Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.27, NO.4, Aug. 2017 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2017.27.4.763

기계학습 기반 내부자위협 탐지기술: RNN Autoencoder를 이용한 비정상행위 탐지*

하 동 욱,† 강 기 태, 류 연 승‡
명지대학교 대학원 보안경영공학과

Detecting Insider Threat Based on Machine Learning: Anomaly Detection Using RNN Autoencoder*

Dong-wook Ha,† Ki-tae Kang Yeonseung Ryu‡
Department of Security and Management Engineering, Myongji Univ.

요 약

최근 몇 년 동안 지속적으로 개인정보유출, 기술유출 사고가 빈번하게 발생하고 있다. 조사에 따르면 이러한 유출 사고의 주체로 가장 많은 부분을 차지하고 있는 것이 조직 내부에 있는 '내부자'로, 내부자에 의한 기술유출은 조직에 막대한 피해를 주기 때문에 점점 더 중요한 문제로 여겨지고 있다. 본 논문에서는 내부자위협을 방지하기 위해 기계학습을 이용하여 직원들의 일반적인 정상행위를 학습하고, 이에 벗어나는 비정상 행위를 탐지하기 방법에 대한 연구를 하고자 한다. Neural Network 모델 중 시계열 데이터의 학습에 적합한 Recurrent Neural Network로 구성한 Autoencoder를 구현하여 비정상 행위를 탐지하는 방법에 대한 실험을 진행하였고, 이 방법에 대한 유효성을 검증하였다.

ABSTRACT

In recent years, personal information leakage and technology leakage accidents are frequently occurring. According to the survey, the most important part of this spill is the 'insider' within the organization, and the leakage of technology by insiders is considered to be an increasingly important issue because it causes huge damage to the organization. In this paper, we try to learn the normal behavior of employees using machine learning to prevent insider threats, and to investigate how to detect abnormal behavior. Experiments on the detection of abnormal behavior by implementing an Autoencoder composed of Recurrent Neural Network suitable for learning time series data among the neural network models were conducted and the validity of this method was verified.

Keywords: Insider threat, Machine learning, Neural network, Anomaly detect, Information security

I. 서 론

글로벌 시대로 접어들면서 기업들은 글로벌 시장에서 우위를 선점하고 고객 확보를 위해 새로운 기술 개발에 많은 노력을 기울이고 있다. 이와 동시에 지속적으로 기술유출 사고와 개인정보유출 사고가 발생하고 있는데, 국가정보원 산업기밀보호센터에서 조사한 '내부자료 유출 실태조사'에 따르면 2005년부터 2012년 까지 국내 첨단기술 유출 또는 유출 시도 사건은 총 294건으로 집계되었고, 적발 건수는 더 늘어나고 있다. 유출사고의 형태는 여러 가지가 있지만 그중에서 가장 빈번히 일어나고 있고 막기 힘든 것은

Received(03. 10. 2017), Modified(04. 28. 2017), Accepted(05. 31. 2017)
* 이 연구는 2016년도 명지대학교 교책중점연구소 지원으로 연구되었음
† 주저자, hdu0105@gmail.com
‡ 교신저자, ysryu@mju.ac.kr(Corresponding author)

