지능정보연구 제15권 제4호 2009년 12월(pp.23~36)

LBG 알고리즘 기반 데이터마이닝을 이용한 네트워크 침입 탐지율 향상

박성철
동국대학교 컴퓨터공학과
(scpark@dongguk.edu)

김준태
동국대학교 컴퓨터공학과
(jkim@dongguk.edu)

네트워크 침입 탐지는 데이터마이닝 기법을 활용하면서 지속적으로 발전하여 왔다. 데이터마이닝에 의한 침입 탐지 기법에는 클래스 레이블을 이용한 감독 학습과 클래스 레이블이 없는 비감독 학습 방법이 있다. 본 논문에서는 클래스 레이블이 없는 비감독 학습 방법인 LBG 클러스터링 알고리즘을 이용하여 네트워크 침입 탐지 정확도를 높이는 방법을 연구하였다. 임의의 초기 중심값들로 시작하여 유클리디언 거리 기반에 의해 클러스터링을 수행하는 K-means 방법은 잡음(noisy) 데이터와 이상치(outlier)에 대하여 취약하다는 단점이 있다. 비균일이진 분할에 의한 클러스터링 알고리즘은 초기값 없이 이진분할에 의해 클러스터링을 수행하며 수행 속도가 빠르다. 본 논문에서는 이 두 알고리즘의 장단점을 통합한 EM(Expectation Maximization) 기반의 LBG 알고리즘을 네트워크 침입 탐지에 적용하였으며, KDD 컵 데이터셋을 대상으로 한 실험을 통하여 LBG 알고리즘을 이용함으로써 침입 탐지의 정확도를 높일 수 있음을 보였다.

논문접수일 : 2009년 09월 15일 논문수정일 : 2009년 09월 28일 게재확정일 : 2009년 10월 10일 교신저자 : 김준태

1. 서 론

인터넷은 불순한 의도를 가진 해커(hacker)와 크래커(cracker)에 의해 항상 위협을 받고 있다. 인터넷이 위협을 받는 근본적인 문제점은 TCP/IP 프로토콜의 취약점에 기인하며, 따라서 인터넷에 연결된 각종 정보시스템을 보호하기 위해서는 시스템이 갖고 있는 자체 취약성을 분석하여 대처하거나, 네트워크의 접근 제어를 위한 침입차단시스템 및 네트워크의 패킷을 감시하는 침입탐지시스템 등 보안시스템을 구축해야 한다.

침입차단 시스템은 외부 침입을 차단하기 위해 기업 내부 네트워크와 외부 네트워크 사이에 위치하며 오고 가는 패킷의 상태를 파악하여 차단하는 기술이다. 그러나 컴퓨터 기술의 발전은 해킹 기술의 발전에 영향을 주어 크래커 및 해커의 침입을 차단하는 데는 한계가 있다. 또한 공격의 행태를 분석할 때 외부 침입보다 내부 사용자에 의한 공격이 더욱 더 심각하며 이러한 내부 사용자에 의한 공격은 침입차단시스템에 있어서 아무런 대책도 마련할 수 없다. 그러므로 내부 및 외부 사용자에 관계없이 공격을 효과적으로 막아낼 수 있는 침입탐지 시스템이 필요하다.

현재 침입탐지시스템은 공격 탐지에 있어서 시그너처(signature)를 이용한 패턴 매칭을 주로 사용하는데, 이러한 방식은 오탐율이 높다. 시그너처

* 이 논문은 2008년도 동국대학교 연구년 지원에 의하여 이루어졌음.

기반의 침입탐지 시스템은 침입이나 공격을 연구하는 전문가에 의해 공격 패턴이 연구된 후, 그 공격 패턴에 의지하여 침입탐지가 이루어지기 때문에 연구와 활용 사이에 시간 차이를 극복하기 어려운 것이 하나의 문제이며, 완전히 새로운 공격 패턴뿐 아니라 약간의 변형된 공격 패턴에 대해서도 잘 인식할 수 없는 구조로 되어 있다는 것도 또 다른 문제이다(Denning, 1987; Han et al., 2002).

전문가에 의해서 공격 패턴을 분석해야하는 문제를 어느 정도 극복하고 변형된 공격이나 새로운 공격에 대해서도 침입탐지를 수행할 수 있는 방법으로 비감독 학습(unsupervised learning)인 클러스터링(clustering)을 적용하는 방법이 연구되어 왔다(Lee, 1999; Portnoy et al., 2001). 대표적인 클러스터링 알고리즘인 K-means는 임의의 초기 중심값들로 시작하여 유클리디언 거리 기반에 의해 클러스터링을 수행하는 방법으로서 침입탐지에 적용하기가 용이하고 확장성(scalability)이 뛰어나지만 잡음(noisy) 데이터와 이상치(outlier)에 대하여 취약하다는 단점이 있다(Kanungo et al., 2002) 비균일이진 분할(non-uniform binary split)에 의한 클러스터링 알고리즘은 초기값 없이 이진분할에 의해 클러스터링을 수행하며 수행 속도가 빠르지만 K-means에 비하여 일반적으로 정확도가 떨어지는 경향이 있다. 본 논문에서는 이 두 알고리즘의 장단점을 통합한 EM (Expectation Maximization) 기반의 LBG 알고리즘을 네트워크 침입탐지에 적용하여 정확도를 높이는 방법을 제시하였다. LBG 알고리즘(Kinde et al., 1980)은 Y. Linde, A. Buzo, R. Gray(세 명의 이름 첫 자를 따서 LBG라 명명함)등에 의해 제안된 벡터 양자화(vector quantization) 기반 알고리즘으로서 J. Zheng 등(Zheng et al., 2006)에 의해 침입탐지에 사용될 수 있는 한 방법으로 소개된 바는 있으나 실제 적용 방안과 실험 결과가 제시된 바는 없다. 성능 분석을 위하여 KDD 컵 데이터셋을 대상으로 실험을 수행하였으며, 실험 결과에 의해 LBG 알고리즘을 이용함으로써 침입 탐지의 정확도와 재현율을 모두 높일 수 있음을 보였다.

