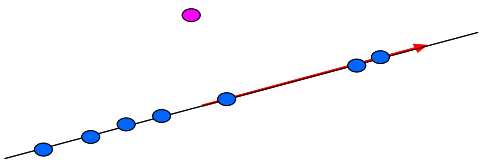

그림 2. 자료집합에 새로운 자료가 추가된 상태
Fig. 2. A new point is added into the data set.

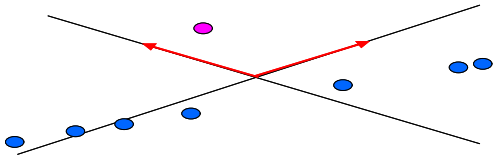

그림 3. 새로운 자료의 추가에 대해 고유벡터를 추가하여 모든 학습 자료를 표현
Fig. 3. We increased the dimensionality in order for all the points to be fully described by the two eigenvectors.

그림 1, 2, 3은 이러한 고유공간 갱신 기법을 개념적으로 설명하고 있다. 갱신된 고유 벡터 집합을 구하기 위해서는 이전의 고유벡터를 회전행렬(rotational matrix)에 적용하여야 하는데 이를 위해 먼저 직교잔차벡터(orthogonal residual vector)를 구해야 한다. 직교잔차벡터는 식 (2) 와 같이 계산된다.

$$\hat{h} = (U_{nk} a_{N+1} + \bar{x}) - x_{N+1} \qquad (2)$$

식 (2) 에서 구한 $\hat{h}$을 정규화한 것은 식 (3)과 같이 표시된다.

$$h_{N+1} = \frac{h_{N+1}}{|h_{N+1}|_2} \quad \text{for} \quad |h_{N+1}|_2 > 0 \quad \text{and}$$
$$h_{N+1} = 0 \quad \text{otherwise} \qquad (3)$$

여기서 $|h_{N+1}|_2$ 은 노름(norm)을 의미한다. $q = k+1$ 라 정의하면 새로운 고유벡터 U'_{nq} 은 식 (3)에 의해 구해진 h_{N+1}과 이전의 고유벡터 U_{nk}에 의해 생성된 행렬을 회전행렬 $R_{(k+1)(k+1)}$에 적용하여 구할 수 있으며 식 (4)와 같이 구한다.

$$U'_{nq} = \left[U_{nk} \quad h'_{N+1} \right] R_{(k+1)(k+1)} \qquad (4)$$

여기서 $R_{(k+1)(k+1)}$ 은 회전행렬 이며 식 (5)의 고유공간의 해이다.

$$D_{(k+1)(k+1)} R_{(k+1)(k+1)} = R_{(k+1)(k+1)} \Lambda'_{(k+1)(k+1)} \qquad (5)$$

여기서 $\Lambda'_{(k+1)(k+1)}$ 은 새로운 고유치들의 대각행렬이다. 행렬 $D_{(k+1)(k+1)}$는 다음과 같이 구성

할 수 있다.

$$D_{(k+1)(k+1)} = \frac{N}{N+1}\begin{bmatrix} \Lambda_{kk} & 0 \\ 0^T & 0 \end{bmatrix} + \frac{N}{(N+1)^2}\begin{bmatrix} aa^T\gamma a \\ \gamma a^T \gamma^2 \end{bmatrix} \quad (6)$$

여기서 $\gamma = h_{N+1}^T(x_{N+1} - \bar{x})$,$a = U^T(x_{N+1} - \bar{x})$ 와 같이 구하며 0는 k차원의 영벡터(zero vector)이다. 행렬 $D_{(k+1)(k+1)}$를 구성하는 몇 가지 방법이 제안 되었는데 Hall이 제안한 방법만이 평균을 갱신할 수 있도록 제공하는데 이 기법은 평균의 갱신을 허용하지 않는 기법에 비해 성능이 우수한 것으로 알려져 있다[3]. 하지만 Hall의 방법에서는 학습시 고유공간의 차원 유지에 관한 명확한 규칙이 제안되지 않았다. 새로운 학습 자료를 표현하기 위해 고유공간의 차원을 증가시키면 학습 자료는 잘 표현할 수 있으나 차원의 증가로 인해 더 많은 기억 공간을 필요로 한다. 반면에 고유공간의 차원을 일정 크기로 고정하면 이로 인해 학습 자료에 대해 일정량의 정보 손실을 감수해야 한다. 일반적으로 고유공간의 차원을 유지하기 위한 방법에는 다음과 같은 것이 있다.

