SVM과 의사결정트리를 이용한 혼합형 침입탐지 모델

엄 남 경[†] · 우 성 희[††] · 이 상 호[†††]

요 약

안전한 네트워크의 운영을 함에 있어 네트워크 침입 탐지에서 오탐지율을 줄이고 정탐지율을 높이는 것은 매우 중요한 일이라 할 수 있다. 최근에 얼굴 인식과 생물학 정보칩 분류 등에서 활발히 적용 연구되는 SVM을 침입탐지에 이용하면 실시간 탐지가 가능하므로 탐지율의 향상을 기대할 수 있다. 그러나 기존의 연구에서는 입력값들을 벡터공간에 나타낸 후 계산된 값을 근거로 분류하므로, 이산형의 데이터는 입력 정보로 사용할 수 없다는 단점을 가지고 있다. 따라서 이 논문에서는 의사결정트리를 SVM에 결합시킨 침입 탐지 모델을 제안하고 이에 대한 성능을 평가한 결과 기존 방식에 비해 침입 탐지율, F-P오류율, F-N오류율에 있어 각각 5.6%, 0.16%, 0.82% 향상이 있음을 보였다.

키워드 : SVM, 침입 탐지 시스템(IDS), 데이터마이닝, 의사결정트리

The Hybrid Model using SVM and Decision Tree for Intrusion Detection

Um nam-kyoung[†] · Woo sung-hee[††] · Lee sang-ho[†††]

ABSTRACT

In order to operate a secure network, it is very important for the network to raise positive detection as well as lower negative detection for reducing the damage from network intrusion. By using SVM on the intrusion detection field, we expect to improve real-time detection of intrusion data. However, due to classification based on calculating values after having expressed input data in vector space by SVM, continuous data type can not be used as any input data. Therefore, we present the hybrid model between SVM and decision tree method to make up for the weak point. Accordingly, we see that intrusion detection rate, F-P error rate, F-N error rate are improved as 5.6%, 0.16%, 0.82%, respectively.

Key Words : SVM, Data Mining, Intrusion Detection System, Decision Tree

1. 서 론

오늘날 컴퓨터와 통신 기술의 급속한 진전은 각종 정보를 공유하게 함으로써 사회 전반에 큰 변화를 가져오게 하였다. 그러나 역기능으로써 정보의 위조나 변조 또는 허락 없이 유출되는 불법 행위가 발생되는 등 폐해 또한 심각하다. 침입 방지 기술이나 침입탐지기술(Intrusion Detection System: IDS) 기술은 꾸준히 발전하고 있으며, 이진 분류 능력이 뛰어난 SVM(Support Vector Machines)을 이용한 연구가 행해지고 있다. 그러나 SVM은 입력 값을 벡터 공간에 나타낸 후 계산된 값을 근거로 분류를 수행하므로 벡터 값으로 표현이 불가능한 연속형데이터는 취급할 수 없다[1]. 따라서 이 논문에서는 기존의 SVM 기반의 침입탐지 시스템에서 입력정보로 사용하지 못했던 이산형(Discrete type)의 데이터를 의사결정트리 방법을 이용하여 추가적으로 탐지함으로써 침입 탐지율을 향상시키는 모델을 제안하고자 한다.

† 준 회 원 : 충북대학교 전자계산학과 박사수료
†† 정 회 원 : 충주대학교 멀티미디어학과 교수
††† 종신회원 : 충북대학교 전기전자컴퓨터공학부 교수

논문접수 : 2006년 9월 2일, 심사완료 : 2006년 11월 1일

2. 관련 연구

2.1 SVM을 적용한 침입 탐지

SVM은 1995년 Vladimir Vapnik에 의해 이원 패턴 인식 문제를 해결하기 위해 제안된 학습 방법으로 부정 예제로부터 긍정 예제를 분류해 낼 수 있는 결정면(Hyperplane)을 찾아내는 분류모형이다[1]. 이진 레이블을 목표 변수로 갖는 데이터의 분류작업에 있어서 매우 좋은 성능을 보이는 SVM은 명료한 이론적 근거와 뛰어난 인식성능을 바탕으로 SVM 기반의 침입탐지시스템들은 시스템에 입력되는 특징들의 수를 줄임으로써 문제를 간결하게 하며, 침입 판정 시간을 줄일 수 있고, 침입 탐지 결과의 정확성을 높일 수 있다[2]. 또한 SVM의 다중 클래스 분류자를 적용한 일대일

방법을 사용함으로써 비정상 공격을 정밀하게 탐지할 수 있으며 Multivariate Bernoulli 모델과 Multinomial 모델, One-class SVM 알고리즘을 One-class Naive Bayes에 적용하여 각각의 학습 알고리즘의 성능을 평가하였다[3-5].