본 논문의 제 2장에서는 네트워크 침입 탐지와 관련된 기존 연구들을 정리하며, 제 3장에서는 LGB 알고리즘을 적용한 침입 탐지 방법을 설명한다. 제 4장에서는 실제 데이터셋을 대상으로 제안한 침입탐지 방법의 성능을 실험한 결과를 분석하며, 제 5장에서 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

침입탐지 기법은 네트워크 시스템의 보안을 위하여 매우 중요한 요소이다. 초기의 침입탐지는 전문가를 활용한 오용탐지 방식에 많이 의존하였는데, 이 방식은 시그너처에 근거하여 필터링을 거쳐 탐지 엔진으로부터 올라온 패킷들을 비교하는 방식으로 이루어졌다. 오남용 탐지 방식은 패킷과 알려진 시그너처를 비교하는 방식이기 때문에 공격의 방식이 바뀌거나 새로운 방법의 공격이 가해지면 공격을 포착하지 못한다(Denning, 1987; Han et al., 2002).

이러한 문제점을 보완한 다른 방법으로 비정상 탐지 방식이 있는데, 정상적이고 평균적 상태를 기준으로 하여 탐지된 패킷이 이 기준에 상대적으로 급격한 변화를 일으키거나 확률이 낮은 사건이 발생할 경우 침입 탐지를 알리는 방법이다. 비정상 탐지를 수행하는 방법에는 여러 가지가 있다. 첫째, 가장 일반적인 탐지 방법은 정량분석으로서 경험적 임계치를 사용하는데, 정상적인 행위도 침입으로 간주하는 등 오탐율이 높은 편이다. 둘째, 통계적 방법은 사용자의 행위를 감시하며 프로파일

을 생성하여 저장한 뒤, 침입탐지 엔진에서 네트워크 패킷과 프로파일 데이터를 비교하여 이상행위를 평가한다. 셋째, 비특성 통계 분석으로서, 이벤트 데이터의 축약을 수행하며, 특성 통계 분석보다 탐지 속도와 정확성이 높으나 초과된 자원 사용으로 분석의 효율성과 정확성이 떨어진다는 단점이 있다(Ghosh, 1989; Lazarevic et al., 2003).

침입 패턴을 학습하여 분류에 사용하는 방법으로 많은 종류의 인공지능 또는 데이터마이닝 기법들이 존재한다. 침입탐지를 위해 사용된 감독학습으로는 분류 및 예측 기법들로는 의사결정나무에 의한 분류, 베이지안 분류, 신경망 학습, 서포트 벡터 머신, 연관규칙 분석, 퍼지집합 접근법 등이 있다(Chavan, 2004; Ghosh, 1998; Kruegel, 2003; Mukkamala et al., 2002).

침입탐지 데이터를 감독학습하기 위해서는 정보보호 전문가에 의해 클래스 레이블이 부여된 대량의 데이터가 필요하므로 비용과 시간이 많이 들게 된다. 또한 병합된 여러 네트워크를 다루는 경우 네트워크 별로 침입탐지 속성이 차이가 날 수 있고, 새로운 형태의 침입을 탐지하는 능력도 필요하기 때문에 비감독학습 방법에 대한 연구가 많이 이루어지고 있다(Lee, 1999; Lee, 2001; Portnoy, 2001; Zheng et al., 2006). 분할 방법(partitioning method)의 클러스터링은 K-means 알고리즘이 대표적인데 침입탐지에 적용하기가 용이하고 대용량의 데이터에서 적용할 수 있도록 확장성(scalability)이 뛰어나지만 미리 클러스터의 수를 정의해야 하며 잡음 데이터 및 이상치에 민감하다는 단점이 있다. 계층적 방법(hierarchical method)의 클러스터링은 클러스터 수를 미리 정의 할 필요가 없으며 나무구조로 군집화 되는 병합방법과 분할방법이 있다. 밀도기반 방법(density based method)의 클러스터링은 밀집 영역에서 객체 그룹화하며 임의의 모양을 가진 군집 발견이 용이하며 잡음 데이터 또한 잘 다룰 수 있다. 그러나 클러스터링을 시작하기 전 매개변수인 반경과 반경(ε) 내의 최소 객체 개수를 미리 정의해야 하는 문제 때문에 매개변수의 선택이 학습에 상당히 영향을 미친다. 격자기반 방법(grid based method)의 클러스터링은 객체 공간을 격자구조로 만드는 유한개의 셀로 분할하는 방법을 이용하여 벡터를 클러스터링 한다.

그외에 모형 기반 방법(model based method)의 클러스터링은 수학적 모델을 정의해 놓고 데이터를 그것에 맞춰 학습하는 기법으로써 통계적 방법(statistical method), 개념적 방법(conceptual method), 그리고 신경망 방법(neural network method) 등이 있다(Lazarevic, 2003; Lichodzijewski et al., 2002). 통계적 방법인 EM(Expectation Maximization) 알고리즘은 확률 분포에 따라 각 객체를 군집에 할당하며 새로운 평균은 가중된 측도에 기반을 두어서 계산한다. 개념적 방법은 모형이 없는 객체들의 집합이 주어졌을 때, 모델에 적합하도록 객체들의 분류 구조를 생성하는 방식이다. 신경망 방법인 SOM(self organizing map)은 경쟁적 학습 구조로서 현재 객체에 대해 경쟁하는 몇 개의 유닛을 가지며 가중치 벡터는 현재 객체에 가장 가까운 유닛이 선택되어 가중치를 수정하며 계속적으로 학습을 해나가는 방식이다.

