LBG-SVDD을 이용한 침입탐지 기법

유승도, 박귀태
고려대학교 전자전기공학과

Intrusion Detection System Using LBG-SVDD

Seong-Do Yoo, Gwi-Tae Park
Dept. of Electrical Engineering, Korea University

Abstract - 최근 유비쿼터스 네트워크에 대한 관심이 높아지고 있다. 하지만 유비쿼터스 네트워크는 무선으로 데이터를 전송함으로서 특성상 쉽게 침입자들로부터 침입을 당할 수 있는 보안 문제가 중요하게 대두되고 있다. 이에 따라 강력한 침입탐지 기술에 대한 요구가 증가되고 있다. 본 논문에서는 갈수록 늘어나는 새로운 변형 공격에 대한 탐지를 위하여 LBG-SVDD을 이용한 침입탐지 기법을 제안한다. LBG-SVDD은 새로운 변형 공격 침입 탐지가 발견되었을 때, 새로운 변형 공격 형태에 대한 빠른 학습 훈련을 통해 공격 침입 탐지를 할 수 있다.

1. 서 론

최근 유비쿼터스가 대두 되면서 무선 네트워크 환경이 급속도로 확산되고 있다. 유비쿼터스 네트워크는 무선으로 단말기들이 통신하며, 이동, 설치, 유지 보수의 확장성과 네트워크 구축이 용이 등 유선 네트워크 보다 많은 장점을 가지고 있다. 이러한 유비쿼터스 네트워크는 많은 수의 노드들을 가지고 있으며, 많은 라우팅 프로토콜 방식으로 데이터를 전송, 수집한다. 하지만 무선 통신으로 데이터를 주고받기 때문에 무방비 상태로 노출되어 보안에 취약하다. 무선은 유선과 달리 언제, 어디서, 누구든지 쉽게 접근하여 침입할 수 있게 된다.

비정상적인 침입 탐지 기법 중 신경망(Neural Network)을 이용한 기법이 있지만 신경망은 학습을 위해 참고로 하는 명령어 개수가 적으면 긍정적 결함이 발생할 확률이 증가하고, 명령어 개수가 많으면 부정적 결함이 발생할 확률이 증가하며, 또한 많은 연산량을 요구하는 단점이 있다[1]. 이에 반해 SVDD(Support Vector Data Description)은 명료한 이론적 근거에 기반하고 있다. 간단하고 명료한 알고리즘을 통하여, 학습을 성공적으로 수행하는데 영향을 미치는 요소들을 규명할 수 있다. 또한 실제 응용문제에서 높은 인식 성능을 나타낸다. 신경망의 한계점으로 지적되었던 과대적합, 국소최적화와 같은 한계점들을 완화하는 장점을 가진다.

본 논문에서는 단일 클래스 분류 알고리즘 중 대표적인 알고리즘인 SVDD을 적용하고 빠른 학습을 위해 LBG 알고리즘을 추가 시켜 새로운 변형 공격에 대한 효율적인 침입 탐지 모델을 제안한다.

2. 본 론

2.1 기존 연구 사례

Anomaly Detection System에서 주제에 따른 공격, 특정 공격(Dos, FFRR, U2Su, R2L 등), 특정 공격이 아닌 분류를 SVM을 사용한 기존 논문들이 있다[2-3]. 이러한 논문들은 학습시간은 고려하지 않고 공격에 대한 분류에 대해서 언급하였다. 이러한 논문들의 문제점은 새로운 변형 공격에 대해서 다시 학습을 하여야 하는데 많은 양의 학습 데이터를 필요로 하기 때문에 학습 시간이 오래 걸린다. 이러한 학습 시간을 줄이기 위해서 우선 LBG 알고리즘을 이용하여 군집화를 한 다음, SVDD을 이용한다. 그리하여 새로운 공격이 발생하였을 경우, 비교적 빠른 학습으로 인한 침입 탐지 기법을 제안한다.

