Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.2, Apr. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.2.267

앙상블 학습의 부스팅 방법을 이용한 악의적인 내부자 탐지 기법

박 수 연†‡
고려대학교 정보보호대학원 (대학원생)

Malicious Insider Detection Using Boosting Ensemble Methods

Suyun Park†‡
School of Cybersecurity, Korea University (Graduate student)

요 약

최근 클라우드 및 원격 근무 환경의 비중이 증가함에 따라 다양한 정보보안 사고들이 발생하고 있다. 조직의 내부자가 원격 접속으로 기밀 자료에 접근하여 유출을 시도하는 사례가 발생하는 등 내부자 위협이 주요 이슈로 떠오르게 되었다. 이에 따라 내부자 위협을 탐지하기 위해 기계학습 기반의 방법들이 제안되고 있다. 하지만, 기존의 내부자 위협을 탐지하는 기계학습 기반의 방법들은 편향 및 분산 문제와 같이 예측 정확도와 관련된 중요한 요소를 고려하지 않았으며 이에 따라 제한된 성능을 보인다는 한계가 있다. 본 논문에서는 편향 및 분산을 고려하는 부스팅 유형의 앙상블 학습 알고리즘들을 사용하여 악의적인 내부자 탐지 성능을 확인하고 이에 대한 면밀한 분석을 수행하며, 데이터셋의 불균형까지도 고려하여 최종 결과를 판단한다. 앙상블 학습을 이용한 실험을 통해 기존의 단일 학습 모델에 기반한 방법에서 나아가, 편향-분산 트레이드오프를 함께 고려하며 유사하거나 보다 높은 정확도를 달성함을 보인다. 실험 결과에 따르면 배깅과 부스팅 방법을 사용한 앙상블 학습은 98% 이상의 정확도를 보였고, 이는 사용된 단일 학습 모델의 평균 정확도와 비교하면 악의적인 내부자 탐지 성능을 5.62% 향상시킨다.

ABSTRACT

Due to the increasing proportion of cloud and remote working environments, various information security incidents are occurring. Insider threats have emerged as a major issue, with cases in which corporate insiders attempting to leak confidential data by accessing it remotely. In response, insider threat detection approaches based on machine learning have been developed. However, existing machine learning methods used to detect insider threats do not take biases and variances into account, which leads to limited performance. In this paper, boosting-type ensemble learning algorithms are applied to verify the performance of malicious insider detection, conduct a close analysis, and even consider the imbalance in datasets to determine the final result. Through experiments, we show that using ensemble learning achieves similar or higher accuracy to other existing malicious insider detection approaches while considering bias-variance tradeoff. The experimental results show that ensemble learning using bagging and boosting methods reached an accuracy of over 98%, which improves malicious insider detection performance by 5.62% compared to the average accuracy of single learning models used.

Keywords: Network intrusion detection, Ensemble learning, Malicious insider, Insider threat detection, Machine learning

Received(02. 22. 2022), Modified(03. 28. 2022), Accepted(03. 28. 2022)

† 주저자, sypark55@korea.ac.kr
‡ 교신저자, sypark55@korea.ac.kr(Corresponding author)

I. 서 론

최근 내부자 위협으로 인해 데이터 침해, 기술 유출, 영업 비밀/지적 재산권 도난 등 다양한 형태로 피해가 나타나고 있다. 예를 들어, 민감한 자료를 인가되지 않은 응용프로그램에 업로드하거나 클라우드 파일 공유 앱에 기밀 정보를 복사하는 위협들이 있다. Gurucul에서 발간한 2021년 사이버 보안 보고서[1]에 따르면 응답자의 50%가 외부 사이버 공격보다 내부자 공격이 탐지 및 예방이 더 어렵다고 하였다. 내부자 공격은 조직의 정보 시스템에 대한 접근 권한이 있는 사용자에 의해 수행되는 악의적인 행위이기 때문이다. 내부자에 의한 보안 사고를 감소시키기 위해서 악의적인 내부자 탐지의 필요성은 커지고 있으며, 내부자 위협에 능동적으로 대응하는 것이 중요해지고 있다.

악의적인 내부자 위협을 방어하고 있는 조직 중 인공지능과 기계학습을 활용하고 있는 조직이 26%로 가장 많고 빅데이터 분석이 뒤를 잇는다[1]. 기존 연구들 또한 시스템의 일반적인 동작을 학습하고, 이에서 벗어나는 동작이 발생할 시 탐지하는 기계학습 기반 탐지 기법을 주로 제안하였다. 이 기법은 내부자의 다양한 악의적인 행위를 자동으로 식별하는 모델을 만들고 특징을 학습한 알고리즘을 지속해서 업데이트하기 때문에 탐지 성능이 정확하고 안정적이다. 하지만, 사용된 데이터셋(dataset) 내의 데이터 분포가 매우 불균형하거나 과소적합(underfitting) 및 과대적합(overfitting)이 발생할 때 기계학습 기법을 사용한 탐지의 결과는 편향될 수 있고 정확도가 낮아질 수 있다. 과소적합의 경우 데이터를 제대로 학습하지 못해 모델의 성능이 낮아지며, 과대적합의 경우 지나치게 학습 데이터에 맞춘 모델을 생성하여 모델의 일반화 성능이 낮아진다.

이에 본 연구에서는 부스팅 유형의 앙상블 알고리즘이 악의적인 내부자 탐지에 미치는 영향을 비교하고 편향과 분산의 균형을 고려하여 악의적인 내부자를 탐지한다. 악의적인 내부자 탐지 오류를 감소하기 위한 배깅 및 부스팅 알고리즘들의 성능을 비교하며, 분류할 클래스들이 균일하게 분포되지 않은 비대칭 데이터셋에 서로 다른 세 가지 샘플링(sampling) 기법을 적용하여 탐지 성능을 비교하고 면밀히 분석한다. 그 결과, 악의적인 내부자를 탐지하기 위해 사용한 앙상블 학습 알고리즘 중 정확도가 높고 속도가 빠른 모델을 확인한다. 다양한 조건을 설정하여 실험을 수행한 결과, 악의적인 내부자 탐지 문제에서 앙상블 학습을 사용하였을 때, 과적합 감소 효과를 통해 탐지 정확도가 향상될 수 있음을 보이며 그와 동시에 편향-분산 트레이드오프 역시 함께 고려하여 기존 연구들과 유사한 정확도를 달성한다.

