실시간 침입탐지 시스템에 관한 연구

김병주*

A Study on Realtime Intrusion Detection System

Byoung-joo kim

요 약

인공지능, 기계학습 및 데이터마이닝 기법들을 침입탐지 시스템에 적용하는 연구가 활발히 진행되고 있다. 그러나 많은 연구가 공격패턴의 분류를 위한 분류기(classifier)의 학습 알고리즘 성능 개선에 목적을 두고 있다. 그리고 이러한 학습 알고리즘은 대부분 일괄처리(batch) 방식으로 동작하여 실시간 침입탐지 시스템의 적용에는 적합하지 못하다. 본 논문에서는 실시간 침입탐지 시스템을 위한 점증적 특징 추출 기법과 분류가 가능한 실시간 침입탐지 시스템을 제안한다. 제안된 방법을 KDD CUP 99 자료에 적용한 결과 실시간 기법임에도 불구하고 일괄처리 방식과 비슷한 결과를 나타내었다.

Abstract

Applying artificial intelligence, machine learning and data mining techniques to intrusion detection system are increasing. But most of researches are focused on improving the performance of classifier. These classifiers are performed by batch way and it is not proper method for realtime intrusion detection system. We propose an incremental feature extraction and classification technique for realtime intrusion detection system. Applying proposed system to KDD CUP 99 data, experimental result shows that it has similar capability compared to batch way intrusion detection system.

키워드

실시간 침입탐지시스템, 점증적 특징추출, 기계학습

I. 서 론

침입탐지 시스템(intrusion detection system)은 네트워크상의 컴퓨터에 대해 내부 사용자의 악의적인 사용이나 외부의 공격에 대해 자동으로 감지하여 필요한 대응을 취하는 시스템을 말한다. 초기의 침입탐지 시스템들은 이미 알려진 공격 패턴들을 기반으로 전문가 시스템(expert system)에 인코딩(encoding)하여 침입 여부를 판단하였다. 이러한 방식에 의한 규칙의 생성 및 추가되는 공격패턴에 대한 확장은 매우 어려우며 또한 효율성도 떨어진다. 이러한 문제를 해결하기 위해 인공지능, 기계학습 및 데이터마이닝 기법들을 침입탐지 시스템에 적용하는 연구가 활발히 진행되고 있다. 그러나 많은 연구가 공격패턴의 분류를 위한 분류기(classifier)의 학습 알고리즘 성능 개선에 목적을 두고 있다. 그리고 이러한 학습 알고리즘은 대부분 일괄처리(batch) 방식으로 동작하여 실시간 침입탐지 시스템의 적용에는 적합하지 못하다.

현실세계에 적용 가능한 침입탐지 시스템은 실시간으로 발생하는 공격패턴에 대하여 침입탐지를 수행하여야 하므로 다음과 같은 요소들을 갖추어야 한다. 첫째 전처리(preprocessing)과정에서 점증적 특징추출(incremental feature extraction) 방법에 의해 패턴의 크기를 줄여 학습 시간을 최소화하여야 할뿐 아니라 기하급수적으로 늘어나는 침

* 영산대학교 정보통신학과

접수일자 : 2004. 9. 10

입 패턴에 대해서도 적용 가능하도록 확장성(scalability)이 있어야 한다. 공격 패턴분류 시 효과적인 특징추출은 학습시간의 감소, 분류기의 성능 향상 및 복잡도(complexity)를 감소시키는 등 효율적인 패턴분류를 가능하게 한다. 실제 패턴인식 시스템에서 많은 수의 특징을 사용하는 것은 분류비용(classification cost) 측면에서 볼 때 매우 비효율적이다. 따라서 효율적인 분류기를 구현하기 위해서는 가능한 최소의 특징들을 사용해야 하고 이와 함께 분류에 필요한 정보의 손실을 최소화하여야 한다.