2.2 의사결정트리를 적용한 침입 탐지

의사결정트리는 결과를 예측하거나 자료를 분류하고자 할 때 효과적인 방법으로 이해하기 쉬운 규칙을 형성하고, 많은 컴퓨팅 작업 없이 분류과정이 형성되며, 연속형 변수와 범주형 변수에 모두 사용가능하다는 장점을 가진다. 의사결정트리의 하나인 C4.5와 신경망 모델을 결합한 하이브리드 방식의 침입 탐지 모델은 신경망과 C4.5 알고리즘이 분류할 수 있는 공격이 다르다는 점을 이용하여 둘의 방식을 결합한 모델로 서로의 장점을 이용하면 침입 탐지율을 높일 수 있다[6].

의사결정트리는 데이터마이닝의 한 기법으로도 사용되며, 데이터마이닝은 데이터베이스에 존재하는 방대한 양의 자료로부터 사전에 알려지지 않은 암시적이고 유용한 정보를 추출하는 것으로 인공지능 분야의 기계학습 이론에 그 뿌리를 두고 있다. 알고리즘으로는 의사결정방법 이외에도 신경망, 의사결정트리, 클러스터링, 연관성 분석 등을 들 수 있다. 특히, 침입 탐지 분야에서의 데이터마이닝 기법은 프로그램과 사용자 행위를 설명하는데 필요한 특징 패턴을 추출하는 등에 사용된다. 이 과정에서 결과에 대한 유용성과 불확실성을 정량화 할 수 있게 되며, 수행 결과로 패턴이나 새로운 정보를 얻을 수 있어 의사결정트리와 클러스터링 기법 등을 이용한 침입탐지기법 등이 현재 연구되고 있다.

3. SVM과 의사결정트리를 이용한 혼합형 침입탐지 모델

3.1 제안 모델의 개요

이 논문에서 제안하는 침입 탐지 모델의 프레임워크는 (그림 1)과 같다. 제안 모델의 프레임워크는 크게 탐지 모듈과 학습 모듈로 나누어진다. 학습 모듈에서는 침입 감사데이터를 이용하여 SVM과 의사결정트리 학습이 이루어지고 탐지 모듈에서는 학습 모듈의 학습 결과를 바탕으로 침입 탐지를 수행한다.

SVM 기반의 침입 탐지 모델은 총 41개의 속성을 가지는 침입 감사 데이터 중 32개의 연속형(Continuous type)의 데이터만을 고려하였다. SVM의 특성 상 이산형 데이터는 SVM에 적용할 수 없고, 적용한다 하더라도 학습 결과에는 영향을 미치지 않기 때문이다. 연속형 데이터와 이산형 데이터는 다음과 같이 구분된다.