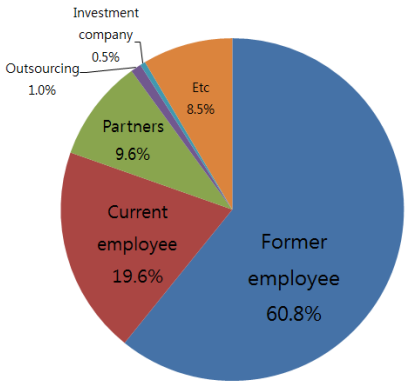

Fig. 1. The main agent of Industry Confidential Spill

내부자에 의한 내부 정보 유출 이다. 실제로 우리나라의 사례를 들어보면 2012년 6월에 발생한 삼성, LG의 AMOLED관련 핵심기술 국외 유출 사건, 2015년 10월 국내 석유시추선 핵심정보유출 시도 사건, 과거 발생한 금융권 대규모 정보 유출사고 등이 내부자에 의해 발생한 유출 사고이다. 내부자란 조직 내의 인원, 시설, 정보자산에 대해 합법적인 접근 권한을 가지고 있는 인원을 의미하며, 좁게는 전·현직 정규직원, 계약직 직원, 내부 상주 협력업체 직원을 포함할 수 있고, 넓게는 양도·양수, 합병 담당자, 공동·위탁 연구자등 법률·규정에 의거 내부 정보를 열람 가능한 사람을 의미한다. 이들은 악의적 의도를 품은 외부인에 의한 공격보다 손쉽게 내부의 자산에 접근할 수 있고, 더 큰 확률로 조직에 막대한 피해를 입힐 수 있다[1]. 실제로 국정원 조사에 따르면 산업기밀 유출사고의 주된 원인이 전직 혹은 현직직원에 의한 유출로 조사되고 있고 그 뒤로 협력업체, 연구 용역으로 역시 내부자라고 정의할 수 있는 사람에 의해 유출 되었다는 것을 알 수 있다. 금융보안원에서는 2017년 금융 IT보안 10대 이슈 전망 보고서에서 제3자 및 내부자 보안관리 중요성 증대에 대해 얘기 하고 있으며, 여기에는 글로벌 사이버 보안 전문가 280명을 대상으로 실시한 사이버 리스크 관리에 따른 비용절감 효과에 대한 설문조사에서 이용자 행위 분석이 가장 큰 비용절감 효과를 가져다주었다고 조사 됐다는 내용을 소개하고 있다[2].

이처럼 내부자 위협은 현재 가장 주목 할 보안 위협중 하나로 여겨지고 있으며, 여러 산업에 걸쳐 점점 큰 이슈로 자리 잡고 있다. 또한, 사이버 리스크 관리 비용을 줄이기 위해서도 내부자의 행동을 분석

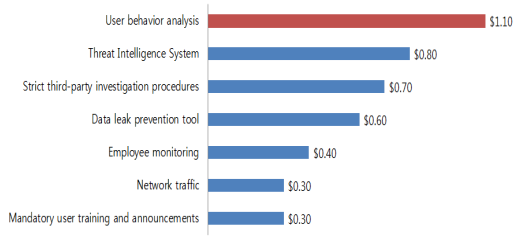

Fig. 2. Cost reduction due to cyber risk management

하여 내부자 위협에 대해 적극적으로 대응하는 것이 중요해지고 있다.

본 논문에서는 이러한 상황을 고려하여 내부자 위협을 미연에 방지하기 위한 방법에 대하여 연구를 한다. 기계학습을 이용하여 사용자의 정상행위를 학습하고 그 후 사용자의 정상행위와 다른 행위를 비정상행위라고 가정하고 탐지하고자 한다. 실험에 사용한 데이터는 카네기멜론 대학의 CERT부서에서 내부자 위협 연구를 위해 제공하는 CERT 데이터를 이용하였다.

본 논문의 구성은 다음과 같다. 2장에서 내부자 위협과 이를 탐지하기 위한 관련연구에 대하여 소개를 하고 3장에서 실험에 사용한 데이터 소개와 전처리 방법에 대하여 서술한다. 4장에서는 RNN Autoencoder를 직접 구현하여 CERT데이터에 적용하고 유효성을 검증한다. 5장에서는 본 논문에 대한 결론과 한계, 향후 연구에 대한 방향을 제시하고자 한다.

II. 관련 연구

2.1 내부자 위협 탐지

내부자 위협에 대한 문제는 과거부터 꾸준히 연구되어 오던 분야로 주요한 이슈로 떠오르면서 점점 더 활발히 연구되고 있는 상황이다. 최근 기계학습의 발전과 함께 이를 내부자 위협 탐지 분야에 적용한 여러 선행 연구들이 있다. 가장 대표적인 방법으로는 Hidden Markov Model(HMM)을 사용하여 탐지하는 모델이다. HMM은 일련의 순차적인 성질을 내포하고 있는 데이터를 다루는 문제에 잘 적용 될 수 있는 모델로써, 음성데이터와 같은 순차적 데이터를 인식하는데 널리 사용되어 왔다. Tabish Rashid 등[3]은 HMM을 이용한 내부자위협 탐지