둘째 분류기에서는 실시간으로 공격패턴을 분류해 내기 위해 일괄처리 방식이 아닌 점증적 분류(incremental classification)가 가능해야 한다. 일괄처리 방식의 분류기는 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 논문에서는 현실세계에 적용 가능한 침입탐지 시스템을 구현하기 위해 실시간 공격패턴에 대해 점증적으로 특징 추출 및 분류를 할 수 있는 침입탐지 시스템을 제안한다. 논문의 구성은 다음과 같다. 2장에서는 침입탐지 시스템에 대한 기존의 연구를 살펴본다. 3장 에서는 기존 연구의 문제점을 살펴본 후 이를 바탕으로 4장에서는 점증적인 특징 추출 및 분류가 가능한 침입탐지 시스템을 제안한다. 마지막으로 5장에서는 결론 및 향후 연구에 대해서 이야기 한다.

II. 기존 연구

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

II-1. 특징추출에 관한 기존 연구

먼저 특징 추출을 위한 연구를 살펴보면 Andrew H. Sung[1]은 신경망과 지지벡터기계(support vector machine:SVM)를 결합한 침입탐지 시스템을 제안하였다. 이는 전체 학습 및 테스트 패턴에서 하나의 특징을 제거한 후 나머지 특징들을 바탕으로 학습을 한 후 이를 테스트 데이터에 적용하여 그 결과를 성능평가 행렬(performance metric)에 기록하는 방식으로 모든 특징들에 대해 이 과정을 반복한다. 이러한 특징 추출 기법은 특징 추출을 위해 많은 시간을 필요로 하며 점증적 방식이 아닌 일괄처리 방식의 특징 추출 기법이다. 또한 SVM이 최근 여러 분야에서 우수한 분류기로 각광을 받고 있지만 일괄처리 방식의 분류기 이므로 점증적 침입탐지 시스템에는 적합하지 못하다.

II-2. 분류기에 관한 기존 연구

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 다음과 같다. Nong Ye[2]은 침입탐지를 수행하기 위해 지도학습 방법에 의해 군집(clustering)과 분류를 할 수 있는 CCA-S 기법을 제안하였다. 이 기법은 기존의 일괄처리 학습에 비해 점증적 학습(incremental learning)이 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.

군집화 기법에 의한 연구를 살펴보면 Leonid Protnoy[3]는 KDD CUP 99 데이터 항목 중 숫자형 데이터에 대해서 군집화를 시행하였다. 각 항목별 범위를 맞추기 위해 정규화를 수행하고 이를 K-Means 방법에 의해 정상 혹은 비정상 군집으로 분류 하였다. 침입 여부는 새로이 추가되는 자료에 대해 각 군집의 중점을 기준으로 계산한 결과에 의해 판단한다. 이 기법은 분류되지 않은 데이터에 대해서 학습을 시도한 점에서는 이전의 접근과는 다른 시도 이지만 K-Means 알고리즘을 사용함으로 인해 실시간 군집화를 수행하지 못하는 단점이 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.

II-3. 기존 연구의 문제점

앞에서 열거한 선행 연구의 문제점을 정리하면 다음과 같다.

① 전처리 과정을 고려하지 않음

침입탐지 시스템의 성능을 향상시키기 위해서는 패킷에서 중요한 특징을 잘 추출하여 이를 분류기의 학습 자료로 사용하는 전처리(preprocessing) 과정이 반드시 필요하나 이제까지 대부분의 연구에서는 이 과정이 없는 경우가 많다.

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.

③ 점증적 분류를 할 수 있는 분류기 개발이 필요

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 개발이 필요하다.

III. 제안된 시스템

II장의 내용을 바탕으로 본 논문에서 제안하는 시스템은 크게 특징 추출과 분류를 수행하는 두 부분으로 구성된다. 먼저 특징 추출을 수행하는 방법은 다음과 같다. 주성분분석(Principal Component Analysis: PCA) 방법은 학습 자료의 특징추출을 위해 사용하는 대표적인 기법중의 하나이다[4]. 하지만 PCA 방법은 다음과 같은 단점이 있다. 첫째 PCA는 일괄처리(batch) 방식으로 동작한다. 이는 새로운 학습 자료가 추가 되면 고유공간(eigenspace)을 다시 계산 하여야 하는 단점이 있다. 또한 고차원 데이터를 처리하기 위해서는 많은 양의 메모리를 필요로 한다. 이러한 문제점을 해결하기 위한 기존의 연구를 살펴보면 크게 두 가지로 나눌 수 있다. 첫 번째는 일괄처리 방식의 문제점을 해결하기 위해 학습 자료를 순차적으로 받아들이며 이전의 고유공간과 새로이 추가된 학습 자료에 의해 새로운 고유공간을 계산하는 방식이다. 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 감수해야한다. 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.