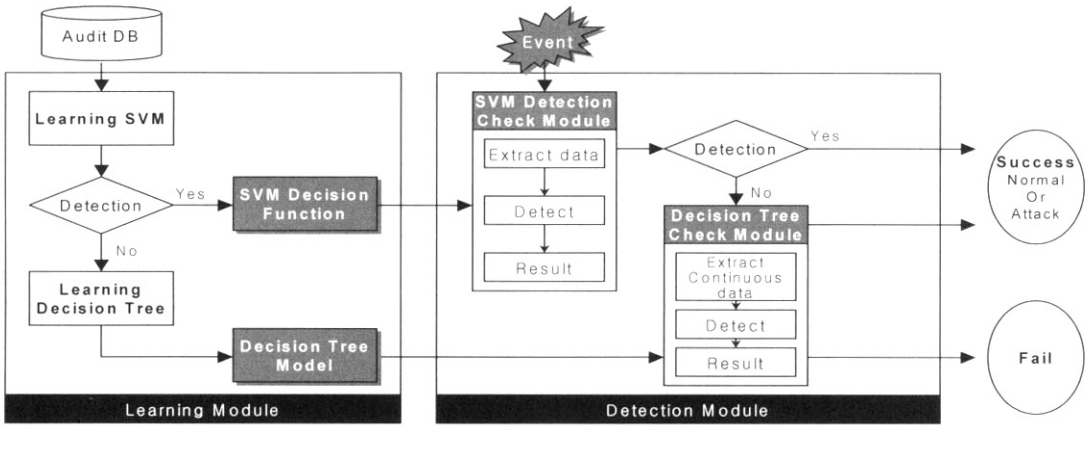
- 연속형 데이터 : 가능한 측정 결과를 셀 수 없는 경우. 즉, 어떤 임의의 구간을 택하고 그 구간 내의 하나 이상의 값이 측정 가능한 경우로 침입 모델 데이터의 속성 중 연결시간, 소스로부터의 데이터 길이, 파일 접근 횟수 등이 여기에 해당한다.
- 이산형 데이터 : 측정 결과를 셀 수 있는 경우로 침입 감사 데이터의 속성 중 프로토콜 타입, 서비스 종류, 정상/에러 플래그 등이 여기에 해당한다.

SVM을 이용한 침입 탐지 실험에 이용하지 못한 이산형 데이터들도 침입 탐지 결과에 충분히 영향을 미칠 수 있으므로 이들을 실험에서 배제하여서는 안 된다. <표 1>에서 연속형 데이터와 이산형 데이터의 예를 보인다.

이산형 데이터 중, 다음 항목들은 침입 탐지 판정에 유용한 정보를 제공한다.

- Protocol Type : TCP, UDP 등의 프로토콜 유형
- root_shell : 루트 접근을 위해 쉘(shell)의 획득 시도 여부
- su_attempted : 루트 전환 명령어 'su'의 이용 여부

따라서 SVM을 이용한 침입 탐지 실험에서 미처 탐지하지 못한 연결들은 그들의 이산형 데이터를 추출, 의사결정트리 방법을 적용하여 재 탐지 하는 방안을 제안한다. 제안 모델은 크게 학습 모듈과 탐지 모듈로 구성되어 있다. 학습 모듈은 침입 감사 데이터를 SVM과 의사결정트리에 적



(그림 1) 제안 프레임워크

〈표 1〉 연속형/이산형 데이터의 예

데이터 타입	데이터 속성	특징 설명
연속형	duration	연결시간
	src_bytes	소스로부터의 데이터길이
	urgent	Urgent 패킷 개수
	hot	"Hot" indicator 개수
	num_root	root 접근 회수
	num_access_files	파일 접근 회수
	num_failed_logins	로그 실패 회수
이산형	protocol type	프로토콜 타입(TCP, UDP등)
	service	서비스 종류(HTTP, FTP등)
	flag	정상 또는 에러 플래그
	land	1:같은 소스/목적지 주소, 0
	logged_in	로그 성공/실패 여부
	root_shell	1:root shell 획득, 0
	su_attampted	1:"su root"명령 시도, 0

용시켜 학습 모델과 결정함수를 생성한다. 탐지 모듈은 침입탐지시스템으로부터 수집된 이벤트들을 SVM 결정함수와 의사결정트리 모델로 탐지 실험을 하여 침입을 판정한다.

3.2 학습 모듈

(1) SVM 학습 모듈

SVM 학습은 침입과 정상을 구분할 수 있는 서포터 벡터와 가중치 벡터 값으로 이루어지는 결정함수를 구하는 과정이다. 학습과정을 통해 입력 벡터 값에 따라 고차원 공간에 침입과 정상을 구분할 수 있는 최대 마진을 가지는 결정면을 가진다.

① 침입 감사 데이터 셋으로 부터 학습을 위한 데이터 셋 추출
② 추출된 데이터 셋을 SVM 머신의 입력 포맷에 맞게 변환
③ 커널을 사용하여 SVM학습
④ 학습 후 결정함수 생성

(2) 의사결정트리 학습 모듈

의사결정트리는 많은 컴퓨팅 작업 없이 분류과정을 형성하며 이산형 변수와 연속형 변수에 모두 사용할 수 있다. 때문에 SVM 학습 결과 탐지하지 못한 데이터의 이산형 데이터 부분만을 추출하여 의사결정트리 방법을 적용한다. 의사결정트리의 학습 과정은 다음과 같다.