본 논문에서는 클러스터링 방법인 K-means와 비균일이진 분할의 장단점을 통합한 LBG 알고리즘을 침입탐지에 사용하는 방법을 제안하였으며 실험을 통해 성능을 분석하였다. LBG 알고리즘의 장점은 데이터의 양과 클러스터의 수에 관계없이 정확도(precision), 재현율(recall), 그리고 이 두 측정치를 단일하게 계산한 F-측도(F-measure) 등에 있어서 안정적으로 높은 성능을 보인다는 것이다.

3. LBG 알고리즘을 이용한 침입탐지

3.1 침입탐지 데이터 값의 정규화

침입탐지 데이터 집합 x = {x1, x2, …, xn}에는 TCP/IP의 프로토콜 자체 특성, 네트워크의 규모, 및 서비스의 종류에 따라 다양한 편차들을 가진 값들이 존재할 수 있다. 이러한 편차를 줄이기 위해 침입탐지 데이터 값을 정규화 함수식 (1)에 의해 정규화(Normalization)한 다음 클러스터링 기법인 K-means와 LBG 알고리즘을 적용하였다.

$$N(x) = \frac{x-\mu}{\sigma}$$

x : 관측값, μ : 평균, σ : 표준편차 (1)

침입탐지 데이터는 초기 단계인 정규화가 끝난 후 바로 클러스터링 알고리즘에 적용되며, 클러스터링 알고리즘들에 사용될 기본 함수들의 정의는 다음과 같다. 식 (2)는 침입탐지 데이터 집합 $X = \{x_1, x_2, \cdots, x_n\}$과 클러스터 중심 집합 $Y = \{y_1, y_2, \cdots, y_k\}$에서 어떤 데이터 벡터 x와 중심 벡터 y간의 거리(distance, d)를 계산하는 유클리디언 거리 함수인 $d(x, y)$를 나타내며, 식 (3)은 새로운 클러스터 집합 $X = \{X_1, X_2, \cdots, X_K\}$의 각 클러스터 중심(centroid, c)을 계산하는 함수 $c(X_j)$를 나타낸다. 식 (4)는 클러스터링을 얼마정도까지 진행할지를 결정하기 위해 사용하는 각 클러스터의 왜곡(distortion) D_j를 계산하는 함수이다. 각 식에서 i는 데이터 집합에 있는 어떤 벡터의 인덱스, j는 클러스터 중심 집합 y 또는 클러스터 집합 X의 원소 인덱스를 가리키며, $y_{j(i)}$는 데이터 집합의 원소 i가 속하는 클러스터 j의 중심 벡터를 나타낸다.

$$d(x, y) = \sqrt{(x-y)^T(x-y)} \quad (T: \textit{transpose}) \quad (2)$$

$$c(X_j) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (3)$$

$$D_j = \sum_{i=1}^{n} d(x_i, y_{j(i)}) \quad (4)$$

3.2 K-means와 비균일이진 분할 알고리즘에 의한 클러스터링

침입탐지 데이터를 클러스터링(clustering)하는데 있어서 중요한 문제는 최적의 클러스터들의 중심 집합(centroid set; codebook) 찾기, 양자화(quantization), 그리고 코드북 거리의 인식 등이다. 침입탐지 데이터는 아주 방대하여 감독 학습을 수행하기 위해서는 막대한 비용과 시간이 소요된다. 클러스터링에 의한 비감독 학습 방법은 클래스 레이블이 존재하지 않아도 가능하기 때문에 최적의 코드북을 찾아낼 수 있다면 감독 학습보다 더 뛰어난 성능을 보일 수 있다. 본 연구에서는 침입탐지 데이터에 대해 위의 문제들을 잘 해결하여 적용할 수 있도록 K-means 알고리즘과 비균일(non-uniform)이진 분할 알고리즘을 결합한 LBG 알고리즘을 사용하여 클러스터링을 수행하였다(Patan, 2001; Zheng et al., 2006).

K-means 기법은 침입탐지 데이터 집합 $X = \{x_1, x_2, \cdots, x_n\}$으로부터 임의의 k개의 벡터를 선택하여 k개의 초기 중심 집합 $y = \{y_1, y_2, \cdots, y_k\}$을 생성함으로써 시작한다. 분류를 최적화하기 위해 EM (Expectation Maximization)을 이용하여, E-단계에서는 데이터 x_i가 y_j에 가장 가깝다면 클러스터 X_j에 속하도록 레이블링하여 데이터 집합을 K개의 클러스터들 $X = \{X_1, X_2, \cdots, X_K\}$로 나누고, M-단계에서는 E-단계에서 구한 새로운 클러스터들

에서 각각의 중심을 갱신한다.

$$X_i = \{x_i \mid d(x_i, y_i) \leq d(x_i, y_{i'})\} \quad (5)$$

$$y_i = c(X_i) \quad (6)$$

이러한 EM 단계를 거치면서 데이터 벡터 집합 x가 새로운 클러스터 중심 벡터 집합 y로 클러스터링 되는 과정에서 특정 벡터 간 정의된 거리 척도에 의해 클러스터링 오차가 발생하게 된다. 오차를 측정하는 왜곡 척도는 일반적으로 유클리디언 거리 함수를 가장 많이 사용하게 되고 이 척도를 이용하여 집합의 멤버들과 거리의 합을 최소화하는 점들로 구성하는 과정으로 클러스터 집합 X의 중심 c(X)를 계산하여 클러스터 중심 벡터 집합 y를 결정하게 된다. 유클리디언 거리를 척도로 사용할 경우 클러스터 중심 벡터 y_i는 클러스터 집합 X_i의 중심 평균이 된다. K-means는 총 왜곡이 변치 않거나 설정된 반복 횟수에 도달 할때까지 계속 반복하는 방법으로 진행되어 결과인 클러스터 중심 벡터 y를 내놓게 된다.