본 논문의 구성은 다음과 같다. 2장에서는 내부자 위협과 앙상블 학습을 기반으로 한 네트워크 침입 탐지 관련 연구들을 살펴본다. 3장에서는 악의적인 내부자 탐지를 위한 앙상블 기법과 사용된 데이터에 대해 설명한다. 4장에서는 분석 결과들을 기술하고, 마지막 5장에서는 향후 연구에 대한 논의와 함께 결론을 맺는다.

II. 관련 연구

본 연구와 관련된 기존 연구를 기계학습을 이용한 악의적인 내부자 탐지 기술과 앙상블 학습 기반 네트워크 침입 탐지, 크게 두 분류로 나누어 살펴본다.

2.1 기계학습 기반 내부자 탐지

악의적인 내부자는 내부자에 대한 조직의 신뢰와 기밀 자료에 접근할 수 있는 권한 때문에 탐지하고 예측하는 것이 어렵다. 최근 몇 년 동안 악의적인 내부자 공격수가 증가함에 따라 악의적인 내부자 문제를 해결하기 위한 여러 연구가 발표되었다. 내부자 탐지 분야에서는 기계학습 학습 방법이 많이 이용되었으며, 제안된 연구들의 대부분은 이상 탐지(anomaly detection)를 기반으로 한다. 기계학습 기반 탐지 기술에서 사용자의 일반 프로파일은 정상 동작을 기반으로 작성되는데, 이상 탐지는 이러한 정상 동작에서 일탈이 발생할 시 이를 식별하는 방법이다. 악의적인 내부자 탐지에 대한 기계학습 기반의 연구로는 비지도 학습 기반 접근법[2], 심층 학습 기반 접근법[3, 5], 자연어처리 기반 접근법[4] 등이 있다.

악의적인 내부자 탐지를 하는 방법으로 [2]에서는 서로 다른 작동 원리를 가진 네 가지 비지도 학습 방법을 사용한다. [3]에서는 시스템 로그를 조직화된 순서로 모델링하기 위해 LSTM Autoencoder를 포함하는 심층 학습 모델을 설계하여 내부자 위협을 탐지하며, [5]에서는 사이버 보안 분석가를 지원하기 위해 다양한 수준의 데이터 세분성(data granularity) 및 훈련 시나리오에 대해 서로 다른

네 가지 기계학습 알고리즘들로 악의적인 내부자 행동을 탐지한다. [4]에서는 주로 음성 인식 및 텍스트 분석에 적용되어 온 Levenshtein distance 측정 기법을 활용하여 IoT 환경 내 악의적인 내부자 공격을 탐지한다.

편향 및 분산 문제와 같은 중요한 요소를 고려하지 않고 내부자 위협을 탐지하는 데 초점을 맞췄던 기존 접근 방식과 달리, 본 연구는 이 두 가지 요소를 고려하여 악의적인 내부자 탐지 시 보다 향상된 성능을 달성할 수 있는 앙상블 학습 방법을 사용한다. 특히 배깅 및 부스팅 알고리즘들의 탐지 성능을 비교하고 성능 향상을 위해 샘플링 기법을 적용하여 면밀히 분석할 예정이다.

2.2 앙상블 학습 기반 네트워크 침입 탐지

앙상블 학습은 기계 학습의 메커니즘 중 하나로 일반적으로 여러 모델을 결합함으로써 평균 예측 성능을 향상시킨다. 기존의 단일 모델을 사용하는 것보다 앙상블 학습을 사용할 때 성능이 향상되는 이유는 모델에 의해 발생하는 예측 오류의 감소 때문이다.

[6]에서는 지능형 지속 공격(advanced persistent threat)의 대상이 되는 취약한 네트워크의 호스트를 여러 기계학습 분류기를 통해 탐지하였다. [7]에서는 인증 로그의 신뢰 여부를 분류하기 위해 다수결 투표(majority voting)를 사용하여 거짓 양성(false positive)과 거짓 음성 값(false negative)이 모두 0임을 보였다. [8]에서는 다양한 유형의 트래픽에 대해 XGBoost와 심층 신경망(deep neural network)을 통합하여 결과를 예측하였다. 또한, [9]에서는 하이브리드 특징 선택(hybrid feature selection)과 2단계 메타 분류기(two-stage meta classifier)의 조합을 기반으로 하는 이상 징후 기반 침입 탐지 시스템을 제안하였다.

위와 같이 기존 연구에서는 외부자의 네트워크 침입 탐지 시 앙상블 학습 방법의 적용을 통해 정확도를 향상한 사례는 다수 존재하나, 외부자가 아닌 악의적인 내부자 탐지 시 이를 적용하는 방안에 대해서는 아직 연구가 부족하다. 본 연구는 앙상블 학습을 사용하여 최종적으로 악의적인 내부자를 판별하고 정확도를 높이고자 한다.

III. 악의적인 내부자 탐지를 위한 앙상블 학습

Fig.1.은 앙상블 학습을 기반으로 악의적인 행동과 내부자 탐지를 하는 기법의 기본 동작을 나타낸다. 악의적인 내부자를 탐지하는 방법은 다음의 단계를 걸쳐 동작하게 되는데, 처음 사용자 데이터가 다양한 CSV(Comma-Separated Variables) 파일에서 수집되면 데이터가 집계, 인코딩 등 전처리 되고 로그온/로그오프 이벤트, 파일 접근 등을 포함하는 특징이 추출되며, 선택한 특징들은 모델 학습 및 평가에 사용된다. 이후 보팅, 배깅, 및 부스팅 알고리즘을 적용하여 오류를 최소화하면서 정확도가 높은 탐지 모델을 확인한다.