① SVM 탐지 결과를 바탕으로 학습을 위한 데이터 셋 추출
② 추출된 데이터 셋을 의사결정 트리 알고리즘을 사용하여 학습
③ 학습 후 의사결정트리 모델 생성

3.3. 탐지 모듈

(1) SVM 탐지 모듈

SVM 탐지 모듈은 SVM 학습을 통해 생성된 결정함수에 침입 감사 데이터를 적용하여 침입 여부를 판정하는 모듈로 탐지 과정은 다음과 같다.

① 실험 데이터 셋을 SVM 입력 포맷에 맞게 변환
② 실험 데이터 셋을 SVM 학습으로 얻은 결정함수에 적용
③ 침입인지 정상인지 판정

(2) 의사결정트리 탐지 모듈

의사결정트리 탐지 모듈은 의사결정트리 학습을 통해 생성된 모델에 SVM 탐지 모듈에서 탐지하지 못한 데이터만을 적용하여 침입을 판정하는 모듈로 탐지 과정은 다음과 같다.

① SVM 탐지 모듈에서 미 탐지된 데이터 셋을 추출
② 추출된 데이터 셋을 의사결정트리 학습으로 얻은 의사결정트리 모델에 적용
③ 침입인지 정상인지 판정

4. 실험 및 평가

4.1 실험 환경

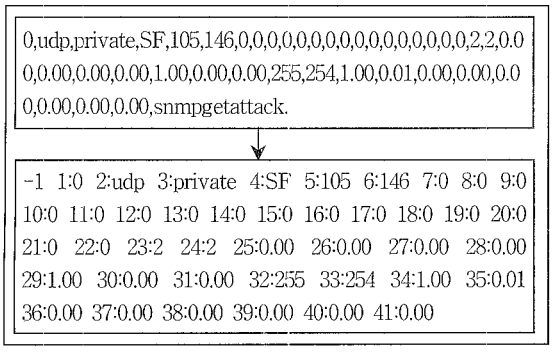
제안모델은 SVMChen2.0과 Clementine7.0을 이용하여 Windows 2000Pro에서 실험되었으며 실험에 사용한 침입 감사 데이터는 KDD Cup 99 데이터 셋이며, 각각의 연결기록은 41개의 독립적인 속성과 공격 유형 레이블로 이루어져 있으며 상세한 데이터 구성은 <표 2>와 같다.

침입탐지시스템은 탐지대상으로부터 생성되는 시스템 사용 내역, 컴퓨터 통신에 사용되는 패킷 등과 같은 데이터를 수집하는 침입 감사 데이터(Audit Data) 수집 과정을 거친다. 이후 수집 된 침입 감사 데이터는 침입 판정이 가능할 수 있도록 데이터 가공 및 축약(Data Reduction and Filtering) 과정을 거쳐 의미 있는 정보로 전환 된다. 이렇게 생성된 침입 감사 데이터는 SVM에 입력되기 전에 SVM 머신의 표준 입력 형식에 합당하도록 포맷을 변환하여야 한다. SVM 학습과 실험을 위한 데이터 변환의 예는 (그림 2)와 같다.

데이터 포맷 변환은 각 연결들의 앞부분에 정상일 경우 '1', 공격일 경우 '-1'을 레이블링한 뒤, 41개의 속성마다 'No: 속성 값'과 같이 넘버링 함으로써 얻어진다.