비균일이진 분할은 클러스터 개수 $k=1$로 한 개의 클러스터 X_i와 중심이 $y_i = c(X_j)$인 모든 데이터 벡터 집합 x로부터 시작하여 아래와 같은 단계들을 (K-1)번 반복함으로써 K개의 클러스터 중심 집합 y를 얻는다.

단계 1 : 클러스터 집합 X의 벡터들과 클러스터 중심 집합 y의 평균 거리로 측정된 가장 큰 왜곡을 가진 클러스터 X_j를 선택한다. 만약 왜곡 변화가 없으면 양자화를 끝내게 된다. 식 (7)에서 D_j는 X_j에 해당하는 왜곡 값이며, k_j는 클러스터 X_j의 원소의 수, $X_{i(j)}$는 클러스터 X_i에 소속된 데이터 집합의 벡터 x_i이다.

$$D_j = \frac{1}{k_j}\sum_{i=1}^{k_j} d(x_{i(j)}, y_j) \quad (7)$$

단계 2) $K=2$로 클러스터 X_j상에서 K-means를 행하거나, 클러스터 X_j의 주고유벡터(principal eigenvector) v_j를 결정하고 클러스터 X_j에 있는 점들 중 $(y_j + v_i)$에 가까운 점들을 X_a로 $(y_j + v_i)$에 가까운 점들을 X_b로 하여 클러스터 X_j를 X_a와 X_b 두 개로 나눈다. 주고유벡터는 클러스터 X_j의 평균과 공분산을 구하고 고유분석을 통하여 고유치(eigenvalues)중 값이 가장 큰 고유벡터를 구하면 된다.

단계 3 : 중심 y_j를 대치하는 새로운 중심을 다음과 같이 둔다.

$$y_j = c(X_a), \quad y_{k+1} = c(X_b) \quad (8)$$

단계 4 : 클러스터 카운터(counter)를 증가 시킨다($k \leftarrow k+1$k).

3.3 LBG 알고리즘에 의한 클러스터링

K-means는 초기 중심값들을 임의적으로 선택하기 때문에 잡음 데이터 및 이상치에 민감하다는 특성을 가지고 있다. 이 알고리즘과 비균일 이진분할 알고리즘을 결합한 알고리즘이 LBG 알고리즘이다. K-means 알고리즘은 비균일 이진분할 알고리즘보다 속도는 느리지만 정확성 면에서는 우수한 특성을 가지고 있고, 비균일 이진분할 알고리즘

은 정확도는 떨어지지만 속도 면에서는 K-means 알고리즘보다 더 우수한 특성을 가지고 있다. LBG 알고리즘은 이렇게 상호 보완적인 특성을 지닌 두 알고리즘을 병합하였다. 빠른 속도를 가진 비균일 이진분할에 의해 초기 클러스터를 이진으로 분할하는 전략을 사용하고 나누어진 두 클러스터를 클러스터 중심 벡터에 최적으로 되도록 왜곡 값을 계속 계산하는 K-means 알고리즘 방식을 사용하였다.

LBG 알고리즘에서는 K-means의 초기값을 랜덤 데이터 벡터로 하는 대신에 비균일 이진분할에서 사용한 이진분리로 구한 중심을 사용하기 때문에 K-means 방법을 사용한 경우보다 더 좋은 클러스터링 결과를 얻을 수 있다. 따라서 LBG 알고리즘을 사용하면 더 우수한 중심들(centroids)의 생성이 가능하며 따라서 침입 탐지의 정확도를 높일 수 있다. LBG 알고리즘의 실행 단계는 다음과 같다.

단계 1 : 침입탐지 데이터 집합 x를 전체 평균을 내어 하나의 클러스터 중심 벡터를 생성한다.

$$y_0 = c(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (9)$$

단계 2 : 클러스터 중심 벡터에 아주 작은 값(예 : ε = 0.01)을 한번은 더하고, 한번은 빼어서 두 개의 중심 벡터를 만든다. ε 값을 아주 작은 값으로 잡은 이유는 왜곡 한계를 극복하여 정교하게 나눌 수 있기 때문이다. 너무 작은 값을 설정할 경우 클러스터링 시간이 길어질 수 있다.

$$y_a = y_0 + \varepsilon \qquad (10\text{-}1)$$

$$y_b = y_0 + \varepsilon \qquad (10\text{-}2)$$

단계 3 : 침입탐지 데이터 집합 x는 이 두 중심 벡터 y_a, y_b와 거리 계산을 하여 자신이 속할 클러스터를 찾는다.

$$X_a = \{x_i \mid d(x_i, y_a) \le d(x_i, y_b)\} \qquad (11\text{-}1)$$

$$X_b = \{x_i \mid d(x_i, y_b) \le d(x_i, y_a)\} \qquad (11\text{-}2)$$

단계 4 : 클러스터 X_a와 X_b를 평균하여 새로운 각 클러스터 중심 벡터를 계산한다.

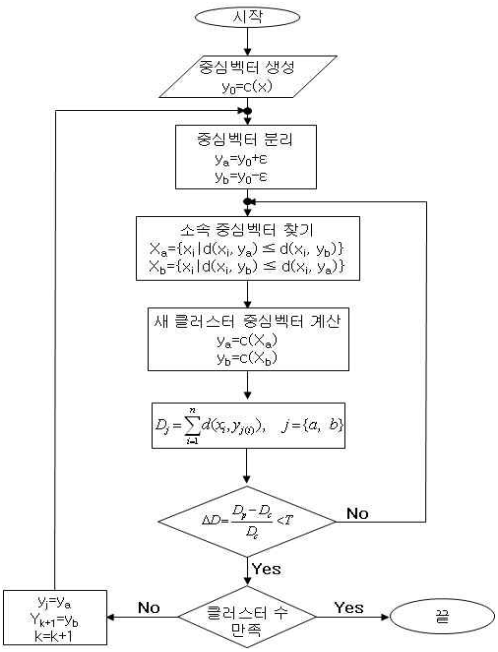
$$y_a = c(X_a) \qquad (12\text{-}1)$$

$$y_b = c(X_b) \qquad (12\text{-}2)$$

단계 5) 단계 3과 단계 4를 반복하여 왜곡이 기준치(threshold, T) 이하가 되도록 한다. $y_{j(i)}$는 데이터 집합의 원소인 인덱스 i가 속하는 클러스터 j의 중심 벡터라는 의미이다.