(1) 잔차벡터가 일정 역치(threshold)를 초과하면 고유벡터를 추가.

(2) 최근에 구해진 고유치(eigenvalue)가 이전에 구한 고유치의 일정 비율을 초과할 경우에 고유벡터를 추가.

(3) 구해진 고유치의 값이 첫 번째 고유치의 일정 비율보다 작은 경우 고유벡터를 제거(차원을 줄임).

(4) 고유공간의 차원을 학습시 일정하게 고정.

본 논문에서는 (2) 방법을 채택하며 새로 구한 고유치의 값이 이전에 구한 고유치의 70% 이상을 초과하면 고유공간의 차원을 증가하는 방법을 사용한다.

실시간 고유공간 갱신 방식은 그 적용 범위가 학습 자료간의 선형적인 관계가 존재할 때 적용이 가능하다. 또한 고유공간의 차원을 유지하는 안정적인 방법이 아직까지 밝혀지지 않아 상용화된 실시간 탐지 시스템에 적용하기에는 실효성에 있어 문제가 될 수 있다. 대부분의 실세계 자료는 학습 자료 간에 비선형성이 존재하는 경우가 많으며, 고유공간의 안정적 차원유지를 위해서 안정적인 실시간 특징 추출에 대한 검토가 필요하다. 이러한 해결책의 하나로 신경망 기반의 특징 추출 기법이 대안이 될 수 있다. 이 방법을 간략히 기술하면 다음과 같다. Kung과 Diamantaras는 식 (7)과 같은 입출력 관계를 가지는 신경망기반의 주성분기법(APEX)을 개발하였다. APEX 모델에서 입력과 출력의 관계는 아래 식 (7)과 같다.

$$z(t) = W^T(t)x(t) \text{ and } y(t) = z(t) + H^T(t)y(t) \quad (7)$$

여기서 $x(t) \in R^p$ 는 입력 벡터이며 $y(t) \in R^m$ $(m \leq p$ $)$는 출력벡터를 나타낸다. 그리고 $W(t)$는 입력노드와 출력노드간의 직접 연결된 $p \times m$ 차원의 연결가중치 행렬을 나타내며 $H(t)$는 $m \times m$ 차원의 출력노드들 끼리 측면 연결된 연결가중치 행렬을 나타내며 상삼각행렬의 형태를 띤다. APEX 학습 모델의 구조는 그림 4에 나타나 있다. 행렬 W 와 H의 열벡터는 다음과 같이 나타낸다 $W = [w_1\ w_2 \ldots\ w_m]$, $H = [0\ h_2\ \ldots\ h_m]$. APEX 모델의 연결가중치 W에 대한 학습 규칙은 식 (8)과 같으며

$$W(t+1) = W(t) + \eta[X(t)\,\bar{Y}(t) - W(t)\,\bar{Y}^2(t)] \quad (8)$$

H 에 대한 학습규칙은 식 (9)와 같다.

$$H(t+1) = H(t) - \eta SUT[Y(t)\,\bar{Y}(t)] = -\eta H(t)\,\bar{Y}^2(t) \quad (9)$$

여기서 η 는 학습률을 나타내며 X 는 $p \times m$ 차원의 행렬, Y 와 $\bar{Y}$ 는 $m \times m$ 크기의 행렬이며 식 (10)과 같이 정의된다.

$$X = [x_1\ x_2\ \ldots\ x_m], \quad Y = [y_1\ y_2\ \ldots\ y_m], \quad \bar{Y} = diag(y_1.y_2,\ \ldots\ ,y_m) \quad (10)$$

연산자 SUT는 행렬 [ ]안의 요소 중 상삼각 부분을 되돌려준다. Kung과 Diamantaras 는 APEX 모델의 수렴성을 증명하여 제안된 모델의

안정성을 보였다[4].

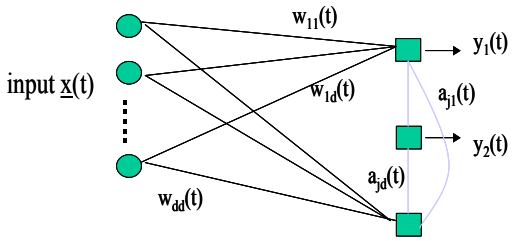

그림 4. APEX 학습모델의 구조
Fig. 4. APEX learning architecture

따라서 본 연구 개발에서는 비선형 자료에도 적용 가능하며 수렴성이 보장된 실시간 특징 추출 방법으로 APEX 알고리즘을 사용하고자 한다.