에 대한 연구를 하였다. HMM을 이용하여 사용자의 정상적 행위를 학습하고 이에 벗어나는 비정상 행위를 탐지할 수 있다는 것을 보여줬고, 추가적으로 앞으로 더 해야 할 연구를 제시하고 있다. Pallabi Parveen 등[4]의 연구에서는 압축과 점진학습을 결합하여 내부자 위협을 탐지하는 실험을 진행하였고, 통계적 방법보다 더 좋은 성능을 나타냈다고 밝히고 있다. 장현성[5]의 연구는 한 기업의 문서관리 시스템을 대상으로 일정 기간 콘텐츠 중심의 사용 현황을 분석하여 조직 내부의 정상적인 사용자의 비정상적인 사용 행위를 탐지할 수 있는 모델을 정립하고 실험 및 결과를 제시했다. 비지도학습의 대표적인 방법인 K-means알고리즘과 SOM(Self Organizng Map)을 이용하였으며 실험 결과를 통해 인증된 사용자의 비정상적인 행위를 클러스터링에 의해 탐지될 수 있음을 확인하였고, 정상적인 행위도 비정상적인 행위로 판단될 수 있음을 확인 하였다. 김정홍 등[6]의 연구에서는 K-means, Gaussian density estimation 등의 one-class classification 방법을 이용하여 비정상행위와 정상행위를 분류하는 방식으로 비정상 행위를 탐지하는 방법과 이메일의 토픽 분포를 분석하여 비정상 행위를 탐지하는 방법에 대한 실험을 하였다. 기계학습 기반의 탐지 방법 외에도 방법론 또한 계속적으로 연구되고 있다. Oliver Brdiczka 등[7]은 사람의 행동과 성격양식이 내부자 탐지에 있어 똑같이 중요하다는 점을 강조하며 사람의 심리 프로파일 정보를 탐지 모델에 결합하여 내부자 위협 탐지의 정확도를 높이는 방법에 대한 연구를 하였다. 미국 국토 안보부[8]는 내부자의 성격과 행동양식에 대한 보고서를 발간 하였는데, 주로 어떤 특징을 갖고 있는 성격이 내부자 위협을 저지를 가능성이 높은지에 관한 연구를 다루고 있다. 임용환 등[9]은 주요 비즈니스 혹은 병원 고객들을 분류하기 위한 스코어링 모델인 RFM(Recency Frequency Monetary)모델을 응용하여 내부자의 활동 수준을 평가하기 위한 SFI(Span Frequency Importance) 분석 기법을 활용 하여 내부자들의 활동에 대한 스코어링을 통해 위협을 탐지하는 방법을 실제 기업에 적용하여 사례 연구를 수행하였다.

2.2 Recurrent Neural Network(RNN)와 Autoencoder

RNN은 인공신경망의 모델 중 한 종류로써 입력으로 이전의 입력과 함께 현재의 입력을 고려하게 되는 신경망 모델로, 시계열 데이터 학습에 적합한 알고리즘 이다. 기존의 일반적인 신경망 모델은 입력으로 현재의 하나의 입력만 처리하였기 때문에 입력순서에 독립적이라고 말할 수 있지만 RNN은 현재의 입력과 이전의 입력을 함께 고려하기 때문에 입력순서에 종속적인 성질을 나타낸다. 그렇기 때문에 최근 자연어 처리 문제와 같이 입력 순서를 고려하여야 하는 문제에서 많이 이용되고 있고, 뛰어난 성능을 보이고 있는 모델이다. RNN의 R을 나타내는 단어인 Recurrent는 동일한 동작을 모든 시퀀스 요소마다 적용하고, 이전 시퀀스의 아웃풋을 현재 시퀀스의 인풋으로 함께 고려한다는 것을 의미하게 된다. RNN은 구현하는 방법에 따라 몇 가지로 다시 나뉠 수 있는데, LSTM(Long Short Term Memory)과 GRU(Gated Recurrent Unit)가 그 중 하나이다.

