네트워크 침입탐지를 위한 밀도함수 기반 아웃라이어 탐지 기법

박종명° 김한준
서울시립대학교 전자전기컴퓨터공학부
{dbmaster,khj}@uos.ac.kr

Density Function-based Outlier Detection Algorithm for Detecting Network Intrusion

Jongmyoung Park° Han-joon Kim
Department of Electrical and Computer Engineering, The University of Seoul

요 약

네트워크 기반 오용 탐지 시스템은 이미 알려진 공격기법만 탐지할 수 있기 때문에 새로운 공격에 대한 탐지를 하기 위해서는 수시로 새로운 침입패턴을 추가시켜야 하는 어려움이 있다. 본 논문에서는 이런 어려움을 해결하기 위해 네트워크 데이터를 분석하여 새로운 침입패턴을 생성해 내는 자동화 시스템과 제안된 시스템의 성능을 결정짓는 밀도 함수 기반의 아웃라이어 탐지 알고리즘을 제안한다. 알고리즘의 성능 평가는 정확도, 재현율을 결합한 조화평균의 측정값을 이용하여 사용하여 평가했으며 기존 알고리즘보다 성능이 향상되었음을 보인다.

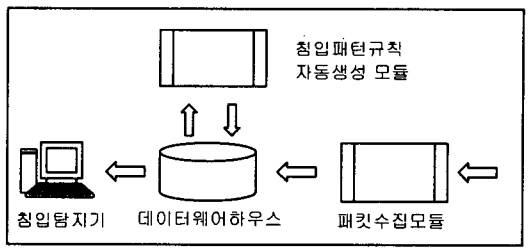
1. 서 론

네트워크 침입탐지 시스템(Network Intrusion Detection System)은 네트워크에서 발생하는 이벤트들을 모니터링하고, 침입발생 여부를 탐지하고, 대응하는 시스템을 말한다. 침입탐지의 분류는 이미 알려진 침입 행위에 대한 정보를 이용하여 공격을 탐지하는 오용탐지(Misuse Detection)와 사용자의 정상 행위를 기반으로 정상적인 행동 패턴에 어긋나는 경우를 침입으로 탐지하는 비정상행위탐지(Anomaly Detection)로 나뉜다.[1]

현재 침입탐지시스템의 가장 일반적인 형태는 네트워크 기반 오용탐지시스템이다. 이는 알려져 있는 침입패턴을 탐지하는데 적합하며 비교적 구현비용이 저렴하고 정확한 탐지결과를 가진다.[2] 네트워크 기반 오용탐지시스템은 알려진 공격기법에 대한 탐지 능력만을 가지고 있기 때문에 새로운 공격에 대한 침입탐지가 불가능하며 침입 패턴을 추가시켜야 하는 어려움이 있다. 물론 기계학습(Machine Learning)을 통하여 자동으로 침입패턴을 생성할 수 있지만[3], 기계학습을 위해 필요한 학습 데이터를 만드는 과정에서 전문가의 노력이 많이 요구된다.

본 논문에서는 이러한 문제점들을 해결하기 위하여 네트워크 패킷 데이터를 분석하여 새로운 침입 패턴을 자동으로 생성해 내는 시스템을 제안하며, 그 시스템의 성능을 결정짓는 새로운 아웃라이어(outlier) 탐지 기법에 대하여 논한다.기존의 아웃라이어 탐지는 유클리디언 거리함수에 의존하고 있어 정확도가 높지 못하다. 기존 유클리디언 거리함수기반 아웃라이어 탐지를 개선하기 위해 본 논문에서는 밀도함수를 이용하여 보다 정확한 아웃라이어 군집을 찾아내는 알고리즘을 제안한다.