〈표 2〉 KDD Cup 99 데이터 셋의 구성

항 목		개 수
총 데이터 수		311,029
총 속성의 수	이산형	9
	연속형	32
	공격 유형 레이블	1
	합계	42
공격유형 클래스		4
총 공격 유형		38



(그림 2) SVM 머신을 위한 데이터 변환의 예

4.2 성능 평가 기준

제안 방법의 성능을 평가하기 위한 항목으로 탐지율과 False Positive 오판율, False Negative 오판율을 사용하며 계산방법은 다음과 같다.

$$\text{탐지율} = \frac{\text{시스템에 의해 침입으로 판정된 침입 데이터의 개수}}{\text{전체 침입 데이터 개수}} \times 100 \quad \text{식-①}$$

$$\text{F-P오류율} = \frac{\text{시스템에 의해 침입으로 오판된 정상 데이터의 개수}}{\text{전체 정상 데이터 개수}} \times 100 \quad \text{식-②}$$

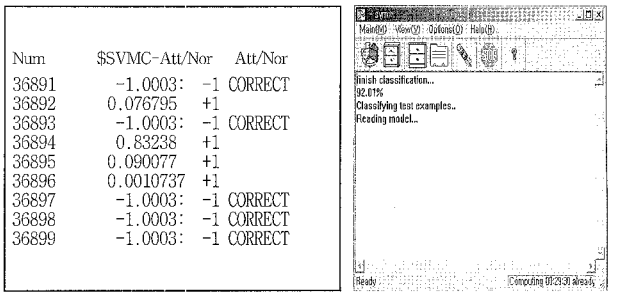
$$\text{F-N오류율} = \frac{\text{시스템에 의해 정상으로 오판된 침입 데이터의 개수}}{\text{전체 침입 데이터 개수}} \times 100 \quad \text{식-③}$$

식-①은 탐지율로써 전체 침입 데이터 중 시스템에 의해 침입으로 정확히 판정된 데이터의 비율을 백분율로 나타낸 값이다. 식-②는 펄스 포지티브(False Positive)에 대한 오판율로 전체 정상 데이터 중 시스템에 의해 침입으로 오 판정된 데이터의 비율을 백분율로 나타낸 값이며, 식-③은 펄스 네거티브(False Negative) 오판율로 전체 침입 데이터 중 시스템에 의해 정상으로 오 판정된 데이터의 비율을 백분율로 나타낸 값이다.

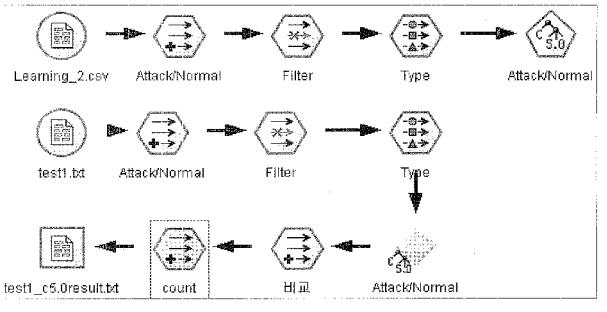
4.3 실 험

(1) 기존 방식 실험

기존 방식 실험은 SVM 학습을 통해 생성된 결정함수에 침입 감사 데이터를 적용하여 침입 여부를 판단하는 과정이다. 기존 방식의 학습은 학습용 데이터 셋과 학습에 사용되는 내부 커널함수, 정규화 매개변수인 C값에 의존하며, 다음과 같은 과정을 거친다. KDD Cup 99 데이터 셋의 연결들 중 6만 건을 비례추출 하여 정상(1)/공격(-1) 으로 레이블링 한 후 $e^{-}|x-y|^2/2\sigma^2$ 으로 나타내어지는 RBF커널(C=0)을 사용하여 학습시킴으로써 이진 분류를 수행할 수 있는 결정함수를 얻는다. 기존방식 학습 후 생성된 결정함수를 이용하여 침입 탐지 실험을 수행하기 위한 실험 데이터는 침입



(그림 3) 기존 방식의 침입 탐지 과정과 결과



(그림 4) 제안 모델의 침입 탐지 과정

KDD Cup 99 데이터 셋 중 10만 건을 랜덤 추출하여 생성하였다. SVM 탐지 실험 과정에서는 30만 건의 침입 감사 데이터를 모두 커버하기 위해 총 3번에 걸쳐 실험 데이터 셋을 추출하여 SVM 탐지 실험을 진행하였다. 추출 된 3건의 실험 데이터를 이용한 탐지 과정과 결과의 일부가 (그림 3)에 나타나 있다.

기존 방식의 침입 탐지 과정에서는 학습 데이터를 추출하여 SVM 결정함수에 적용시켜 탐지를 수행하며 그 결과 생성된 결과 파일에는 탐지 여부가 기록되게 된다.