2.3.2 실시간 분류 시스템

SVM은 신경망의 단점인 과적합과 지역최소화에 빠지는 단점을 해결하기 위해 Vapnik과 공동연구자들에 의해 개발되었다. SVM에서 신경망에서 발생하는 과적합을 막아주는 것은 Vapnik-Chervonenkis(VC) 이론으로 설명 할 수 있으며 SVM의 학습은 polynomial 시간 내에 유일해가 발견될 수 있다는 것이 증명되어 안정적으로 사용할 수 있다. 신경망처럼 학습이 진동하지 않고 전역해를 보장하는 이유로 SVM은 많은 분야에서 분류기로 적용되고 있으나 몇 가지 단점이 존재한다. SVM을 학습 시키기 위해서는 QP(Quadratic programming) 과정이 필요한데 이에는 복잡한 계산이 요구되며 또한 QP 부분을 구현하기도 쉽지 않다. 더구나 학습시에 필요한 기억공간은 학습하고자 하는 자료 수의 제곱에 해당하며 이로 인해 학습 속도 또한 느리다는 문제점이 있다. 이러한 문제를 해결하기 위해 SVM의 분류기를 구하는데 있어 SVM처럼 QP 문제를 해결하는 것이 아닌 선형방정식을 푸는 형태인 LS-SVM모델을 제안하였는데[5] 이는 역행렬을 계산하는 계산의 단순함 뿐 만 아니라 대용량 학습 데이터에도 적용할 수 있는 장점이 있다.

LS-SVM 모델은 계산상 문제에 있어서 효율적이며 SVM 방법에 비해 확장이 쉬워 유리하지만 일괄처리 기법의 LS-SVM 모델은 N 이 학습하고자 하는 패턴의 개수라고 할 경우 (N+1) * (N+1) 크기의 행렬을 저장해야 하는 문제점이 있다. 이 것은 침입탐지와 같은 빅데이터를 다루는 경우에는 문제가 된다. 이러한 문제점을 해결하기 위해 Mu[6] 는 점증적 LS-SVM 모델을 제안하였다. 이는 새로운 역행렬을 계산하지 않아도 되는 계산상의 장점이 있으며 성능 면에서도 LS-SVM과 비슷한 유용한 기법이다. 따라서 본 연구에서는 Mu에 의해 제안된 수정된 LS-SVM 기법을 실시간 침입탐지 시스템의 분류기로 사용하고자 한다.

3. 실 험

제안된 방법의 성능을 비교하기 위해 침입탐지 시스템의 성능 평가를 사용하는 대표적인 자료인 KDD CUP 99 자료[7]를 바탕으로 성능을 평가한다. 실험 자료는 크게 DOS(denial of service), R2L(unauthorized access from a remote machine), U2R(unauthorized access to local superuser), Probing(port scanning) 등의 4가지 공격 유형으로 나뉜다. 실험은 전체 자료를 각각 학습 자료와 테스트 데이터로 70%, 30%의 비율로 나눈 다음 성능 평가를 실시하며 탐지 정확률은 아래와 같은 식 (11)에 의해서 평가한다.

$$\text{detection ratio} = \frac{correct \;\; \text{detection} \;\; data}{total \;\; data} \times 100 \tag{11}$$

3.1 특징 추출 실험

먼저 특징 추출 성능 결과를 보기 위해 모든 특징을 사용하는 경우와 APEX를 이용하여 특징을 추출한 경우의 탐지 성능을 비교하였다. 표 1과 표 2의 결과에서 알 수 있듯이 모든 특징을 사용한 경우와 특징 추출한 경우의 탐지율은 큰 차이를 보이지 않는다. 이 결과를 통해서 알 수 있는 것은 일반적으로 실시간으로 특징을 추출하는 방법의 경우 일괄 처리 방법에 비해 특징 추

출 성능이 떨어지는 것으로 알려져 있다. 그러나 APEX를 사용한 특징 추출 결과는 모든 특징을 사용한 경우와 비교해서 탐지율이 비슷하므로 APEX의 특징 추출 성능이 우수하다는 것을 나타낸다. 또한 추가 메모리의 요구 없이 실시간으로 특징 추출이 가능하므로 본 연구와 같이 상용화제품 개발을 염두에 둔 경우에는 매우 의미 있는 실험 결과라 할 수 있다.