2.2 LBG 알고리즘

비균일 이진분리와 k-means를 결합된 알고리즘이다. k-means 알고리즘은 좋은 코드북을 얻을 수 있지만 속도가 느린 단점이 있다. 이진 분할 알고리즘은 코드북은 k-means에 비교하여 떨어지지만 분할 속도가 빠르다. 그러므로 훨씬 빠른 비균일 이진 분할법을 접목시킨다. k-means를 통하여 클러스터들을 찾고 어떠한 두 클러스터의 중심을 연결하여 이 직선을 수직 이등분하는 선분이 최적 경계이다. 그리고 각 클러스터의 중심에서 가장 가까운 모든 점들을 해당 클러스터에 포함시키면 된다. k-means는 초기 중심들의 선택에 민감한 특징을 가진다. 그러므로 초기값을 이진 분리로 구한 중심을 사용하면 표준 k-means 방법을 사용하는 경우보다 더 나아진다. 비균일 이진분리와 k-means를 결합한 알고리즘을 LBG 알고리즘이라 하고, 이 방법을 이용하면 표준 k-means보다는 더 빠르고 더 높은 질의 코드북을 만들어 낼 수 있다[4].

2.2.1 k-means 알고리즘

k-means 알고리즘은 임의의 초기값에서부터 추정(E)-최대화(M) 과정을 수렴할 때까지 반복시키면서 중심을 찾는다. 즉, E단계에서 중심으로부터 클러스터를 선택하여 결정하고, M단계에서 거꾸로 클러스터로부터 중심을 결정하는 과정을 반복하는 추정 알고리즘이다. k-means의 계산 절차는 다음과 같다[4].

우선, 데이터 집합$[x_1,\cdots x_N]$으로 임의의 k개의 벡터를 선택하여 k개의 초기 중심 집합 $[y_1,\cdots y_K]$을 만든다. E단계로 만약 데이터 x_N이 y_i에 가장 가깝다면 클러스터 X_i에 속하도록 라벨링한다. 결국 데이터 집합을 K개의 클러스터들 $\{X_1,\cdots X_K\}$로 나누어진다[4].

$$X_i=\{x_n|d(x_n,y_i)\le d(x_n,y_j)\},\quad j=1,\cdots,K \tag{1}$$

다음은 M단계로 E단계에서 구한 새로운 클러스터들에서 각각의 중심을 갱신한다[4].

$$y_i=c(X_i), i=1,\cdots,k \tag{2}$$

그리고 데이터와 가장 가까운 클러스터 중심들과 거리의 합으로 총 왜곡을 구한다[4].

$$D=\sum_{n=1}^{N} d(x_n,y_{i(n)})\ \text{ 여기서, } i(n)=k,\ \text{if } x_n\in X_k \tag{3}$$

마지막으로 왜곡이 적절하게 변하지 않거나 설정된 반복 횟수에 도달할 때까지 E단계부터 총 왜곡 계산 과정까지 반복한다.

이 때, 왜곡이 안정되었는지 아닌지를 확인하는 효과적인 검정은 다음 수식과 같다[4].

$$\triangle D=\frac{D_{prev}-D_{curr}}{D_{prev}}<10^{-4} \tag{4}$$

2.2.2 비균일 이진 분할

이진 분할 알고리즘은 초기에 데이터 집합을 두 개의 클러스터로 나눈 다음 K개의 클러스터가 남을 때까지 반으로 클러스터를 나누는 과정을 반복한다. 즉, logK번의 이진 검색을 통하여 가장 가까운 중심을 찾게 되는 것이다. 수식은 다음과 같다.

한 개의 클러스터 X_1과 관련된 중심이 $y_1=c(X_1)$인 모든 데이터 점들 x_n으로 시작한다. 클러스터 카운터는 k=1로 둔다. K개의 중심들이 얻기 위해서 다음 과정을 (k-1)번 반복한다. 클러스터 내의 점들과 중심의 평균거리로 측정된 가장 큰 왜곡을 가진 클러스터 X_i를 선택한다[4].

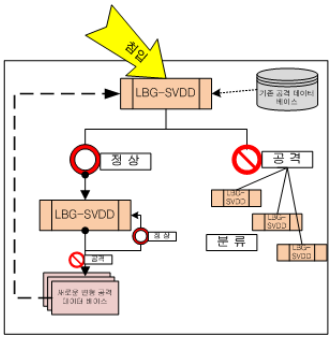
$$D=\frac{1}{N_j}\sum_{n=1}^{N_i} d(x_n^{(j)},y_i),\qquad X_j=x_n^{(j)}|n=1,\cdots,N_j \tag{5}$$

$$D_i\ge D_j,\ j=1,\cdots,k$$

선택된 클러스터 X_i를 다음 방법들 중에 하나를 선택하여 두 개의 부클러스터 X_a와 X_b로 나눈다. 그런 다음 K=2로 집합 X_i상에서 k-means를 수행하고 집합 X_i의 주 고유벡터 v_i를 결정하고 부클러스터 X_i에 있는 점들이 y_i+v_i에 가장 가까운 점들을 X_a로 y_i-v_i에 가장 가까운 점들을 X_b로 한다. 그 후 중심 y_i를 대치하고 새로운 중심을 다음과 같이 둔다.