Autoencoder란 기계학습 알고리즘중 하나로 별도의 답을 가르쳐주지 않은 상태에서 인공지능이 스스로 학습을 하게 되는 비지도학습 방법이다. Autoencoder는 개념적으로 encoder와 decoder가 연결되어 구성된 모델로 생각 할 수 있다. 입력으로 받은 x를 encoder를 통해 보이지 않는 출력 값 z를 만들어내고, z를 decoder의 입력으로 하여 최종 출력 값 x'을 출력하게 된다. 이 과정을 거쳐 나온 최종 출력 값 x'과 입력 값 x를 같은 값을 갖게 하도록 하는 것을 Autoencoder라고 할 수 있다. 학습과정에서 encoding과 decoding 과정을 반복하면서 입력 값의 압축된 표현을 담은 모델을 생성하는 것이 Autoencoder의 목적이다. 이러한 특징을 이용하여 Autoencoder는 이상 탐지, 데이터 생성

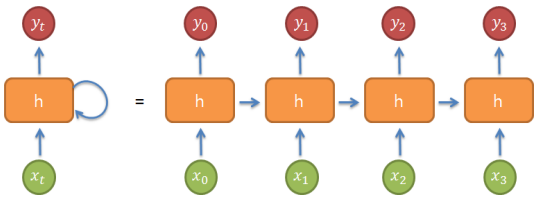

Fig. 3. RNN structure

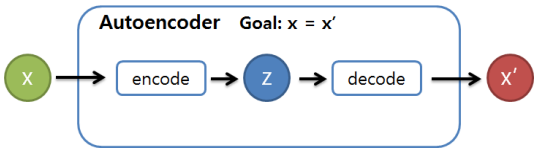

Fig. 4. Autoencoder

모델 학습 등 여러 분야에서 활용되고 있다.

기존 연구에서는 사용자 행동을 예측하기 위한 방법으로 대부분 HMM을 사용하거나 클러스터링 혹은 분류 방법을 이용하여 정상행위에서 벗어나는 비정상행위를 분류하려고 하였다. 본 논문에서는 내부자위협 탐지 분야에서 비교적 많이 다루지 않았던 Neural Network를 이용한 탐지에 대한 연구를 진행 한다.

2.3 연구의 차별성

HMM모델은 Markov 가정에 의해 현재의 행동에 대한 예측이 바로 직전의 행동에만 의존을 한다는 한계가 있다[3]. 클러스터링을 이용한 탐지 모델의 문제점은 정상 클러스터와 이상 클러스터를 구분하기 어렵다는 점이다. 예를 들어 K-means 방법의 경우 아래 fig.5와 같이 어떤 클러스터가 정상 클러스터인지 구분하기 어렵다. HMM을 이용한 [3]의 실험에서는 사용자의 행위를 1주일 단위로 탐지를 해내는 방식으로 진행하였기 때문에, 탐지의 효율성이 떨어지는 문제가 발생 할 수 있다. 예를 들어, 월요일에 발생한 위협 행위가 금요일이 지나서야 탐지가 되기 때문에 이미 중요 데이터는 외부로 유출될 가능성이 크다.

본 논문에서는 앞선 연구들의 단점을 보완하기 위해 주 단위가 아닌 일 단위로 탐지를 해냄으로써 위협 행위가 일어난 후 신속하게 조치를 취할 수 있도록 하고, RNN을 이용하여 더 긴 의존성을 가지도록 하여 더 정확한 탐지를 할 수 있도록 할 것이다. 또한 탐지 값을 이용한 탐지로 이상 행위와 정상 행위를 더욱 직관적으로 구분 할 수 있는 모델을 제안하고자 한다.