(2) 제안 방식 실험

제안 방식 실험은 의사결정트리 학습을 통해 생성된 모델에 침입 감사 데이터를 적용하여 침입 여부를 판단하는 과정이다. 기존 방식 실험 후, 탐지하지 못한 약 7천 건의 연결 데이터를 추출하여 실험 데이터로 사용한다. (그림 4)에서는 실험 데이터를 제안 모델에 적용하여 침입을 탐지하는 과정은 나타낸다. 즉, 제안 방식의 침입 탐지 실험은 실험 데이터를 추출하여 제안 모델에 적용시켜 탐지 결과를 파일로 저장하게 된다. 텍스트 파일로 저장된 탐지 결과에는 탐지 여부와 정탐지 횟수를 카운트 한 내용이 기록된다.

4.4. 평가

(1) 기존 방식 실험 결과

SVM 학습 후 생성된 결정함수로 기존 방식 실험을 수행한다. 이때 실험 데이터는 KDD CUP 99 데이터로부터 랜덤하게 추출한 10만개의 연결 데이터를 사용하였으며 총 3번에 걸쳐 실험을 반복하였다. 실험 결과는 <표 3>에 기술하였다.

〈표 3〉 기존 모델의 탐지 실험 결과

평가척도 \ 실험	실험 1	실험 2	실험 3	평균
학습 시간	45분 45초			45분45초
탐지소요시간	29분 30초	31분 19초	31분 6초	30분 39초
탐지율	92.01%	92.90%	92.89%	92.60%
F-P오류율	3.03%	3.07%	3.01%	3.03%
F-N오류율	9.02%	9.09%	9.09%	9.06%

〈표 4〉 제안 모델의 탐지 실험 결과

평가척도 \ 실험	실험 1	실험 2	실험 3	평균
학습 시간	45분 47초			45분 47초
탐지소요시간	29분 39초	31분 28초	31분 16초	30분 39초
탐지율	97.76%	98.35%	98.78%	98.20%
F-P오류율	2.88%	2.93%	2.81%	2.87%
F-N오류율	8.29%	8.21%	8.22%	8.24%

SVM을 이용한 침입 탐지의 경우 평균 탐지율은 약 92.60%로 나타났다. 실험 데이터 100건당 학습에 소요된 시간은 2.745초이며, 탐지에 소요 된 시간은 1.839초이다.

(2) 제안 방식 실험 결과

의사결정트리 학습 후 생성된 결정함수로 제안 방식 실험을 수행한다. 이때 실험 기존 방식 실험을 통해 탐지 하지 못한 데이터를 추출하여 사용하였으며 총 3번에 걸쳐 실험을 반복하였다. 실험 결과는 <표 4>에 기술하였다.

제안 방식을 이용한 침입 탐지의 경우 평균 탐지율은 약 98.20%로 나타났다. 실험 데이터 100건당 학습에 소요된 시간은2.747초, 탐지에 소요된 시간은 1.848초이다.

(3) 비교 평가

<표 5>는 기존 방법과 제안 방법의 실험 결과이며, (그림 5)는 SVM과 제안방식을 비교한 그래프이다.

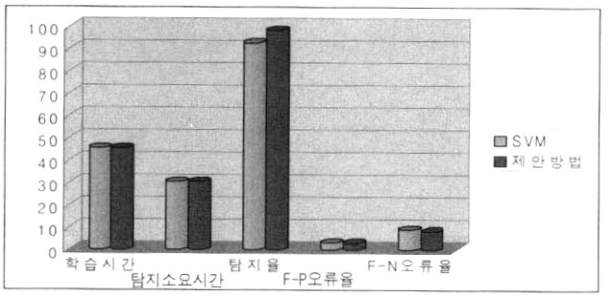
실험 데이터 100건당 학습에 소요된 시간은 2.747초, 탐지에 소요된 시간은 1.848초이며 전체 학습에 추가로 소요된 시간은 2초, 탐지 실험에 추가로 소요된 시간은 9초로 제안 방법에 추가로 소요된 시간은 전체 학습, 탐지시간에 미치는 영향이 미미하다.