$$y_i=c(X_a)y_{k+1}=c(X_b) \tag{6}$$

클러스터 카운트를 증가 시킨다[4].



<그림 1> LBG-SVDD를 이용한 침입 탐지 시스템

2.3 SVDD(Support Vector Data Description)

SVM은 관측되지 않은 영역을 포함하여 결정 경계면을 생성함으로 새로운 학습 데이터에 대해서 오분류 할 가능성이 크다 그러므로 해당 클래스만을 독립적으로 표현하는 단일 클래스 분류기로서 대표적인 알고리즘인 SVDD를 기반으로 한다.

d-차원 입력 공간 상에서 존재하는 N-개의 데이터의 집합 $D=\{x_i|i=1,\cdots,N\}$이 주어졌을 경우, 각 클래스의 학습 데이터를 포함하면서 중심이 a, 반경이 R인 구체의 체적을 최소화 하는 문제로 정의 되며, 최적화 문제를 통하여 수식화 된다[5].

$$\min L_0(R^2,a,\xi)=R^2+C\sum_{i=1}^{N}\xi_i \tag{7}$$

$$s.t.\ \ \|x_i-a\|^2\le R^2+\xi_i,\quad \xi_i\ge 0,\ \forall i$$

학습 데이터 x_i와 중심 a 사이의 거리가 R을 초과하는 경우 적절한 벌점을 부과하는 전략으로 ξ_i는 i번째 학습 데이터 x_i가 원형체에서 벗어나는 벌점이다. C는 반지름과 벌점항의 상대적 중요성을 조정하는 상수이다. 쌍대문제를 구하기 위해서 라그랑제 함수 L을 도입하고 다음과 같은 QP문제로 정리된다[5].

$$\min_a \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j<x_i,x_j>-\sum_{i=1}^{N}\alpha_i<x_i,x_i> \tag{8}$$

$$s.t.\ \sum_{i=1}^{N}\alpha_i=1,\ \alpha_i\in[0,C],\ \forall i$$

학습 후, 결정함수는 다음 수식으로 정의 된다[5].

$$\begin{aligned} f(x)&=R^2-\|x-a\|^2 \\ &=R^2-\left(\langle x,x\rangle-2\sum_{i=1}^{N}\alpha_i\langle x_i,x\rangle+\sum_{i=1}^{N}\sum_{j-1}^{N}\alpha_i\alpha_j\langle x_i,x_j\rangle\right) \\ &\ge 0 \end{aligned} \tag{9}$$

특히, 커널함수 k를 통하여 정의되는 고차원의 특징 공간 F위에서 정의 되는 원형체를 사용하는 방향으로 확장될 수 있다. 가우시안 커널을 사용할 경우, 다음 수식과 같이 단순화 될 수 있다[5].

$$\min_a \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j k(x_i,x_j) \tag{10}$$

$$s.t.\ \sum_{i=1}^{N}\alpha_i=1,\ \alpha_i\in[0,C],\ \forall i$$

$$\begin{aligned} f(x)&=R^2-\|x-a\|^2 \\ &=R^2-\left(1-2\sum_{i=1}^{N}\alpha_i k(x_i,x)+\sum_{i=1}^{N}\sum_{j-1}^{N}\alpha_i\alpha_j k(x_i,x_j)\right) \\ &\ge 0 \end{aligned} \tag{11}$$

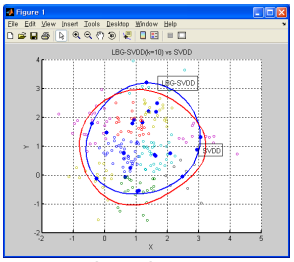
2.4 실험 결과

본 논문에서는 LBG 알고리즘은 패턴인식 개론[4]에 있는 MATLAB 실습 예제의 바탕과, Tax's data description toolbox[6]를 이용하여 SVDD를 통하여 성능을 비교 평가하였다.