다음으로 점증적 분류기에 대해 알아보면 다음과 같다. 최근의 침입탐지 연구에서 SVM은 신경망에 비해 우수한 성능을 나타내었다. 유일한 전역해를 보장하는 SVM이 많은 분야에서 성공적으로 적용되고 있으나 몇 가지 단점이 존재한다. SVM의 문제를 해결하기 위해 Suykens은 SVM의 분류기(classifier)를 구하는데 있어 SVM처럼 QP 문제를 해결하는 것이 아닌 선형방정식을 푸는 형태인 최소제곱 SVM(LS-SVM) 모델을 제안하였는데 계산상의 단순함 및 대용량 학습 데이터에도 적용할 수 있는 장점이 있다[5]. 하지만 LS-SVM은 일괄처리 방식으로 작동하므로 실시간 침입탐지 시스템에 적용하기에는 어려움이 있다. Liu[6]는 점증적 LS-SVM 모델을 제안하였다. 이 기법은 새로 추가 되는 자료에 대해 새로운 역행렬을 구할 필요 없이 이전 단계의 역행렬만으로 새로운 역행렬을 계산하는 방법에 의해 점증적 LS-SVM 모델을 개발 하였다. 이는 새로운 역행렬을 구하지 않아도 되는 계산상의 이점이 있으며 성능 면에서는 Suykens에 의해 제안된 LS-SVM과 비슷한 유용한 기법이다. 따라서 본 논문에서는 Liu에 의해 제안된 점증적 LS-SVM 기법을 침입탐지 시스템의 분류기로 사용하고자 한다. 제안된 침입참지 시스템은 다음과 같다. 먼저 전처리 과정에서는 점증적 주성분 분석 방법에 의해 점증적 특징 추출을 한다. 점증적 분류를 위해 전처리 과정에서 생성된 데이터를 점증적 LS-SVM의 입력으로 하여 분류를 시행한다. 그림 1은 제안된 시스템의 전체적인 구성도이다.

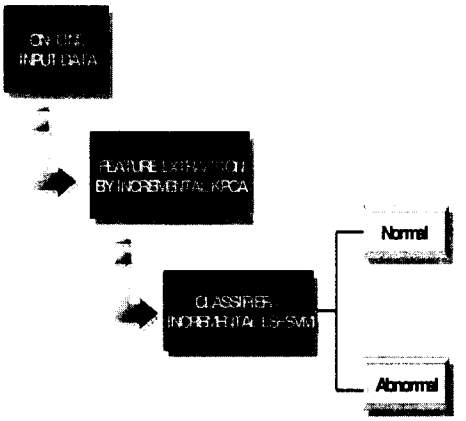

그림 1. 제안된 침입탐지 시스템
Fig. 1 Proposed intrusion detection system

IV. 실 험

제안된 방법의 성능을 비교해 보기 위해 침입탐지시스템 연구에서 대표적으로 사용하는 분류기에 대해 KDD CUP 99 자료[7]를 바탕으로 성능을 평가한다. 자료의 구성은 다음과 같다. KDD CUP 99 전체 자료중 표 1과 같은 총 1500개의 데이터를 추출하여 실험을 수행한다. 또한 자료는 전체 자료를 랜덤하게 학습과 테스트 자료로 나눈다. 시스템의 성능평가방법은 탐지율에 의해 평가하며 탐지율은 다음과 같이 구한다.

$$detection\ ratio = \frac{correct\ detection\ data}{Total\ data} \times 100$$

표 2 실험에 사용된 각 클래스의 구성
Table. 1 Data composition ratio on each class used for experiment..