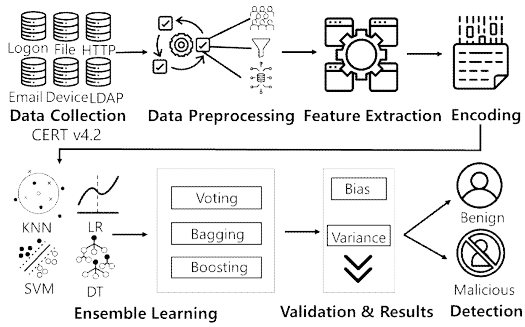

Fig. 1. Process of malicious insider detection using ensemble learning

3.1 데이터 수집

본 연구에 사용된 데이터셋은 CMU CERT 내부자 위협 데이터셋[10] r4.2 버전이다. CERT 데이터는 Carnegie Mellon University의 CERT 프로그램에서 제공하는 내부 위협 연구 자료이며, 가상 조직에서 근무하는 내부자의 데이터를 포함하고 있다. 이 데이터셋은 일반 내부자와 악의적인 내부자를 포함하며 총 1000명의 사용자 중 70명이 악의적인 내부자이다. 현재 버전 r1부터 r6까지 접근할 수 있으며, 우리는 실험에 데이터셋 내 각 시나리오별 인스턴스가 1개씩 존재하는 r4.2 버전을 활용하였다.

본 논문에서 사용한 CERT 데이터셋의 각 파일의 필드(field)는 Table 1.과 같다. CERT r4.2 데이터셋은 7개의 로그 파일로 나뉘고, 이 로그 파일은 18개월에 걸친 사용자 활동을 기록한다. CERT 데

Table 1. CMU CERT r.4.2 dataset fields

File	Fields
logon.csv	id, date, user, pc, activity (logon/logoff)
file.csv	id, date, user, pc, filename, content
http.csv	id, date, user, pc, url, content
email.csv	id, date, user, pc, to, cc, bcc, from, size, attachment_count, content
device.csv	id, date, user, pc, activity (connect/disconnect)
LDAP	employee_name, user_id, email, role, business_unit, functional_unit, department, team, supervisor
psychometric .csv	employee_name, user_id, O, C, E, A, N

이터셋은 악의적인 내부자에 대해 시나리오 세 가지를 포함하고 있으며, 내부자 모두 조직의 현재 직원이다. 첫 번째 시나리오에서는 직원이 근무 시간이 끝난 후 조직의 시스템에 로그인하여 조직의 시스템에서 중요한 데이터를 이동식 디스크에 수집 및 저장한다. 두 번째 시나리오에서는 직원이 구인 사이트를 탐색하고 다른 조직에 고용을 요청함으로써 새로운 일자리를 찾기 시작한다. 마지막으로, 세 번째 시나리오에서는 직원이 키로거(keylogger)를 다운로드하고 이동식 디스크를 사용하여 상사의 컴퓨터에 전송한 후 상사로 로그인하여 조직에 많은 양의 이메일을 보낸 후에 곧바로 조직을 떠난다.

3.2 데이터 전처리

데이터셋의 각 파일에는 미가공 데이터가 포함되어 있어 기계학습 알고리즘에 바로 입력될 수 없고 전처리 과정을 필요로 한다. 웹 액세스, 이메일 및 파일 액세스 로그와 같은 조직의 일반적인 모니터링 데이터를 가정하였고, 데이터 전처리 단계에서는 이를 집계하여 사용한다.

한편, 데이터가 실수형과 같은 연속적인 수치가 아닌 범주형 값에 해당할 수 있으므로 이에 대한 처리 역시 필요하다. Scikit-learn에서 제공하는 기계학습 알고리즘의 모든 입출력 변수는 정수형으로 변환되어야 한다[11]. 따라서 본 논문에서는 모든 사용자와 할당된 역할을 설명하는 파일 집합인 LDAP 파일의 역할(role)들에 레이블 인코딩을 적용한다.

3.3 특징 추출

내부자 시나리오에 따라 사용자 동작을 효율적으로 학습하고 예측할 수 있는 관련 데이터 필드를 식별한다. 특징 추출에 psychometric.csv는 사용되지 않는다. 이 범주는 주로 사용자의 심리측정과 같은 보다 복잡한 데이터로 구성되기 때문이다. 모든 CSV 파일의 필드에는 날짜와 시간, PC, user_id가 공통으로 있으며 총 21개의 특징을 추출하였다. 총 21개의 특징을 추출한 이유는 시나리오에 따른 새로운 미가공 데이터를 조합하는 것이 필요하기 때문이다. 특징들의 개수가 충분하지 못할 경우, 모델이 단순해져 정확도가 낮아질 수 있고, 지나치게 많은 특징을 선택할 경우 과적합을 초래할 수 있다. 우리는 추출 특징 개수와 조합을 번갈아 수행하며 최종적으로 본 CERT 데이터셋에 대해 가장 높은 성능

Table 2. Extracted list of features

Features	Working hours(07-19)
logon.csv	(1)Number of logons
	(2)Number of logoffs
	(3)Logon after working hours
	(4)Total number of PCs accessed
file.csv	(1)File access status
	(2)File access after working hours
http.csv	(1)Number of visited websites
	(2)Visting a job searching site
	(3)Number of wikileaks.org visits
	(4)Downloads a keylogger
	(5)Visiting sites after working hours
email.csv	(1)Number of total emails sent
	(2)Number of internal emails
	(3)Number of external emails
	(4)Emails with attachments
device.csv	(1)Connection after working hours
	(2)Number of connections in the month before quitting
	(3)Markedly higher rates than their previous activity
LDAP	(1)Role
	(2)Month of service
	(3)Whether quitting

을 달성하는 것으로 확인된 21개의 특징 추출을 적용한다. 추출된 특징에 대한 전체 목록은 Table 2.에 나타나 있다. 예를 들어, 사용자의 정상 근무 시간은 오전 7시부터 오후 7시이며 이 시간 동안 사용자가 로그인/로그아웃하면 정상이지만 사용자가 업무 시간 이후에 로그인하여 이동식 디스크를 사용하고 잠시 후에 나가는 경우는 악의적인 행동으로 간주한다. 파일별로 사용자가 수행한 이메일 전송 수, 업무 시간 이후 파일 접근 수, 또는 PC에서 방문한 웹 사이트 수를 추출한다. 데이터의 평균, 표준 편차 등과 같은 통계적 특징으로 요약되는 데이터의 예로는 이메일 첨부 파일 수, 퇴사 직전 달의 외부 장치 연결 횟수 등이 있다.