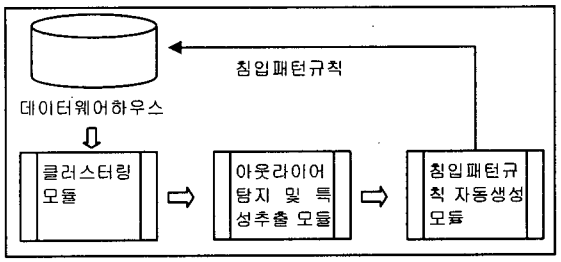
2. 제안 시스템



(그림 1) 네트워크 침입탐지 시스템

제안하는 네트워크 침입 탐지 시스템은 그림 1과 같이 패킷 수집모듈, 침입탐지기, 패턴규칙 자동 생성기, 데이터웨어하우스로 구성된다.

첫째, 패킷 수집 모듈은 네트워크망에 접속하여 실시간으로 네트워크 패킷을 스캐닝 하면서 시스템 내부에서 미리 정의한 포맷으로 변환하여, 그 데이터를 데이터웨어하우스에 적재한다. 둘째, 침입탐지기는 실제 네트워크 패킷 데이터를 이미 생성된 침입 패턴과 실시간 대조작업을 통해 불순한 것으로 판단되는 침입 패킷을 구분하여 판단 정보를 관리자에 자동으로 경고하며, 자세한 네트워크 데이터의 분석을위해 OLAP(실시간 자료분석) 기능을 제공 한다. 셋째, 패턴규칙 자동 생성기는 그림 2에서 보는바와 같이, 클러스터링 모듈과 아웃라이어 탐지 및 특성 추출 모듈 그리고 침입패턴규칙 자동 생성 모듈로 구성된다. 클러스터링 모듈은 클러스터링 알고리즘[4]에 입각하여 유사 인스턴스를 가지는 클러스터(집합)을 생성한다. 아웃라이어 탐



(그림 2) 침입패턴 규칙 자동 생성 시스템

지 및 특성 추출 모듈은 주변으로부터의 영향력이 미미한 아웃라이어 클러스터를 결정하고, 그 아웃라이어 클러스터내에 포함된 인스턴스들로부터 다른 클러스터와 구별짓는 주요 특성을 추출한다. 침입패턴규칙 자동생성 모듈은 아웃라이어 클러스터로부터 찾아낸 침입관련 특성을 사용하여 침입감시모듈이 사용할 수 있는 정형화된 형식의 침입패턴규칙을 생성한다. 넷째, 데이터웨어하우스는 여러 개의 패킷수집기로부터 전달된 데이터와 네트워크 침입패턴 데이터베이스를 보관하기 위한 통합 데이터 저장소이다.

본 제안 시스템에서 가장 중요한 부분은 침입패턴 규칙 자동생성 모듈 이며, 특히 아웃라이어 탐지 및 특성 추출 모듈에 초점을 맞춘다. 다음 절에서 기존의 유클리디언 거리함수에 의존한 아웃라이어 탐지 기법을 설명하고 새로운 유형의 밀도함수 기반 아웃 라이어 탐지 알고리즘을 소개한다.

3. 아웃라이어 탐지 알고리즘

3.1 유클리디언 거리함수를 이용한 아웃라이어 탐지 알고리즘

아웃라이어(outlier) 탐지 알고리즘의 기본 아이디어는 그 크기가 가장 큰 클러스터가 정상적인 네트워크 패킷일 것이라는 예측하에, 그 최대 클러스터와 가장 거리가 먼 클러스터를 찾아내어 이를 비정상적인 침입행위에 해당하는 패킷으로 간주하는 것이다. 이때 기존기법은 거리함수로서 유클리디언 거리함수(식 1 참조)를 사용한다. 즉 가장 많은 인스턴스를 포함하고 있는 클러스터를 선택한 후 클러스터의 중심점으로부터 유클리디언 거리함수 값이 가장 큰 클러스터를 아웃라이어로 결정 한다. [5]

유클리디언 거리함수는 두 개의 인스턴스가 벡터 $\overrightarrow{d_x}$, $\overrightarrow{d_y}$ 로 주어질 때 식1과 같이 정의한다.

$$\left|\overrightarrow{d_x}-\overrightarrow{d_y}\right| = \sqrt{(\overrightarrow{d_x}-\overrightarrow{d_y})^T \cdot (\overrightarrow{d_x}-\overrightarrow{d_y})} \quad \text{(식 1)}$$

여기서, T는 전치행렬을 의미한다.