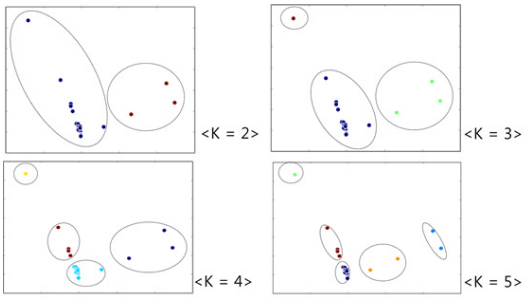

Fig. 5. Limitation of K-means clustering

III. 데이터 소개 및 데이터 전처리

3.1 데이터 소개

본 논문의 실험에 사용한 데이터는 CERT dataset을 이용하였다. CERT dataset은 카네기 멜론 대학의 CERT 부서에서 제공하는 내부자위협 연구 데이터로 가상의 기업에 다니고 있는 내부자들의 데이터들이 담겨져 있고, 내부자들 중엔 악의적 의도를 갖고 회사에 피해를 주는 행위를 하는 악의적 내부자도 존재한다[10]. 현재 r1버전부터 r6버전까지 나왔으며 우리의 실험에서는 r4.2를 이용하였다. r4.2는 다른 dataset에 비해 더 많은 내부자들을

Table 1. CERT dataset insider threat scenario

Scenario 1	User who did not previously use removable drives or work after hours begins logging in after hours, using a removable drive, and uploading data to wikileaks.org. Leaves the organization shortly thereafter.
Scenario 2	User begins surfing job websites and soliciting employment from a competitor. Before leaving the company, they use a thumb drive (at markedly higher rates than their previous activity) to steal data.
Scenario 3	System administrator becomes disgruntled. Downloads a keylogger and uses a thumb drive to transfer it to his supervisor's machine. The next day, he uses the collected keylogs to log in as his supervisor and send out an alarming mass email, causing panic in the organization. He leaves the organization immediately.
Scenario 4	A user logs into another user's machine and searches for interesting files, emailing to their home email. This behavior occurs more and more frequently over a 3 month period.
Scenario 5	A member of a group decimated by layoffs uploads documents to Dropbox, planning to use them for personal gain.

포함하고 있기 때문에 더 다양한 비정상 행위를 보여주고 있다. Table 1은 CERT dataset에서 정의하고 있는 내부자 위협 시나리오를 나타낸다. r4.2를 제외한 다른 버전의 dataset에서는 각 시나리오를 수행하는 내부자가 1~2명 존재하고 있지만, 우리가 사용한 dataset인 r4.2는 시나리오 1, 2 각 30명, 시나리오 3 10명이 존재하고 있고, 실험을 위해 임의적으로 r6버전에 존재하는 시나리오 4, 5의 내부자를 1명씩 추가 시켰다.

시나리오 1은 사용자가 평소에 하지 않던 행위를 하여 위협을 수행하는 시나리오이고, 시나리오 2는 평소보다 더 많은 특정 행위를 하여 데이터를 훔쳐가는 시나리오이다. 시나리오 3과 4는 계정 도용에 관한 시나리오이고 시나리오 5는 중요 데이터를 웹에 업로드 하는 시나리오 이다. 시나리오 1,3,4는 위협 행위를 수행하면서 사용자가 평소에 하지 않던 새로운 패턴으로 위협행위를 수행하게 되고, 시나리오 2는 특정 행위를 포함하는 패턴이 평소보다 패턴에 더 많이 나타나게 된다. 시나리오 5는 웹사이트에 접근하는 패턴이 나타나게 된다.

3.2 데이터 전처리

CERT dataset은 조직의 직원들의 행동 로그를 담고 있는 여러 파일로 구성되어 있다. logon, logoff, 웹사이트 접속, 이메일 전송, 파일을 이동식 디스크에 복사, 이동식 디스크를 연결, 연결해제에 대한 시간과 행위의 주체를 담고 있는 logon.csv, http.csv, email.csv, device.csv 파일과 직원들의 성향에 대한 내용을 담고 있는 psychometric.csv, 사용자들의 직위, 부서, 근무기간, 참가 프로젝트 등을 담은 LDAP 파일로 이루어져 있다. 다음은 device.csv 파일의 필드구조와 내용을 나타낸 표이다. 나머지 파일들도 조금씩 차이가 있지만 거의 유사한 구조로 구성되어 있다.