클래스번호	클래스	데이터갯수
0	Normal	300
1	Dos	300
2	R2L	300
3	U2R	300
4	Probing	300

표 2는 제안된 방법에 의한 실험 결과이다.

표 2 제안된 방법에 의한 성능 평가 결과
Table. 2 Performance results by proposed method

분류성능테스트					
Average case			Best case		
클래스별탐지율		정분류율	클래스별탐지율		정분류율
0	85.24	81.30	0	88.25	86.01
1	90.33		1	90.77	
2	55.02		2	67.38	
3	80.64		3	85.4	
4	95.28		4	98.27	

표 3 Dr. Benhard에 의한 성능 평가 결과
Table. 3 Performance results by Dr. Benhard

분류성능테스트					
Average case			Best case		
클래스별탐지율		정분류율	클래스별탐지율		정분류율
0	87.31	81.41	0	96.59	88.63
1	90.71		1	93.18	
2	55.58		2	73.86	
3	80.77		3	87.50	
4	98.57		4	100	

제안된 방법의 성능 평가를 위해 그 결과를 KDD CUP 99 대회 우승자인 Dr. Bernhard의 결과 8]와 비교하였다. 표 3은 Dr. Benhard의 실험 결과이다. 그는 C5 알고리즘을 이용하여 여러 가지 공격 유형에 대하여 분류를 하였다. 그의 실험에서는 전체데이터 311,029개에 대하여 실험을 수행한 결과이며 본 논문에서는 1500개의 데이터를 사용한 결과이다. 데이터의 개수가 달라 직접적인 비교는 어려우나 전체적인 분류능력에서는 비슷한 성능을 보이고 있다. 더구나 일반적으로 실시간 방법은 일괄처리 방식에 비해 메모리 효율성이 좋고 학습 자료의 추가를 허용하는 유연한 기법이지만 일괄처리 방식에 비해 정확도가 떨어지는 것이 문제가 되는 것이 일반적인 경우인데 제안된 방법은 실시간 기법임에도 불구하고 정확도면에서 일괄처리 방식

과 비슷한 결과를 나타내어 효율성뿐만 아니라 성능 면에서도 우수한 결과를 나타내었다.

V. 결 론

본 논문에서는 실시간 침입탐지 시스템을 위한 점증적 특징 추출 기법과 분류가 가능한 실시간 침입탐지 시스템을 제안하였다. 제안된 방법을 KDD CUP 99 자료에 적용한 결과 기존의 방업에 비해 우수한 특징 추출 성능을 나타내었다. 향후 연구 과제는 실제 네트웍상에서 발생하는 침입탐지 패킷에 대해 제안된 방법을 적용하고자 한다.

참고문헌

[1] A.H. Sung, and S. Mukkamala, "Identifyi-ng Important Features for Intrusion Detection Using Support Vector Machines and Neu-ral Networks," Proceedings of the 2003 Sy-mposium on Applications and the Internet, 2003

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

[3] Leonid Portnoy, "Intrusion detection with unlabeled data using cluster " Undergraduate thesis, Columbia University

[4] I.T. Jolliffe, "Principal Component Analy-sis," New York Springer-Verlag, 1986.

[5] J.A.K. Suykens, and J. Vandewalle, "Least squares support vector machine classifiers," Neural Processing Letters, vol.9, (1999)

[6] J. Liu, J. Chen, S. Jiang and J. Cheng, "Online LS-SVM for function estimation and classification," Journal of Univ. of Science and Tech. Beijing. Vol.10, Num. 5, Oct. 2003.

[7] Accessible at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[8] Accessible at http://www-cse.ucsd.edu/users/elkan/clresults.html.

저자 소개

김병주(Byung-Joo Kim)

경북대학교 이학박사(컴퓨터과학전공)
부산대학교 이학석사(전자계산학)
부산대학교 이학사(전산통계학)
영산대학교 네트워크정보공학부 교수

※관심분야 : 기계학습, 컴퓨터보안