2.4.1 시나리오

본 논문의 주요 목적은 새로운 변형 공격 형태가 발견 되었을 경우, 재빠른 학습으로 새로운 공격 형태에 대해서 감지하여 예방하는 것이다. <그림 1>에서 나타낸 바와 같이 공격적인 침입이면 대상에 대한 분류를, 정상이면 다시 LBG-SVDD 알고리즘을 통해 정상적인 침입인지 새로운 침입인지를 탐지한다. 새로운 유형의 침입이 감지가 되면 새로운 침입의 유형에 따른 데이터 베이스를 구축하여 빠른 학습을 통해 새로운 침입 유형 공격에 대비를 한다.

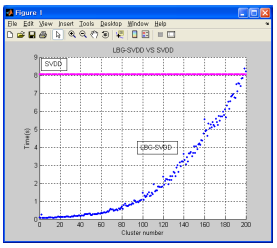
LBG-SVDD 알고리즘은 우선 데이터 들을 k개의 중심으로 클러스터를 구성하고 각 클러스터들을 학습 데이터로 가정하고 SVDD로 학습을 한다.



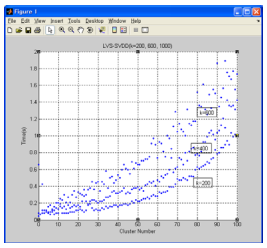
<그림 2> LBG-SVDD(k=10)와 SVDD 성능에 대한 비교

2.4.21 실험 및 결과

<그림 3>에서 a의 결과는 학습 데이터를 200개의 랜덤 생성하여 클러스터 수의 k를 0~200까지 변화를 주어 LBG-SVDD와 SVDD의 학습 시간에 대해서 비교하였고 <그림 3>에서 b의 결과는 학습 데이터를 200, 600, 1000개로 생성하고 클러스터 수 k를 0~200까지 변화시켜 각각에 대해서 학습 시간에 대해서 비교했다.



a. LBG-SVDD와 SVDD 학습 시간에 대한 비교



b. LBG-SVDD에서 k=200,600,1000에 대한 학습 시간에 대한 비교

<그림 3> 학습 시간에 대한 비교

3. 결 론

본 논문에서는 침입 탐지 시스템을 제안하고 갈수록 늘어나는 새로운 변형 공격이 일어났을 경우 빠른 감지를 위하여 LBG-SVDD를 이용하여 학습 시간을 줄였다, 성능 면에서는 비슷하게 나왔다. 따라서 새로운 변형 공격에 대한 침입에 대하여 효율적인 대비로 인하여 손해를 줄일 수 있다. 본 연구의 향후 과제는 시간뿐만 아니라 학습 에러와 오분류를 줄이는 침입탐지 시스템에 연구가 향후 연구과제로 요구된다.

감사의 글

본 논문은 건설교통부가 출연하고 한국 건설교통기술평가원에서 위탁 시행한 첨단융합기술개발사업 [과제번호:06 첨단융합 D01]의 지원으로 이루어졌습니다.

[참 고 문 헌]

[1] 이종성, 채수환, 박종서, 지승도, 이종근, 이장세 , "침입탐지 기술 동향", 한국통신학회지 (정보통신) 제16권 11호, pp. 46~63 , 1999. 11

[2] S. Mukkamala, A. H. Sung, "Detecting Denial of Service Attacks Using Support Vector Machines", IEEE International Conference on Fuzzy Systems, IEEE Computer Society Press, pp. 1231-1236, 2003

[3] H. Deng, Q.-A. Zeng, and D. P. Agrawal, "SVM-based Intrusion Detection System for Wireless Ad Hoc Networks", IEEE Vehicular Technology Conference (VTC'03), Orlando, October 6-9, 2003.

[4] 한학용, 한빛미디어, "패턴인식 개론", 한빛 교재 시리즈, 2006

[5] 박주영, 임채환, "비정상 상태 탐지 문제를 위한 서포트벡터 학습", 퍼지 및 지능시스템학회 논문지, vol. 13, no. 3, pp. 266-274, 2003.

[6] D.M.J.Tax, "dtools, the data description toolbox for matlab,version 1.5.4,"Sept 2006. http://ict.ewi.tudelft.nl/~avidt/dd tools.html.