Table 2. File Structure of device.csv

Field	Content
id	primary key
date	dd/mm/yyyy hh:mm:ss
user	User ID
pc	PC ID
activity	connect/disconnect

우리의 실험에서는 psychometric.csv와 LDAP를 제외한 파일들만을 이용하여 간단한 탐지 지표를 정의 하였다. 여러 파일로 분산되어 저장되어있는 직원 개개인의 로그를 전처리를 통해 사용자의 하루하루의 행위를 시간 순서대로 담은 개별 파일로 만들었고 실험의 편의를 위해 각 행위를 다음과 같이 숫자로 표현하였다.

logon: 0 http: 1 email: 2 file: 3 connect: 4 disconnect: 5 logoff: 6

logon, logoff는 사용자가 pc에 사용자 계정을 이용하여 접속/해제 하는 행위를 나타내고, http는 웹사이트에 접속하는 행위, file은 이동식장치에 파일을 복사하는 행위, connect, disconnect는 이동식 디스크를 pc에 연결/연결해제 하는 행위를 나타낸다. fig. 6.과 fig. 7.은 이러한 데이터 전처리 과정과 전처리 된 각 사용자 파일의 내용을 표현한 그림이다.

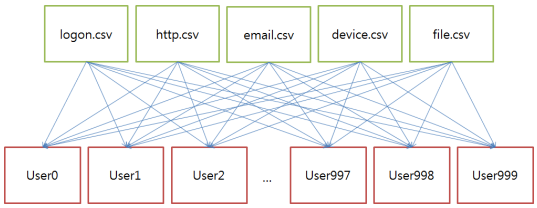

Fig. 6. File preprocess

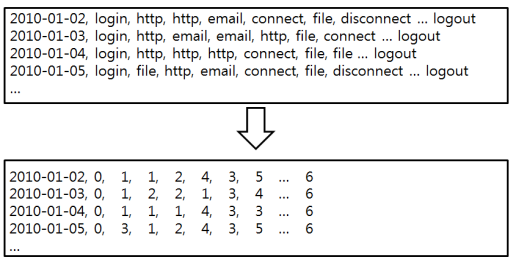

Fig. 7. Example of preprocessed file

IV. RNN Autoencoder를 이용한 내부자 위협 탐지

본 논문에서는 앞서 설명한 RNN을 이용하여 구현한 Autoencoder를 이용하여 사용자의 일반적인 정상 행위를 학습을 하고, 정상 행위에 벗어나는 비정상 행위를 탐지하도록 할 것이다.

4.1 정상행위 학습

비정상 행위를 탐지해 내기 위하여 먼저 정상행위를 학습을 하여야 하는데 그에 앞서 몇 가지 가정이 필요하다. 첫 번째, 정보 유출을 시도하는 내부자는 원래는 악의적 의도가 없는 정상적 내부자였지만, 어느 순간 악의적 의도를 품고 정보 유출을 시도한다. 두 번째, 직원들은 평소 업무를 하는 일정한 패턴이 있다. 이러한 일정 패턴을 우리는 정상행위라고 한다. 예를 들어 이메일을 많이 이용하는 사용자의 정상 패턴은 〔0,1,2,1〕, 〔1,2,1,2〕 등 과 같이 나타날 수 있다. 세 번째, 내부자 위협은 평소에 하던 정상 행위 패턴과 다른 패턴 이거나 평소보다 많은 특정 행위의 반복 이다. USB를 이용하지 않던 사용자가 어느날 갑자기 USB를 이용한다던지, 평균적으로 2~3번 USB를 사용하던 사용자가 갑자기 5~6번 USB를 사용하는 경우가 이에 해당한다. 네 번째, 하루 일과 시퀀스는 몇 가지의 행위 패턴들로 이루어져 있고, 행위 패턴은 연속되는 몇 가지의 행위로 구성된다. 예를 들어 한 사용자의 일과 시퀀스가 〔0,1,1,2,1,4,3,5,6〕로 나타난다면, 이는 행위 패턴 〔0,1,1,2〕,〔1,1,2,1〕,〔1,2,1,4〕,〔2,1,4,3〕, 〔1,4,3,5〕,〔4,3,5,6〕 들로 구성된다. 이 가정들을 바탕으로 직원들의 초기 활동 데이터는 직원들의 정상행위 패턴을 나타내는 악의적 의도가 없는 정상적인 활동이라는 것을 이용하여 정상 행위를 학습 하고, 그 후 이에 어긋나는 행위 패턴을 담은 날을 비정상 행위가 발생 한 날로 판단하고 탐지해 낸다.