3.4 앙상블 학습

이상적인 학습 모델은 낮은 예측 오류율과 높은 정확도를 달성하는 모델이다. 오류는 학습된 모델을 사용하여 추론할 시, 모델의 예측값과 실제 값의 차이이다. 기계학습 기법을 사용하여 목표 변수를 예측하려고 할 때 이러한 오류가 발생하는 주요 원인에 해당하는 것이 편향 및 분산이다. 기계학습 알고리즘의 예측 오류는 편향, 분산 및 잡음(noise)으로 이루어져 있는데 이처럼 3요소로 분해하는 것을 편향-분산 분해(bias-variance decomposition)라 한다[12]. 편향에 따른 오차는 모델의 기대 예측과 예측해야 할 실제 값(ground truth) 사이의 차이고 분산에 따른 오차는 모델의 예측값이 평균으로부터 흩어진 정도이다. 마지막으로, 잡음은 어떤 기계학습 모델로도 줄일 수 없는 오류이다. 따라서 잡음 요인은 종종 무시되고 일반적으로 이 분해 모형에서는 편향과 분산만을 고려한다. 편향-분산 분해에 대한 구체적인 여러 가지 방법 중에 우리는 앙상블 모델이 오류를 줄이는 이유를 설명하고 모든 손실 함수(loss function)에 적용할 수 있는 편향과 분산의 통일된 정의를 제안한 Domingos의 모형을 따른다[12]. 해당 분해 모형에 따르면 어떠한 모델의 예측 오류는 다음과 같이 편향의 제곱, 분산, 그리고 줄일 수 없는 잡음의 합으로 표현할 수 있다. 예측 오류, 편향 및 분산을 수식으로 표현하면 식(1)과 같다. x는 데이터를 의미하며 $f(x)$는 정답을 의미한다. $\hat{f}(x)$는 예측값이며 $E[\cdot]$는 평균값이다. 편향은 예측값들의 평균인 $E[\hat{f}(x)]$와 정답인 $f(x)$의 차이이며, 이를 통해 정답과 예측값이 서로 떨어진 거리를 알 수 있다. 분산은 예측값인 $\hat{f}(x)$값과 예측값들 평균의 차이를 제곱한 값의 평균값이며 이를 통해 예측값들이 얼마나 흩어져 있는지를 알 수 있다[13].

$$Error = Bias^2 + Variance + IrreducibleNoise$$

$$Bias = (E[\hat{f}(x)] - f(x))$$

$$Variance = E[(\hat{f}(x) - E[\hat{f}(x)])^2] \quad (1)$$

일반적으로 학습 모델에서는 편향과 분산의 트레이드오프(trade-off) 관계가 관찰된다. 편향과 분산은 반비례 관계이기에 편향이 감소하면 분산이 증가하고, 편향이 증가하면 분산이 감소한다. 이 문제는 지도 학습 모델이 편향과 분산을 각각 최소화하려 할 때 겪는 문제이다. 높은 편향을 가지는 모델은 과거에 입력된 적이 없었던 데이터에 대해 일정 부분만 일반화할 수 있다. 이러한 모델의 단순성은 학습 데이터의 과소적합으로 이어진다. 즉, 모델의 높은 편향은 학습 모델에 의한 추론 결과가 실제 값과 상당히 동떨어진 상태를 말하며, 학습 데이터와 테스트 데이터에서 모두 높은 오류를 가진다. 반면에, 높은 분산을 가지는 모델은 학습 모델 자체가 학습 데이터에만 특화되어 있다. 모델이 학습 데이터에 지나치게 맞추어져 있어 그 외의 다양한 변수에는 대응하기 어려워 이전에 입력된 적이 없었던 새로운 데이터에 대해서는 올바른 예측을 하지 못하는, 학습 모델의 일반성이 낮아지는 결과로 귀결된다. 결론적으로 편향과 분산 모두 낮아야 좋은 모델이라고 할 수 있다. 일반적으로 Fig.2.와 같이 모델의 복잡도에 따른 편향과 분산을 곡선으로 나타낼 수 있으며, 편향과 분

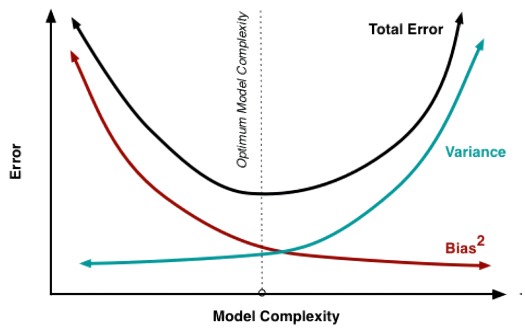

Fig. 2. Bias-variance trade-off according to the complexity of the model

산 곡선이 교차하는 지점에 해당하는 모델 복잡도를 적용하는 것이 유효한 전략으로 알려져 있다[13].

앙상블 학습은 줄일 수 없는 오류인 잡음을 제외한 모델의 편향과 분산을 최대한 줄이려 시도한다. 앙상블 학습의 기본 아이디어는 단일 기계학습 모델들을 연결하여 강력한 모델을 만드는 기법으로 하나의 기계학습 모델을 사용하였을 때 보다 더 좋은 일반화 성능을 달성한다. 본 연구는 기본 모델로 간주할 수 있는 단일 기계학습 기법들을 다수결 보팅을 사용하여 정확도를 높이고, 배깅 알고리즘을 사용하여 분산을 낮추며, 부스팅 알고리즘을 사용하여 편향을 낮출 수 있음을 확인하고 성능을 비교, 분석한다.