기존 알고리즘은 클러스터의 중점간의 거리가 멀다는 이유만으로 비정상적인 침입행위 패킷이라고 간주하기 때문에 탐지율이 높지 못하다. 다시 말해서, 클러스터 크기가 크더라도 크기가 최대인 클러스터의 중심과 멀리 떨어져 있다면 아웃라이어로 결정될 수 있다는 것이다.

3.2 밀도함수 기반의 아웃라이어 탐지 알고리즘

유클리디언 거리함수 기반 아웃라이어 탐지 기법의 한계를 극복하기 위해 본 논문에서는 클러스터 밀도함수(density function)를 사용하여 아웃라이어를 탐지하는 기법을 제안한다.

클러스터 밀도함수는 인스턴스간 영향함수(influence function)를 사용하여 정의한다. 영향함수의 정의는 다음과 같다.

인스턴스 d_x, d_y가 주어질 때 d_x에 대한 d_y의 영향 함수 $\Omega^{d_y}(d_x)$는 식 2와 같이 정의 한다. 이는 d_y가 d_x에 미치는 영향력을 정규분포 함수식을 빌어 표현한 것인데, 둘 사이의 유클리디언 거리가 가까울수록 영향력이 급격하게 증가한다.

$$\Omega^{d_y}(d_x) = e^{-\frac{|d_x-d_y|^2}{2\sigma^2}} \quad \text{(식 2)}$$

식 2에서 δ 는 영향함수의 형태를 결정하는 제어인자(control parameter)이고, $\left|\overrightarrow{d_x}-\overrightarrow{d_y}\right|$는 인스턴스 d_x와 d_y간의 유클리디언 거리 값이다.

클러스터링 결과에서, 특정 클러스터 내부의 각 인스턴스(개체)에 대해 주변 클러스터의 영향력(Influence)를 계산 했을 때, 그 클러스터가 아웃라이어인 경우에 그 영향력(나중에 이 값을 클러스터 밀도함수로 정의한다.)이 크지 않을 것이다. 정상행위로 인해 발생한 네트워크 데이터는 네트워크 침입(비정상행위)으로 발생한 데이터와 그 성격이 매우 다를 것이다. 그러므로 비정상행위 관련 데이터를 함유하고 있는 클러스터는 주변으로부터 영향력을 덜 받아 클러스터 밀도함수 값이 적을 것이다.

또한 인스턴스 집합 $D=\{d_i,d_j,\cdots,d_k,\cdots\}$이 주어질 때, 집합 D에 존재하는 모든 인스턴스가 인스턴스 d_x에 미치는 영향력의 합에 해당하는 함수 $\Omega^D(d_x)$는 식 3과 같다. 이 함수는 인스턴스 밀도함수라고 칭한다.

$$\Omega^D(d_x) = \sum_{d_y \in D} \Omega^{d_y}(d_x) = \sum_{d_y \in D} e^{-\frac{|d_x-d_y|^2}{2\sigma^2}} \quad \text{(식 3)}$$

클러스터링 후에 특정 클러스터 C_i에 포함된 모든 인스턴스에 대해서 밀도함수는 식 4와 같이 계산한다. 이 밀도함수를 클러스터 밀도함수라 칭한다.

$$\Omega^D(C_i) = \sum_{d_y \in C_i} \Omega^{d_y}(d_x) \quad \text{(식 4)}$$

위 식은 C_i에 포함된 각 인스턴스 d_x에 대한 영향 함수값을 모두 더한 것이다. 여기서 D는 C_i를 제외한 다른 클러스터에 존재하는 인스턴스들의 집합이다.