5. 결 론

기존의 SVM을 이용한 침입탐지시스템에서는 SVM의 입력 정보로 연속형 데이터만을 고려하여 학습과 실험을 수행하였다. 그러나 SVM에 입력으로 사용될 수 없었던 이산형 데이터는 침입 판정에 상당한 영향을 미치는 중요한 정보들

〈표 5〉 제안 방법과 기존 방법의 성능 비교

방법 \ 평가기준	학습 소요시간	탐지 소요시간	탐지율	F-P 오류율	F-N 오류율
SVM	45분 45초	30분 39초	92.60%	3.03%	9.06%
제안방법	45분 47초	30분 48초	98.20%	2.87%	8.24%



(그림 5) SVM과 제안방식 비교

을 포함하고 있다. 따라서 이 논문에서는 침입을 탐지하는데 있어 분류능력이 탁월한 SVM과 이산형 데이터를 입력정보로 사용하여 동작이 가능한 데이터마이닝의 기법 중의 의사결정트리를 결합하여 침입을 탐지하는 모델을 제안하고 실험 하였다. 이 과정에서 SVM의 입력정보로 사용할 수 없는 이산형 데이터를 침입 탐지에 반영하기 위해 의사결정트리의 C5.0 알고리즘을 적용해 보았다. 실험 결과 기존 방법보다 침입 탐지 소요 시간이 총 11초 더 요구되었으나 5.6%의 탐지율 향상이 있었음을 알 수 있었다. 또한 F-P오류율, F-N오류율도 각각 0.16%, 0.82%향상을 보임으로써 침입 탐지에 제안 방법이 효율적임을 보였다.

향후 연구로는 학습 시간과 탐지 시간을 줄이기 위해 탐지에 적합하면서도 많은 시간을 요구하지 않도록 학습 데이터양을 조절하는 연구가 필요하며, 실시간 탐지를 만족시키기 위한 구현 및 연구, 실험도 병행되어야 하겠다.

참 고 문 헌

[1] N. Cristianini an, J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods", Cambridge University Press, 2000.

[2] Sung, A.H. Mukkamala, S. "Identifying important features for intrusion detection using support vector machines and neural networks", Proceedings of Symposium on Applications and the Internet, Jan., 2003.

[3] Ke Wang, Salvatore J. Stolfo, "One-Class Training for Masquerade Detection", CU Tech Report, April, 2003.

[4] Leslie, Christina, Eleazar Eskin and William Stafford Noble, "The spectrum kernel: A string kernel for SVM protein classification", Proceedings of the Pacific Symposium on Biocomputing, pp.564-575. Jan., 2002.

[5] Srinivas Mukkamala, Guadalupe Janoski, Andrew Sung, "Multi class support vector machine implementation to intrusion detection", Proceedings of the 2002 International Joint Conference, Vol.2, May, 2002.

[6] Zhi-Song Pan, Song-Can Chen, Gen-Bao Hu, Dao-Qiang Zhang, "Hybrid neural network and C4.5 for misuse detection", Proceedings of Machine Learning and Cybernetics International Conference, Vol.4, pp.2463-2467, Nov., 2003.

우 성 희

e-mail : shwoo@chungjoo.ac.kr

1990년 2월 청주대학교 전자계산학과 졸업

1993년 2월 충북대학교 전자계산학과 석사

1999년 2월 충북대학교 전자계산학과 박사

1995년~현재 : 충주대학교 멀티미디어학과 교수

관심분야 : 프로토콜공학, 데이터통신, 컴퓨터네트워크, 네트워크보안

엄 남 경

e-mail : family@netsec.cbnu.ac.kr

1999년 2월 충북대학교 컴퓨터과학과 졸업

2002년 2월 충북대학교 전자계산학과 석사

2004년 2월 충북대학교 전자계산학과 박사수료

관심분야 : 유비쿼터스 네트워크, 네트워크 보안, 침입탐지 시스템, 프로토콜 테스팅

이 상 호

e-mail : shlee@cbnu.ac.kr

1976년 2월 숭실대학교 전자계산학과 졸업

1981년 2월 숭실대학교 전자계산학과 석사

1989년 2월 숭실대학교 전자계산학과 박사

1981년 6월~현재 : 충북대학교 전기전자컴퓨터공학부 교수

관심분야 : 통신 프로토콜 공학, 네트워크 관리, 네트워크 보안