본 논문에서 활용한 앙상블 학습의 유형은 보팅(voting), 배깅(bagging), 부스팅(boosting)이다. 앙상블 학습은 여러 모델의 예측을 결합하여 더 나은 예측 성능을 만들기 위한 기계학습 기법이다. 보팅은 여러 기본 모델의 투표를 통하여 최종 예측에 도달하기 위해 개별 예측을 결합하는 방식이다. 보팅의 종류로는 하드 보팅과 소프트 보팅이 있는데, 일반적으로 소프트 보팅 방식의 성능이 좋기 때문에 본 논문에서도 이를 활용한다. 배깅은 부트스트랩(bootstrap)을 통해 모델을 학습시키고 결과를 집계하는 방법으로 안정성을 확보하면서 부스팅은 오류가 발생한 부분에 연속적으로 가중치를 조절하면서 학습을 수행하는 방법으로 정확도를 향상시킨다. 배깅이 무작위 복원 추출(sampling)을 통해 여러 개의 학습 데이터 표본 추출을 한 번에 수행하여 일반적인 모델을 만드는 것에 집중되어 있다면, 부스팅은 하나의 표본을 추출하여 학습 모델에 입력하고, 맞추지 못한 데이터를 포함하여 다시 표본 추출을 하여 이전 모델의 예측을 수정하여 새로운 모델을 순차적으로 만든다. 배깅은 예측 결과 오류 중 분산을 줄이고 과적합을 해소하는 것에 도움이 된다. 대표적인 배깅 알고리즘의 예로 random forest가 있다. 부스팅은 주로 분산과 더 나아가 편향을 줄이는 역할을 한다. 대표적인 부스팅 알고리즘의 예로는 adaboost, gradient boosting 등이 있다.

Table 3.은 본 연구에서 사용한 알고리즘의 목록이다. 기본(baseline) 학습 방법은 직관적이고 간단한 KNN(K-Nearest Neighbor) 알고리즘을 사용하였고, 보팅에는 LR(Logistic Regression), SVM (Support Vector Machine), DT(Decision Tree)를 KNN과 함께 사용하였다. 배깅에는 random forest를 사용하였으며, 부스팅에는 Gradient Boosting, AdaBoost, XGBoost, LightGBM, CatBoost를 사용하였다.

Table 3. List of algorithms used for detecting malicious insiders

Baseline	K-Nearest Neighbor
Voting	KNN, LR, SVM, DT
Bagging	Random Forest
Boosting	Gradient Boosting, AdaBoost, XGBoost, LightGBM, CatBoost

IV. 실험 및 평가

본 연구에서는 데이터 전처리 단계에 Python 3.7을 사용하고 기계학습 알고리즘 구현에 scikit-learn을 사용한다[14]. 또한, 앙상블 학습 기반 모델은 악의적인 내부자와 일반 사용자의 두 가지 클래스가 있는 설정으로 학습된다. CERT 데이터셋 r4.2 버전의 일부 누락된 데이터로 인해 알고리즘이 작동하지 않아 이 단계에서 결측값(missing value)을 추정 평균값으로 대체하여 데이터를 사전처리하였다.

4.1 평가 지표

본 논문의 해당 모델의 성능을 측정하기 위해 식(2)와 같은 평가 지표와 측정값을 사용한다. 우리는 먼저 기존 연구와의 비교를 위한 평가 지표로 정확도(accuracy)와 정밀도(precision)를 제시한다. 기계학습 모델 정확도는 입력 또는 학습 데이터를 기반으로 데이터셋에서 변수 간의 관계와 패턴 식별 시 가장 적합한 모델을 결정하는 데 사용되는 지표이다. 정확도는 전체 데이터 수 중 예측 결과와 실제 값이 동일한 건수가 차지하는 비율이며, 정밀도는 예측을 양성(positive)으로 한 대상 중 예측과 실제 값이 양성으로 일치한 건의 비율이다. 이와 비슷한 재현율(recall)은 실제 양성인 건 중 올바르게 양성을 맞춘 건으로 실제 정답 비율이며 TPR(True Positive Rate)이라고도 한다. 이진 분류 모델에서는 정확도만으로 판단할 경우 모델의 신뢰도가 낮아질 수 있어서 정밀도와 재현율을 사용하는 것이 더욱 바람직하다. 또 다른 평가 지표로, 기계학습 모델의 성능을 오류 관점에서 특징 지을 수 있는 편향과 분산 값을 제시한다.

또한, F1 score, ROC(Receiver Operating Characteristic) 곡선, AUC(Area Under the ROC Curve)를 추가하여 제시한다. F1 score는 정밀도와 재현율의 조화평균(harmonic mean)에 해당하는데 이는 데이터 클래스가 불균형할 때, 산술평균을 이용하는 것보다 편향을 줄일 수 있어 모델의 성능을 정확하게 평가할 수 있다. ROC 곡선은 여러 임계값에 대해 FPR(False Positive Rate)과 TPR 사이의 관계를 나타낸다. FPR과 TPR은 ROC 곡선에서 각각 가로, 세로축에 대응된다. AUC는 ROC 곡선의 밑면적을 계산한 값이며, 좋은 모델일수록 1에 가까운 값이 도출된다.

$$\begin{aligned}\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}\\ \text{Precision} &= \frac{TP}{TP+FP}\\ \text{Recall} &= \frac{TP}{TP+FN}\\ \text{F1 score} &= 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\end{aligned} \quad (2)$$

4.2 실험 결과

Table 4.는 앙상블 학습 기반 모델의 성능을 측정한 평가 결과를 나타낸다. 개별 기계학습 알고리즘을 사용한 학습 모델들의 평균 정확도는 약 93%, 정밀도는 87%, F1 score은 85%를 달성한다. 이와 비교하였을 때 앙상블 학습 기반 모델의 평균 정확도는 약 98%, 정밀도는 94%, F1 score은 91%를 달성하며 약 6%의 성능 향상을 보였다. 대체로 부스팅 방법을 사용한 학습 모델의 속도가 더 빠르고, 정확도가 높은 것을 알 수 있다. 또한, 배깅 및 부스팅 알고리즘의 성능은 KNN, LR, SVM, DT과 같은 개별 기계학습 알고리즘들과 비교한 결과 편향-분산 값이 모두 2% 이하로 균형을 이룬다. 평균적으로 Gradient Boosting과 LightGBM 학습 모델이 편향-분산 값이 가장 낮아 악의적인 내부자 탐지를 하는데 효율적인 것을 알 수 있다.