$$D_j = \sum_{i=1}^{n} d(x_i, y_{j(i)}),\ j = \{a, b\} \qquad (13)$$

단계 6 : 원하는 수의 중심 벡터가 생성되었으면 종료하고 그렇지 않으면 $y_i = y_a$ 및 $y_{k+1} = y_a$로 변경하고 클러스터 카운트 k를 하나 증가시킨 뒤 단계 2로 간다.



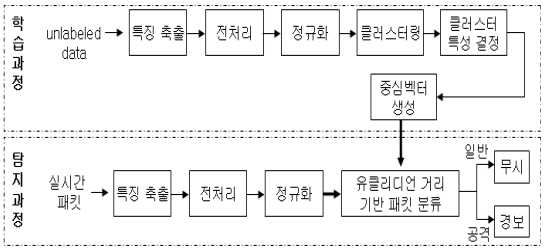
<그림 1> 침입탐지 훈련 벡터를 이용한 LBG 알고리즘 실행 과정

LBG 알고리즘 실행과정에서 왜곡이 안정되었는지 아닌지를 확인하는 테스트는 이전 반복시의 왜곡과 현재 왜곡에 대한 상대적인 감소 값을 점검함으로써 이루어지게 된다. <그림 1>의 알고리즘 실행과정에서 왜곡의 변화율 ΔD를 검사하는 수식은 이전(previous) 왜곡 D_p와 현재(current) 왜곡 D_c만으로 계산하도록 되어있고, 이 값이 10^{-4} 이하가 되면 완전히 수렴(convergence)하지 않아도 실행을 멈추도록 되어 있다.

3.4 LBG 알고리즘에 의한 공격 데이터의 탐지

본 논문에서 적용한 LBG 클러스터링에 의해 공격 데이터를 탐지하는 방법은 다음과 같다. 우선 침입 탐지 데이터를 대상으로 제 3.3절에서 설명한 클러스터링 과정을 거쳐 클러스터 중심 벡터들을 생성한다. 생성된 각 클러스터의 특성은 침입 데이터와 일반 데이터가 어떤 분포 또는 확률 모수(parameter)에 관련된 통계값(statistical value)에 의존적이거나 이상치(outlier)를 찾는 경우 자동적으로 결정될 수 있으며, 경우에 따라서는 보안전문가가 직접 결정할 수도 있다. 본 연구에서는 KDD 컵 1999 데이터셋을 사용하였으며, 비감독 학습에 적합하도록 데이터에 있는 클래스 레이블(class label)을 제외시킨 다음 LBG 알고리즘과 K- means 알고리즘을 수행하여 클러스터 중심 벡터, 즉 코드북을 생성하였다. 클러스터 중심 벡터들이 생성된 후 각 클러스터의 특성은 보안전문가에 의하여 판단된다고 가정하고 데이터셋에 주어진 클래스 레이블을 참고하여 결정되도록 하였다. 클러스터 내에 있는 데이터 벡터들에 따라 공격 패킷의 특성이 높으면 공격 클러스터로 분류되고 일반 패킷의 특성이 높으면 일반 클러스터로 분류하였다. 대부분 한 쪽 특성(공격 또는 일반)이 압도적으로 나타났기 때문에 공격 패킷인지 일반 패킷인지 애매한 경우는 발생하지 않았다. 실제 적용에서는 클래스 레이블이 없는 데이터만을 가지고 클러스터링을 수행한 후 각 클러스터의 특성을 전문가가 또는 통계적 특성에 의해 결정하고 침입탐지를 수행한다.

클러스터 특성이 결정되고 중심 벡터들이 얻어지고 나면, 공격을 탐지하기 위해서는 네트워크 카드로부터 무작위 모드(promiscuous mode)에 의해 감지된 패킷들을 특징 축출, 전처리 및 정규화 과정을 거쳐 저장된 중심 벡터와 유클리디언 거리를 계산하여 비교한 후, 일반 패킷이면 무시하고 공격 패킷이면 경보를 발생시킨다. 이와 같은 학습 및 침입탐지 과정을 <그림 2>에 나타내었다. 본 논문의 실험에서는 제외되었던 클래스 레이블을 참고



<그림 2> LBG 클러스터링에 의한 공격 탐지 과정

하여 클러스터링에 의한 침입탐지의 정확도를 계산하였다.

4. 실험 및 결과

LBG 알고리즘에 의한 침입 탐지 방법의 성능을 평가하기 위해 본 논문에서는 일반적으로 많이 사용되는 KDD 컵 1999 데이터셋들을 이용하였다. 이는 DoS(Denial of Service) 공격 5종류, R2L (Remote to Local) 9종류, U2R(User to Root) 4종류, Probing 4종류 등 모두 22종류의 침입 패턴을 포함한다. 아래 <표 1>은 22종류의 침입 패킷들과 일반 패킷들의 비율을 정리한 것이다.

성능분석은 침입으로 판단되는 클러스터 내에서 실제로 얼마나 많은 침입 데이터들이 포함되었는가를 나타내는 정확도(Precision, P)과 침입 데이터의 전체 패킷 중 얼마나 많은 침입 패킷들이 적용된 알고리즘에 의하여 실제로 검출 되었는가를 평가하는 재현율(Recall, R)을 True Positive, False Positive, False Negative에 의해 계산하여 구하였다.