밀도함수 기반의 아웃라이어 탐지 알고리즘은 단순한 유클리디언 거리에 의한 탐지가 아니라 각 클러스터 간의 영향력을 기반으로 하기 때문에 침입 패턴을 포함할 것으로 판단되는 아웃라이어 클러스터를 보다 정확하게 판별할 수 있다.

4. 성능 평가

4.1 실험 데이터 및 성능 평가 방법

실험 데이터를 위해 1998년 DARPA Intrusion Detection Evaluation Program[6]에서 사용된 것을 가공하여 두 가지 유형인 A형과 B형을 생성 하였다. A형은 공격 패킷이 적게(1%)

포함되 있는 데이터이며, B형은 공격 패킷이 많이(23%) 포함되어 있는 데이터이다. 이 데이터에는 침입의 기본적 분류인 DOS, R2L, U2R, Probing등의 공격 유형 24가지와 정상행위 패킷이 포함 되어 있다.

아웃라이어 탐지 성능을 평가하기 위한 기준은 정확도(Precision)(식 5)와 재현율(recall)(식 6)을 결합한 조화 평균인 F-measure(식 7)가 사용된다.

$$P = \frac{\text{아웃라이어에포함된침입패킷수}}{\text{아웃라이어에포함된전체패킷수}} \quad \text{(식 5)}$$

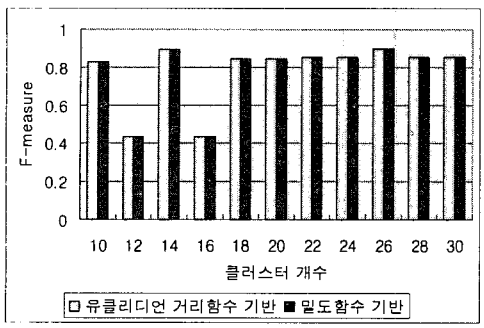
$$R = \frac{\text{아웃라이어에포함된침입패킷수}}{\text{전체데이터에포함된침입패킷수}} \quad \text{(식 6)}$$

$$\text{F-measure} = \frac{2 \cdot P \cdot R}{P + R} \quad \text{(식 7)}$$

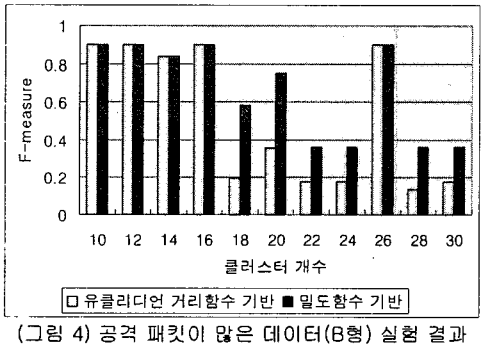
4.2 실험 결과 및 분석

제안 알고리즘의 성능 평가를 위하여 네트워크 패킷 데이터를 K-means 클러스터링 기법[8]을 이용하여 클러스터링 하였다. 클러스터링 후 유클리디언 거리함수를 이용한 아웃라이어 탐지 알고리즘과 제안한 밀도함수 기반의 아웃라이어 탐지 알고리즘을 이용하여 아웃라이어를 탐지하였다. A형 데이터와 B형 데이터의 실험 결과는 각각 그림 3과 그림 4와 같다. 그래프의 X축은 클러스터 개수이며 Y축은 F-measure를 나타낸다.

일반적으로 클러스터링 알고리즘을 사용할 때 최적의 클 러스터 개수를 결정하는 것은 어려운 문제이므로, 본 실험에서는 10~30개 범위에서 클러스터의 개수를 변경하면서 실험하였다. A형 데이터를 사용한 실험에서는 그림 3에서 보는 바와 같이 유클리디언 거리함수를 이용한 아웃라이어 탐지 알고리즘과 제안한 밀도함수 기반의 아웃라이어 탐지 알고리즘이 동일한 성능을 보였다.