Fig. 3. 은 각 대표적인 두 가지 단일 학습 기법과 앙상블 학습 기법을 사용하여 악의적인 내부자를 예측한 결과를 나타낸 검증 곡선이다. 검증 곡선은 모델을 최적화하기 위하여 제어할 수 있는 각 하이퍼파라미터(hyperparameter)에 따른 정확도 변화를 나타낸다. Fig.3.(a)와 (b)는 개별 모델 KNN과 SVM의 검증 곡선이며, Fig.3.(c)와 (d)는 앙상블 학습 기반 모델 random forest와 LightGBM의 검증 곡선이다. 검증 곡선의 훈련 점수(training score)와 검증 점수(cross validation score)는 5-겹 교차 검증(5-fold cross validation)을 통해 평균 정확도를 계산하였다. 훈련 점수는 모델이 얼마나 학습 데이터에 일반화되었는지 나타내는데 일반적으로 훈련 점수가 높고 검증 점수가 낮으면 과대적합이고, 훈련 점수와 검증 점수가 매우 낮으면 과소적

Table 4. Evaluation results of single classification models and ensemble models

Algorithm	Accuracy	Precision	Recall	Bias	Variance	F1	AUC
KNN	0.915	0.73	0.71	0.062	0.024	0.72	0.85
LR	0.945	0.88	0.78	0.065	0.009	0.82	0.96
SVM	0.94	0.93	0.83	0.016	0.005	0.87	0.89
DT	0.935	0.91	0.89	0.016	0.012	0.89	0.93
Voting (KNN, LR, SVM, DT)	0.935	0.91	0.91	0.064	0.003	0.91	0.98
Random Forest	0.98	0.94	0.94	0.008	0.007	0.94	0.99
Gradient Boosting	0.995	0.95	0.92	**0.005**	**0.003**	0.93	0.99
AdaBoost	0.985	0.92	0.95	0.011	0.009	0.93	0.99
XGBoost	0.995	0.95	0.99	0.009	0.005	0.96	0.99
LightGBM	0.995	1.0	0.95	**0.004**	**0.003**	0.97	0.99
CatBoost	0.99	0.94	0.99	0.010	0.007	0.96	0.99

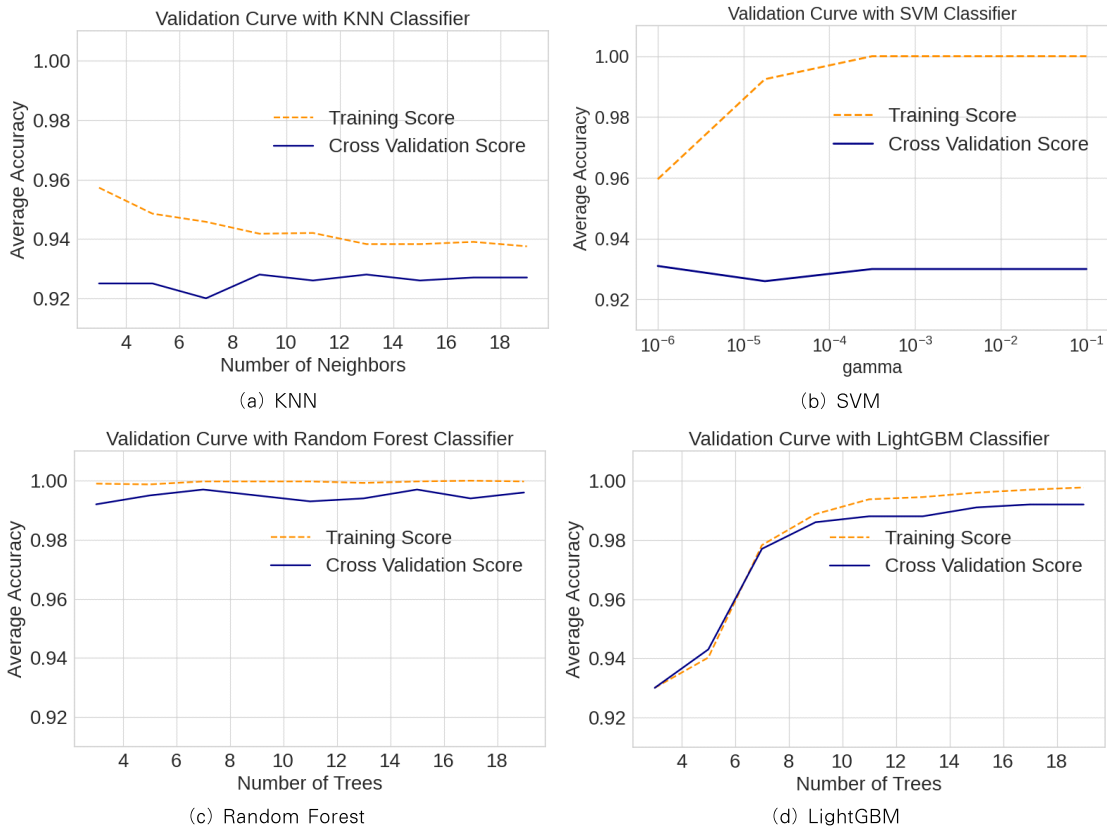

Fig. 3. The validation curve of the algorithms used

합이다. KNN과 SVM 모델을 통해 악의적인 내부자를 탐지한 결과 훈련 점수와 검증 점수가 앙상블 학습 모델인 random forest와 LightGBM보다 확연히 낮게 나타나는 것을 알 수 있다.