전체 침입탐지 데이터 집합 중에서 침입인데 침입이라고 판정한 True Positive의 집합을 TP, 침입이 아닌데 침입이라고 판정한 False Positive의 집합을 FP, 그리고 침입인데 침입이 아니라고 판정한 False Negative의 집합을 FN라고 하자. 정확도 P와 재현율 R의 계산식은 식 (14)와 같다. 또한 단일한 측정치를 위해 정확도와 재현율을 하나로 합한 F-측도(F-Measure)를 식 (15)와 같이 구하였으며, P와 R사이의 중요도를 균등하게 생각하여 β = 1로 설정하였다.

$$R = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \quad P = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{14}$$

$$F = \frac{(1+\beta^2) * R * P}{\beta^2 (R+P)} \tag{15}$$

침입탐지 데이터 7만 건에 대한 실험 결과인 <표 2>을 보면 LBG와 K-means 알고리즘을 비교했을 때 평균적으로 F-측도에 의한 결과는 LBG 알고리즘이 K-means 알고리즘보다 7% 정도 앞섰고, 정확도는 LBG가 0.91 K-means가 0.82로, 재현율은 LBG가 0.99 K-means가 0.87로 모두

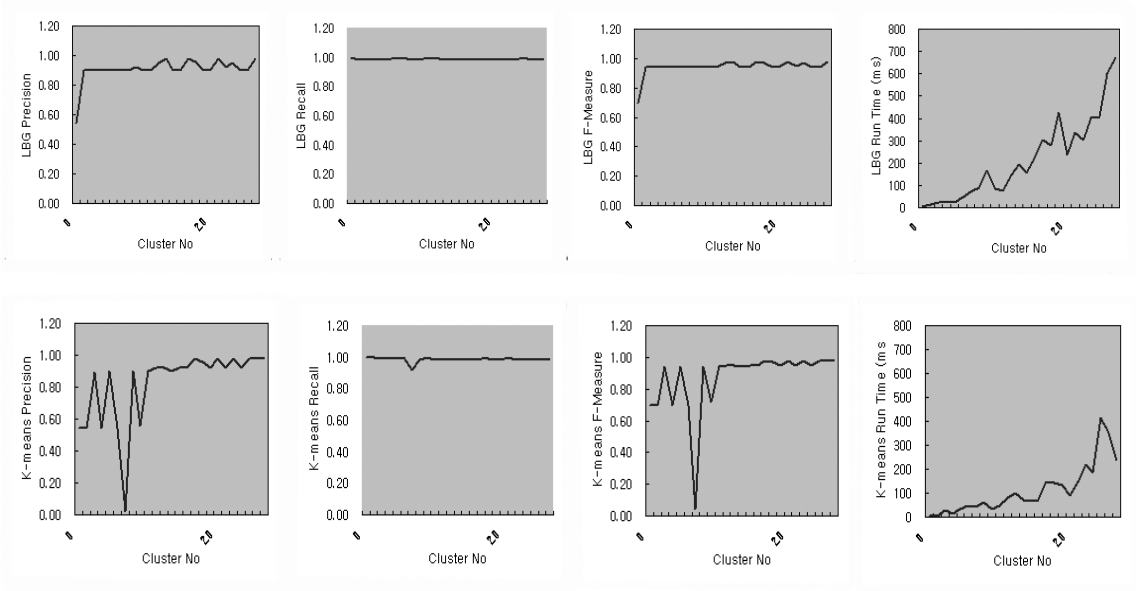
<표 1> KDD 컵 1999 데이터셋의 일반 패킷들과 공격 패킷들의 비율

Seq.	Packet Type	Ratio %	Seq.	Packet Type	Ratio %	Seq.	Packet Type	Ratio %	Seq.	Packet Type	Ratio %
1	smurf.	56.838	7	portsweep.	0.211	13	buffer_overflow.	0.0061	19	ftp_write.	0.0016
2	neptune.	21.700	8	warezclient.	0.206	14	land.	0.0043	20	multihop.	0.0014
3	normal.	19.691	9	teardrop.	0.198	15	warezmaster.	0.0040	21	phf.	0.0008
4	back.	0.446	10	pod.	0.053	16	imap.	0.0024	22	perl.	0.0006
5	satan.	0.322	11	nmap.	0.047	17	rootkit.	0.0020	23	spy.	0.0004
6	ipsweep.	0.252	12	guess_passwd.	0.011	18	loadmodule.	0.0018	Total		100

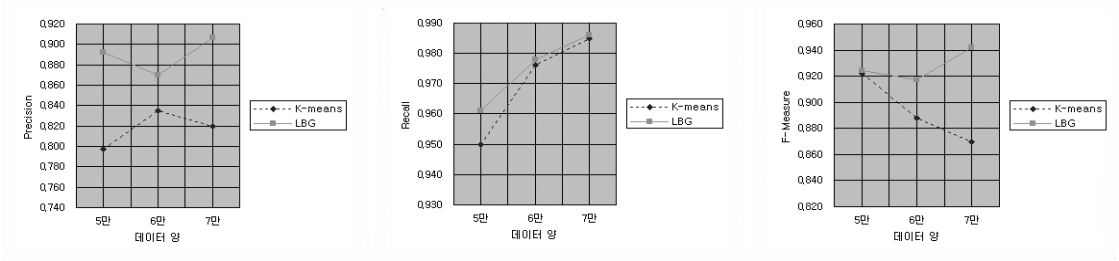
<표 2> 침입탐지 데이터 7만 건과 36 칼럼에 대한 LBG와 K-means 알고리즘의 클러스터 수의 변화에 따른 정확도, 재현율, F-측도, 그리고 훈련시간의 변화