표 1. 모든 특징을 사용한 실험 결과
Table 1. Detection ratio using all features

클래스	탐지율	훈련시간 (Sec)	테스팅시간 (Sec)
Normal	98.55	5.83	1.45
Probe	98.59	28.0	1.96
DOS	98.10	16.62	1.74
U2R	98.64	2.7	1.34
R2L	98.69	7.8	1.27

표 2. APEX를 사용하여 특징 추출한 실험 결과
Table 2. Detection ratio using APEX

클래스	탐지율	훈련시간 (Sec)	테스팅시간 (Sec)
Normal	98.43	5.25	1.42
Probe	98.63	25.52	1.55
DOS	98.14	15.92	1.48
U2R	98.64	2.17	1.32
R2L	98.70	7.2	1.08

3.1 분류 성능 실험

제안한 방법의 분류 성능을 확인하기 위해 분류 문제에서 많이 사용되는 SVM 과의 비교 실험을 하였다.

표 3. APEX를 사용하여 특징 추출한 경우의 SVM과의 분류 성공률 성능비교
Table 3. Classification ratio of proposed systems.

	Normal	Probe	DOS	U2R	R2L
제안한 방법	98.67	98.72	98.56	98.88	98.78
SVM	98.59	98.38	98.22	98.87	98.78

실험 결과 제안한 방법과 SVM을 사용한 경우의 분류 결과는 근사 하거나 일부 클래스에서는 제안한 시스템이 좀 더 우수한 분류 성능을 나타내었다.하지만 SVM을 학습 시키기 위해서는 QP(Quadratic programming) 과정이 필요하며 이에는 복잡한 계산이 요구되며 또한 QP 부분을 구현하기도 쉽지 않다. 더구나 학습시에 필요한 기억공간은 학습하고자 하는 자료 수의 제곱에 해당하며 이로 인해 학습 속도 또한 느리다는 문제점이 있다. 이에 반해 제안한 방법은 실시간으로 동작하기에 SVM에 비해 메모리 효율성이 좋고 학습 자료의 추가를 허용하는 유연한 방식이면서도 비슷한 성능을 나타내고 있어 있다. 이러한 실험 결과로 제안한 실시간 침입탐지 시스템은 현실 세계에서도 적용 가능하다는 실험 결과를 얻었다.

4. 결론

본 논문에서는 비선형 자료에도 적용 가능하며 수렴성이 보장된 실시간 특징 추출 방법으로 APEX 알고리즘과 점증적 LS-SVM 분류기를 결합한 실시간 침입탐지 시스템을 개발하였다. 개발한 시스템의 상용화를 위해 공동 개발 회사와 함께 실제 네트워크상에서의 침입탐지 자료를 수집한 후 이를 바탕으로 성능 평가를 진행하여 제안한 방법의 타당성 검토와 이를 바탕으로 상용화 단계로 진행할 예정이다.

REFERENCES

[1] A.H. Sung, S. Mukkamala, "Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks" *Proceedings of the 2003 Symposium on Applications and the Internet,* 2003.

[2] Nong Ye, "A Scalable Clustering Technique for Intrusion Signature Recognition," *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security,* 2001.

[3] P. Hall, D. Marshall, R. Martin., "Incremental eigenalysis for classification", *In British Machine Vision Conference*, Vol. 1, pp. 286-295, 1998.

[4] H. Chen, R.-W. Liu, "Adaptive distributed orthogonalization processing for principal components analysis", *Acoustics Speech and Signal Processing*, Vol. 2, pp. 293-296, 1992.

[5] J. A. K Suykens, J. Vandewalle : "Multiclass Least Squares Support Vector Machines", *Proc. International Joint Conference on Neural Networks, Washington DC,* 1999.

[6] MU Xin-guo, Hao Wen-ning, Zaho En-Lai, Chen Gang "An incremental LS-SVM learning algorithm ILS-SVM" *International Conference on E-Business and E-Government ICEE*, 2011.

[7] https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[8] Yasir Hamid, M Sugumaran, Ludovic Journaux " Machine Learning Techniques for Intrusion Detection : A Comparative Analysis", *International Conference on Informatics and Analytics* ICIA 2016, 2016.

[9] Mahdi Zamani "Machine Learning Techniques for Intrusion Detection" arXiv, 2013.

저자약력

김 병 주(Byung-Joo KIm) **[정회원]**

- 1992년 2월 : 부산대학교 전자계산학과 대학원 (이학석사)
- 2004년 2월 : 경북대학교 컴퓨터과학(이학박사)
- 2003년 3월 ~ 현재 : 영산대학교 컴퓨터공학과 정교수

<관심분야> 기계학습, 데이터마이닝