Fig.4.는 본 연구에서 사용한 모든 학습 모델들의 ROC 곡선을 나타낸다. 성능 정확도는 ROC 곡선 아래의 면적인 AUC에 의해 측정되며 ROC 곡선이 좌측 위쪽 모서리에 가까울수록 더 정확한 성능을 나타내는 모델이다. 학습 모델을 적용하지 않았다고 가정한 AUC 0.5를 기준으로, 개별 학습 모델 KNN은 0.85, SVM은 0.89의 AUC가 나타난다. 그 외에 다른 학습 모델들은 0.9와 1 사이의 AUC가 나타나며 이는 단일 학습 모델보다 앙상블 학습을 적용한 모델이 높은 성능을 달성하는 것을 알 수 있다.

Table 5.에 의하면 부스팅 유형 알고리즘들의 탐지 성능을 LSTM Autoencoder, GCN(Graph Convolutional networks), random forest, ANN(Artificial Neural Network)과 비교하였을 때, 상대적으로 우수한 정확도(99%)와 정밀도(97%)를 달성하는 것으로 관찰된다. 특히 심층 학습 방법을 사용하는 기존 연구와 비교하였을 때에도 탐지 성능을 향상시키고 오탐율을 낮추어 분명한 이점을 보여준다.

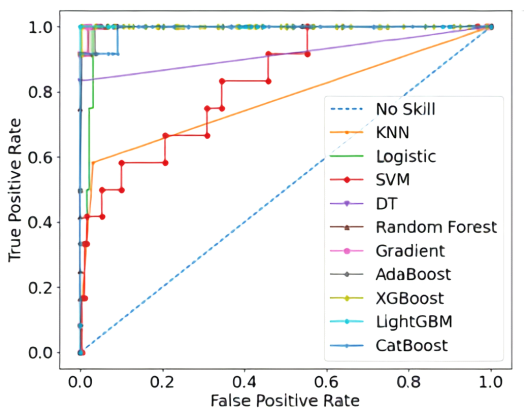

Fig. 4. The ROC curve of the algorithms used

Table 5. Accuracy and precision comparison with previous studies

No.	Reference	Algorithm	Performance Metrics	
			Acc.	Prec
1	[3]	LSTM Autoencoder	0.90	0.97
2	[15]	GCN	0.945	0.97
3	[16]	Random Forest	0.99	0.86
		ANN	0.97	0.52
4	Proposed Approach (Ensemble)	**Random Forest**	**0.98**	**0.94**
		Gradient Boosting	**0.995**	**0.97**
		LightGBM	**0.995**	**1.0**

4.3 클래스 불균형 해결 및 성능 비교

악의적인 내부자 탐지와 같은 이진 분류 문제에서 클래스 불균형은 분류할 클래스들이 균일하게 분포하지 않은 상황이며, 이로 인해 분류 모델의 성능이 저하될 수 있다. 클래스 비율이 매우 불균형하면 더 높은 비율의 클래스를 선택하는 것만으로도 모델의 정확도가 높아져 모델의 성능을 판단하기 어렵다. 이에 따라 본 논문에서는 사용된 데이터셋의 불균형한 상태를 해결하기 위하여 오버샘플링(oversampling) 기법인 SMOTE(Synthetic Minority Over-sampling Technique)[17], 언더샘플링(undersampling) 기법인 ADASYN(Adaptive Synthetic Sampling Approach)[18], 오버샘플링과 언더샘플링을 함께 수행하는 SMOTE-Tomek[19] 기법을 적용한 후, 성능을 정밀도, 재현율, 편향 및 분산의 평가 지표들로 비교 분석하였다.

서로 다른 세 가지 샘플링 기법을 KNN, RF, LightGBM에 적용한 결과는 각각 Table 6., Table 7., Table 8. 과 같다. 본 연구에서 적용한 세 가지 샘플링 기법에 따르면 재현율이 높아짐을 확인할 수 있었다. 특히, SMOTE 기법을 적용했을 때 KNN과 LightGBM은 재현율 측면에서 최적의 성능을 달성하였다. 한편, 평균적으로 정밀도와 재현율을 모두 고려하는 KNN의 F1 score는 평균적으로 SMOTE-Tomek 기법에서 가장 높았다. 여러

Table 6. Data sampling with KNN

KNN	Prec.	Recall	Bias	Var.
Baseline	0.73	0.71	0.062	0.024
SMOTE	0.74	0.93	0.058	0.019
ADASYN	0.71	0.89	0.071	0.027
SMOTE-Tomek	0.76	0.93	0.064	0.017

Table 7. Data sampling with Random Forest

Random Forest	Prec.	Recall	Bias	Var.
Baseline	0.94	0.94	0.008	0.007
SMOTE	0.97	0.95	0.05	0.006
ADASYN	0.78	0.96	0.063	0.005
SMOTE-Tomek	0.82	0.98	0.032	0.002

Table 8. Data sampling with LightGBM

LightGBM	Prec.	Recall	Bias	Var.
Baseline	1.0	0.95	0.004	0.003
SMOTE	0.92	0.99	0.01	0.003
ADASYN	0.92	0.99	0.015	0.003
SMOTE-Tomek	0.91	0.95	0.012	0.007

샘플링 기법을 적용하였을 때, 성능이 향상된 것을 알 수 있는데, 편향-분산 측면에서는 결과가 상이하게 도출되었다. 앙상블 학습에 샘플링 기법을 적용한 것은 오히려 편향을 증가시키는 결과를 야기하였다. 오버샘플링 기법들은 과적합 문제가 발생할 수 있으며, 언더샘플링 기법은 데이터를 일부 제거하기 때문에 정보 손실이 발생할 수 있기 때문이다.