Algorithm / Cluster Number	K-means(70,000 rows, 36 columns)			
	Precision	Recall	F-Measure	Run Time(ms)
2	0.54	1.00	0.70	7
3	0.54	1.00	0.70	7
4	0.89	0.99	0.94	30
5	0.54	1.00	0.70	12
6	0.90	0.99	0.94	31
7	0.54	1.00	0.70	50
8	0.02	0.92	0.04	45
9	0.90	0.98	0.94	64
10	0.56	1.00	0.72	34
11	0.90	0.98	0.94	48
12	0.92	0.98	0.95	82
13	0.92	0.98	0.95	98
14	0.90	0.99	0.94	71
15	0.92	0.98	0.95	69
16	0.92	0.98	0.95	71
17	0.98	0.99	0.98	146
18	0.96	0.99	0.97	145
19	0.92	0.98	0.95	133
20	0.98	0.99	0.98	90
21	0.92	0.99	0.95	148
22	0.98	0.98	0.98	221
23	0.92	0.98	0.95	186
24	0.98	0.99	0.98	416
25	0.98	0.98	0.98	354
26	0.98	0.98	0.98	236
Mean	0.82	0.98	0.87	111.76

Algorithm / Cluster Number	LBG(70,000 rows, 36 columns)			
	Precision	Recall	F-Measure	Run Time(ms)
2	0.54	1.00	0.70	5
3	0.90	0.98	0.94	12
4	0.90	0.99	0.94	22
5	0.90	0.98	0.94	22
6	0.90	0.98	0.94	24
7	0.90	0.99	0.94	45
8	0.90	0.99	0.94	70
9	0.90	0.99	0.94	90
10	0.92	0.98	0.95	167
11	0.90	0.99	0.94	84
12	0.90	0.99	0.94	77
13	0.95	0.99	0.97	149
14	0.98	0.98	0.98	193
15	0.90	0.99	0.94	158
16	0.90	0.98	0.94	225
17	0.98	0.98	0.98	302
18	0.96	0.99	0.97	278
19	0.90	0.98	0.94	425
20	0.91	0.98	0.95	236
21	0.98	0.98	0.98	338
22	0.92	0.98	0.95	304
23	0.95	0.99	0.97	403
24	0.90	0.99	0.94	406
25	0.90	0.98	0.94	603
26	0.98	0.99	0.98	670
Mean	0.91	0.99	0.94	212.32



<그림 3> 침입탐지 데이터 7만 건과 36 칼럼에 대한 LBG와 K-means 알고리즘의 클러스터 수의 변화에 따른 정확도, 재현율, F-측도, 그리고 훈련시간의 변화



<그림 4> 침입탐지 데이터의 양의 변화에 따른 평균 정확도, 재현율, F-측도에 대한 꺽은 선의 비교

LBG 알고리즘이 높은 성능을 나타내었다. 또한 <그림 3>에서 클러스터 수의 변화에 따른 정확도를 보면 LGB가 훨씬 안정적인 결과를 보임을 알 수 있다. 그러나 훈련시간은 평균적으로 LBG가 212.32초, K-means 111.76초로 K-means보다 LBG가 2배 정도로 오래 걸렸다.

또한 데이터의 양을 변화시키면서 실험한 결과인 <그림 4>을 보면 실행 속도 면에서는 K-means 알고리즘이 LBG 알고리즘보다 앞서지만 데이터의 양이 증가할수록 정확도는 LBG가 K-means보다 훨씬 높은 수치를 보였다. 재현율은 두 알고리즘이 비슷하였지만 LBG 알고리즘이 약간은 우세하게 나왔으며, F-측도에 있어서는 K-means는 수치가 하강하고 있지만 LBG 알고리즘은 조금씩 상승하여 데이터가 더 많은 7만 건에 대해서는 훨씬 더 우세하다는 것을 볼 수 있다. 전체적인 관점에서 보자면 수행 시간을 제외한 나머지 측정치인 정확도, 재현율, 그리고 F-측도에 있어서 K-means보다 LBG가 더 우세한 알고리즘이라고 할 수 있다.

5. 결론

본 논문에서는 비감독학습 방법인 K-means 알고리즘과 비균일 이진분할(non-uniform binary split) 알고리즘의 장단점을 통합한 LBG 알고리즘을 적용하여 네트워크 침입탐지 정확도를 높이는 연구를 하였다. K-means 알고리즘의 경우 초기에 임의의 클러스터 중심 벡터를 생성하여 학습을 하기 때문에 정확도가 안정적이지 못한 반면, LBG 알고리즘은 이진분할의 방법에 의해 클러스터를 둘로 나눈 후 왜곡 값의 차이가 적어지도록 계속하여 클러스터 중심 벡터를 변경하기 때문에 대부분의 경우에 안정적이었다. 일반적으로 잡음 데이터와 이상치에 민감한 K-means 알고리즘보다 LBG 알고리즘을 적용한 경우 데이터의 양이 늘어나고 군집의 수가 증가해도 높은 정확성을 보였다.

그러나 대용량 데이터와 침입탐지 데이터의 많은 차원은 학습과 탐지에 있어서 성능을 저하시키므로 높은 차원에서의 정보를 유지하면서 낮은 차원으로 차원을 축소시켜 데이터의 양을 줄이고, 상관이 있는 변량들의 변동 또는 분산을 줄이기 위한 연구가 필요하다. 실험에서 LBG 알고리즘은 그 특성상 K-means 알고리즘에 비하여 수행 시간이 약 2배 정도 더 걸렸으나, 차원 축소의 적용으로 그 차이를 줄일 수 있을 것으로 기대된다. 다음 연구에서는 본 논문과 연계하여 대용량 침입탐지 데이터와 다차원 데이터의 차원 축소를 통해 효율성을 높이는 방법과 이에 따라 K-means에 비해 데이터의 양과 클러스터 수가 증가함에 따라 수행 시간이 점차 늘어나는 문제를 해결하기 위한 연구를 하고자 한다.