우리 실험에서는 패턴의 길이를 4로 정하여 학습 패턴을 구성한다. 패턴의 길이를 4로 정한 이유는 어떤 행위를 할 때 나타날 최소 길이이기 때문이다. 예를 들어 파일을 메일로 보낸다고 했을 때 〔login, file, email, logout〕과 같이 나타나게 되므로 하루에 최소 4가지 행위가 존재하여야 한다. 충분한 학습 데이터를 얻기 위해 사용자의 초기 60일 간의 일과 시퀀스에서 나오는 모든 패턴을 학습 데이터 세트로 구성하고 Autoencoder의 학습을 진행한다.

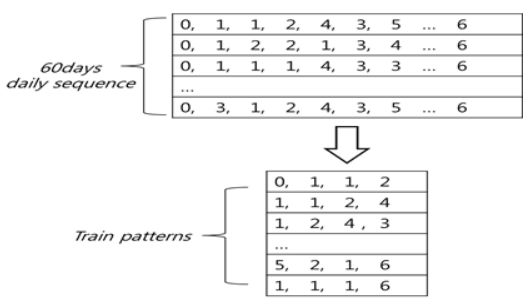

Fig. 8. Extract train patterns

학습과정은 다음 수식들을 이용한다. σ는 activation function을 의미한다.

$$y = f(x) = \sigma(Wx + b) \tag{1}$$

$$x' = g(y) = \sigma'(W'y + b') \tag{2}$$

$$L(x,x') = \|x - x'\|^2 \tag{3}$$

(1)은 Autoencoder의 encoder를 나타내고 (2)는 decoder를 나타낸다. (1)에서 x는 학습 데이터 세트를 나타내고 (1)과 (2)의 W, b, W', b' 값들은 초기에 임의의 값으로 주어지며 학습을 통해 최적의 값을 갖게 된다. 입력 값 x를 (1)의 입력으로 하여 y를 구하고 이것을 (2)의 인풋으로 하여 출력 값 x'을 구하게 된다. 그 후 x와 x'을 (3)의 loss 함수에 입력으로 넣어 loss 값을 구한다. 학습을 거듭하면서 loss 값을 최소로 하는 W, b, W', b' 값을 구하게 된다.

4.2 비정상 행위 탐지

학습을 마치면 각 직원들에 대한 정상 행위에 대한 정보를 압축하여 담고 있는 Autoencoder를 얻을 수 있다. 이 후에 비정상 행위를 탐지하기 위하여 하루 일과 시퀀스를 하나씩 불러와 시퀀스가 갖고 있는 행동 패턴을 추출한다. 그 후 Autoencoder의 입력으로 넣어 얻은 출력을 입력과 비교하여 loss 값을 구하고 이를 이용하여 이상행위를 탐지한다. 만약 입력 패턴이 정상 패턴과 동일하거나 유사한 패턴이라면 출력 패턴 또한 유사한 패턴의 형태를 나타내기 때문에 작은 loss 값을 갖게 되고, 학습과정에서 나오지 않은 다른 패턴, 즉, 비정상 패턴이라면 입력 패턴과 출력 패턴이 서로 다르게 되어 큰 loss 값을 갖게 된다. 이때 평소보다 특정행위를 더 많이 반복하여 발생하는 비정상 행위를 탐지하기 위해 특정 행위가 더 많은 날의 loss 값에 penalty를 주도록 하였다. penalty는 다음의 수식을 이용하였다.

$$Z = \frac{x - \mu}{\sigma} \tag{4}$$

(4)는 표준 값을 구하는 식으로 x는 하루 일과시퀀스에 특정 행위가 들어있는 수를 의미하고 σ와 μ는 학습 데이터 세트에서 특정 행위의 표준편차와 평균값을 나타낸다. 표준 값은 원수치 x가 평균에서 얼마나 떨어져 있는지 나타내는 수치로, 평소보다 특정 행위를 더 많이 할수록 큰 값을 나타낸다. (4)를 통해 구한 표준점수에 임의의 값을 곱한 후 loss 에 더하여 최종 loss 값을 얻을 수 있다. (5)는 최종 loss 값을 구하는 수식이다.