V. 결론 및 향후 연구 방향

오늘날에는 사용자 접근 제어, 사용자 행동 모니터링 및 기계학습 기반 탐지를 시행함으로써 악의적인 내부자 탐지가 실현되고 있으나 모델의 오류에 대한 면밀한 분석 및 반영이 부재하여 성능 측면에서 한계가 존재한다. 본 논문에서는 예측 오류를 낮추고 탐지 성능을 향상시키는 부스팅 유형의 앙상블 학습을 사용하여 악의적인 내부자 탐지 성능을 확인하고 이에 대한 면밀한 분석을 수행하였다. 우리는 다양한

내부자 위협 시나리오를 반영하는 공개 데이터셋을 활용하였으며, 실험을 통해 적용된 알고리즘의 성능을 LSTM Autoencoder, GCN, ANN과 같이 잘 알려진 다른 기법과 비교하였다. 앙상블 학습을 사용한 악의적인 내부자 탐지 기법은 비교적 우수한 정확도(98%), 정밀도(97%) 및 F1 score(91%)를 산출하는 것으로 나타났으며, 적용 가능한 배깅 및 부스팅 알고리즘 중 Gradient Boosting과 LightGBM이 우수한 성능을 보이는 것으로 관찰되었다. 또한 학습에 사용된 CERT 데이터셋의 불균형 문제를 해결하기 위해 SMOTE, Adasyn, SMOTETomek 기법을 적용하였다. 이를 적용함으로써 성능은 향상되지만, 현재는 오히려 편향과 분산이 증가한다는 것을 확인할 수 있었다. 향후 연구에서는 학습 모델을 통해 분류할 클래스들이 균일하게 분포하지 않은 비대칭 데이터를 훈련하는 상황에서, 편향 및 분산을 최소화하는 방법에 관한 확장 연구를 수행할 계획이다.

References

[1] Gurucul, "2021 Insider Threat Report" https://gurucul.com/2021-insider-threat-report, Accessed: 25 Oct. 2021.

[2] D.C. Le and N. Zincir-Heywood, "Anomaly Detection for Insider Threats Using Unsupervised Ensembles," IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1152-1164, Jun. 2021.

[3] R. Nasir, M. Afzal, R. Latif, and W. Iqabl, "Behavioral Based Insider Threat Detection Using Deep Learning," IEEE Access, vol. 9, pp. 143266-143274, Oct. 2021.

[4] A. Khan, R. Latif, S. Latif, S. Tahir, G. Batool, and T. Saba, "Malicious Insider Attack Detection in IoTs Using Data Analytics," IEEE Access, vol. 8, pp. 11743-11753, Dec. 2019.

[5] D.C. Le, N. Zincir-Heywood, and M.I. Heywood, "Analyzing Data Granularity Levels for Insider Threat Detection Using Machine Learning," IEEE Transactions on Network and Service Management, vol. 17, no. 1, pp. 30-44, Mar. 2020.

[6] H. Bian, T. Bai, M.A. Salahuddin, N. Limam, A.A. Daya, and R. Boutaba, "Uncovering Lateral Movement Using Authentication Logs," IEEE Transactions on Network and Service Management, vol. 18, no. 1, pp. 1049-1063, Mar. 2021.

[7] G. Kaiafas, G. Varisteas, S. Lagraa, R. State, C.D. Nguyen, T. Ries, and M. Ourdane, "Detecting Malicious Authentication Events Trustfully," Proceedings of the 2018 IEEE/IFIP Network Operations and Management Symposium, pp. 1-6, Apr. 2018.

[8] T. Yang, Y. Lin, C. Wu, and C. Wang, "Voting-Based Ensemble Model for Network Anomaly Detection," Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8543-8547, Jun. 2021.

[9] B.A. Tama, M. Comuzzi, and K. Rhee, "TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System," IEEE Access, vol. 7, pp. 94497-94507, Jul. 2019.

[10] Software Engineering Institute, Carnegie Mellon University, "CERT Insider Threat Test Dataset" https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=508099, Accessed: 20 Sept. 2021.

[11] Scikit-learn, "6.3. Preprocessing data" https://scikit-learn.org/stable/modules/preprocessing.html, Accessed: 30 Oct. 2021.

[12] P. Domingos, " Unified Bias-Variance Decomposition and its Applications," Proceedings of the 17th International

Conference on Machine Learning, pp. 231-238, Jun. 2000.
[13] Scott Fortmann-Roe, "Understanding the Bias-Variance Tradeoff" http://scott.fortmann-roe.com/docs/BiasVariance.html, Accessed: 31 Nov. 2021.
[14] scikit-learn, "Machine Learning in Python" https://scikit-learn.org/stable, Accessed: 31 Oct. 2021.
[15] J. Jiang, J. Chen, T. Gu, K.R. Choo, C. Liu, M. Yu, W. Huang, and P. Mohapatra, "Anomaly Detection with Graph Convolutional Networks for Insider Threat and Fraud Detection," Proceedings of the 2019 IEEE Military Communications Conference, pp. 109-114, Nov. 2019.
[16] D.C. Le and A.N. Zincir-Heywood, "Machine learning based Insider Threat Modelling and Detection," Proceedings of the 2019 IFIP/IEEE Symposium on Integrated Network and Service Management, pp. 1-6, Apr. 2019.
[17] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, no.1, pp. 321-357, Jun. 2002.
[18] H. He, Y. Bai, E.A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 1322-1328, Jul. 2008.
[19] G.E. Batista, A.L.C. Bazzan, and M. Monard, "Balancing Training Data for Automated Annotation of Keywords: a Case Study," Proceedings of the 2nd Brazilian Workshop on Bioinformatics, pp. 35-43, Jan. 2003.

〈저자소개〉

박 수 연(Suyun Park) 학생회원
2019년 8월: 서울여자대학교 컴퓨터학과 졸업
2020년 9월~현재: 고려대학교 정보보호대학원 석사과정
〈관심분야〉 네트워크 보안, 머신러닝, 침입탐지시스템