참고문헌

Breunig, M., H.-P. Kriegel, R. T. Ng, J. Sander, "LO F : identifying density-based local outliers", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Vol.29, No.2(2000), 93~104.

Chavan, S., K. Shah, N. Dave, S. Mukherjee, A. Abraham, "Adaptive neuro-fuzzy intrusion detection systems", *Information Technology : Coding and Computing*, Vol.1(2004), 70~74.

Denning, D. E., "An intrusion-detection model", *IEEE Transactions on Software Engineering*, Vol.SE-13, No.2(2004), 222~232.

A.K. Ghosh, A. Schwartzbard, "A Study in Using Neural Networks for Anomaly and Misuse Detection", *Proceedings of the 7th USENIX Security Symposium*, 1998.

Han, H., X. L. Lu, J Lu, C. Bo, R. L. Yong, "Data mining aided signature discovery in network-based intrusion detection system", *ACM SIGOPS Operating Systems Review*, Vol.36, No.4(2002), 7~13.

Kanungo, T., D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu., "An efficient K-means clustering algorithm : analysis and implementation", *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No.7(2002), 881~892.

Kruegel, C., D. Mutz, W. Robertson, F. Valeur, (2003), "Bayesian event classification for intrusion detection", *Proceedings of the 19th Annual Computer Security Applications Conference*, (2003), 14~23.

Lazarevic, A., L. Ertoz, V. Kumar, A. Ozgur, "A comparative study of anomaly detection schemes in network intrusion detection", *Proceedings of the Third SIAM International Conference*, (2004), 25~36.

Lee, W., S. J. Stolfo, K. W. Mok, "A data mining framework for building intrusion detection models", *IEEE Symposium on Security and Privacy*, (1999), 120~132.

Lee, W., S. J. Stolfo, P. K. Chan, E. Eskin, W. Fan, "Real time data mining-based intrusion detection", *DARPA Information Survivability Conference* (2001).

Lichodzijewski, P., A. Zincir-Heywood, and M. Heywood., "Dynamic intrusion detection using self-organizing maps", *The 14th Annual Canadian Information Technology Security*, 2002.

Linde, Y., A. Buzo, R. Gray, "An Algorithm for Vector Quantizer Design", *IEEE Transaction on Communications*, Vol.28 No.1(1980), 84~94

Mukkamala, S., G. Janoski, A. Sung, "Intrusion detection using neural networks and support vector machines", *Proceedings of IEEE International Joint Conference on Neural Networks*, (2002), 1702~1707.

Patan, G. and M. Russo., "The enhanced LBG algorithm", *Neural Networks*, Vol.14, No.9(2001), 1219~1237.

Portnoy, L., E. Eskin, S. Stolfo, "Intrusion detection with unlabeled data using clustering", *Proceedings of ACM CSS Workshop on Data Mining* 2001.

Yahia, M. E., B. A. Ibrahim, "K-nearest neighbor and C4.5 algorithms as data mining methods-advantages and difficulties", *Computer Systems and Applications*, 2003.

Zheng, J. and M. Hu, "An Anomaly Intrusion Detection Sys Based on Vector Quantization", *IEICE-Transactions on Information and Systems archive*, Vol. E89-D, No.1, (2006), 201~210.

Abstract

Improvement of Network Intrusion Detection Rate by Using LBG Algorithm Based Data Mining

Seongchul Park* · Juntae Kim*

Network intrusion detection have been continuously improved by using data mining techniques. There are two kinds of methods in intrusion detection using data mining-supervised learning with class label and unsupervised learning without class label. In this paper we have studied the way of improving network intrusion detection accuracy by using LBG clustering algorithm which is one of unsupervised learning methods. The K-means method, that starts with random initial centroids and performs clustering based on the Euclidean distance, is vulnerable to noisy data and outliers. The non-uniform binary split algorithm uses binary decomposition without assigning initial values, and it is relatively fast. In this paper we applied the EM(Expectation Maximization) based LBG algorithm that incorporates the strength of two algorithms to intrusion detection. The experimental results using the KDD cup dataset showed that the accuracy of detection can be improved by using the LBG algorithm.

Key Words : Network Intrusion Detection, Data Mining, LBG Clustering

* The Department of Computer Engineering, Dongguk University

저 자 소 개

박성철

동국대학교 정보관리학과를 졸업하고 2005년 고려대학교 전자컴퓨터공학과에서 공학석사를 취득하였으며 2009년 동국대학교 컴퓨터공학과 박사수료 후 네트워크 보안과 데이터 마이닝을 결합하는 연구와 시간강의를 하고 있다. 주요 경력으로는 원스테크넷 부설보안연구소에서 과장으로 보안관제시스템을 연구 및 개발하였고 넷시큐어테크놀러지에서는 선임연구원으로 침입탐지시스템을 연구 및 개발을 하였다. 그리고 서울호서전문학교 인터넷정보보안과 전임교수를 역임하였다. 한국지능정보시스템학회, 한국정보보호학회, 한국정보과학회 등의 정회원이며, 관심분야로는 기계학습, 데이터마이닝, 네트워크 보안 등이다.

김준태

서울대학교 제어계측공학과를 졸업하였고, 1990년과 1993년에 미국 University of Southern California에서 전기공학 석사 및 컴퓨터공학 박사 학위를 각각 취득하였다. 1994년에 미국 Southern Methodist University에서 Postdoctoral Research Associate로 재직하였으며, 1995년부터 현재까지 동국대학교 컴퓨터공학과 교수로 재직중이다. 2003년에서 2004년까지는 미국 Oregon State University의 방문교수였다. 한국정보과학회, 정보처리학회, 미국 IEEE, ACM 등의 정회원이며, 관심분야는 인공지능, 기계학습, 데이터마이닝, 추천시스템 등이다.