(그림 3) 공격 패킷이 적은 데이터(A형) 실험 결과



(그림 4) 공격 패킷이 많은 데이터(B형) 실험 결과

하지만 B형 데이터의 실험에서는 그림 4와 같이 클러스터의 개수가 많아질수록 밀도 함수 기반의 아웃라이어 탐지 알고리즘이 좋은 성능을 보이고 있다. 그림 4에서 보이는 바와 같이, 클러스터 개수 18의 경우 최대 296%의 개선 효과를 보이고 있다. 밀도함수기반 아웃라이어 탐지 알고리즘은 유클리디언 거리함수 기반 아웃라이어 탐지 알고리즘과 달리 영향력이 적은 클러스터를 탐지 하기 때문에 클러스터의 개수에 상관없이 높은 정확도를 보인다.

5. 결론 및 향후 과제

최근 들어 네트워크 상의 불법적인 행위를 탐지하기 위해서 다양한 침입탐지시스템들이 사용되고 있다. 하지만 패턴 매칭을 기반으로 하는 오용탐지시스템의 경우, 새로운 공격에 대해서는 탐지가 불가능하며 수시로 침입 패턴을 추가시켜야 하는 어려움이 있다. 본 논문에서는 이러한 문제점을 해결하기 위하여 네트워크 데이터를 분석하여 새로운 침입 패턴을 자동 생성하는 시스템을 제안했으며, 그 시스템의 성능에 가장 중요한 역할을 하는 아웃라이어 탐지 기법에 대하여 논하였다. 1998년 DARPA Intrusion Detection Evaluation Program 데이터를 사용한 실험을 통해 클러스터링 후 아웃라이어 탐지기법으로 네트워크 침입 패킷을 찾아 낼 수 있었으며, 밀도함수 기반 아웃라이어 탐지기법이 기존의 유클리디언 거리함수 기반 아웃라이어 탐지기법에 비하여 높은 성능을 보임을 확인하였다. 향후 연구과제는 탐지된 아웃라이어의 특성을 추출해 내는 알고리즘의 연구와 추출된 특성을 정형화된 침입 패턴으로 변경하는 것이다.

참 고 문 헌

[1] J. Frank. " Artificial intelligence and intrusion detection: Current and future directions" , In Proceedings of the 17th National Computer Security Conference, 1994

[2] S. E. Smaha, Tools for Misuse Detection, In proceedings of ISSA' 93, Crystal City, VA, 1993.

[3] S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff. " A sense of self for unix processes." , In Proceedings of the 1996 IEEE Symposium on Security and Privacy, pp. 120-128, Los Alamitos, CA, 1996.

[4] C. Fraley and A. E. Raftery, " How many clusters? Which clustering method? Answers via model-based cluster analysis," The Computer Journal, vol. 41, pp. 578-588, 1988.

[5] Shi Zhong, Taghi M. Khoshgoftaar, and Naeem Seliya. " Evaluating Clustering Techniques for Network Intrusion Detection." , In Proceeding of the 10th ISSAT International Conference on Reliability and Quality Design, pp. 149-155. Las Vegas, Nevada, USA. 2004.

[6] MIT Lincoln Labs, DARPA intrusion detection evaluation. 1999.
In http://www.ll.mit.edu/IST/ideval/index.html

[7] Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Paliouras, G. and Spyropoulos, C. D., " An Evaluation of Naïve Bayesian Anti-Spam Filtering," In Proceeding of the 11th European Conference on Machine Learning, pp.9-17, 2000.

[8] Michael J.A.Berry and Gorden Linoff. " Data Mining, Techniques." John Wiley &Sons, 1997.