$$penalty = \alpha \times Z \qquad (5)$$
$$loss = L(x,x') + penalty$$

우리의 실험에서는 시나리오 2번을 탐지하기 위해 이동식 디스크를 삽입하는 행위를 특정행위로 정하여 실험을 진행하였다. 마지막으로 어느 정도의 loss 값을 갖는 패턴을 비정상 행위로 간주하고 탐지를 할지 정하기 위해 Threshold를 정한다. Autoencoder를 통해 나온 loss 값과 Threshold를 비교하여 그 이상의 loss 값을 갖는 날을 비정상행위로 간주하고 탐지해낸다. fig. 9는 이러한 탐지 절차를 그림으로 도식화한 내용이다.

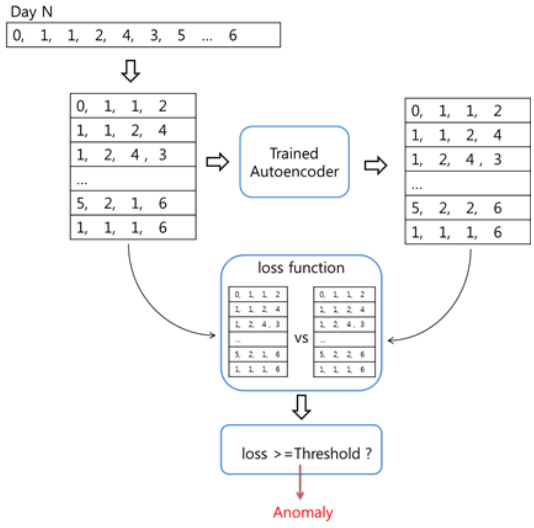

Fig. 9. Process of anomaly detect using Autoencoder

4.3 실험 결과

앞서 설명한 것과 같이 본 논문에서는 RNN으로 구성한 Autoencoder를 구성하여 '순서'에 종속적으로 정상패턴을 학습하도록 하였고, 훈련 데이터를 포함한 각 사용자의 전체 일과 시퀀스를 테스트 데이터로 하여 실험을 진행하였다. fig. 10.은 악의적 행위를 통해 데이터를 유출하는 내부자의 결과 값을 나타낸 그래프로 평소에는 0에 가까운 낮은 loss 값을 갖지만 악의적 행위가 일어난 날은 높은 loss 값을 갖는 것을 확인 할 수 있다.

CERT dataset r4.2는 직원 1000명의 약 1년 6개월 동안의 데이터를 갖고 있고, 그중 악의적 내부자 시나리오 1, 2, 3만을 포함하고 있다. 우리는 시나리오 4, 5에 대한 유효성도 함께 검증하기 위해 r6.2에 포함되어있는 사용자 두 명을 추가하였다. 최종적인 실험에 사용된 데이터 총 1002명의 직원의 데이터를 갖고 있고 이중 72명이 내부자위협을 수행하는 내부자이다. Table 3.은 실험에 사용된 데이터 세트의 구성을 나타낸 표이다.

임의의 값으로 정해지는 Threshold는 매우 중요한 값이기 때문에 가장 알맞은 Threshold를 찾기 위해 여러 값으로 변경하면서 실험을 진행하였고 그 결과는 Table 4.와 같다. 이때 penalty 값의 α값을 0.3으로 하여 진행하였다.

Accuracy는 정확도를 나타내는 항목으로 전체 데이터 중 예측이 맞은 경우에 해당하는 항목으로 Threshold를 이용하여 예측한 예측 값이 적중한

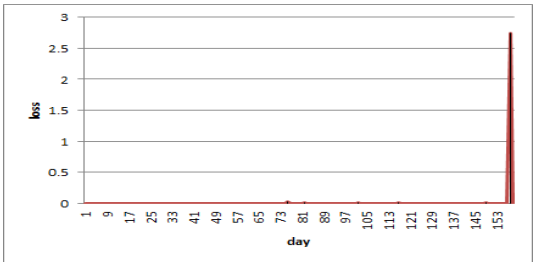

Fig. 10. loss graph of insider

Table 3. Configuration of Experimental Dataset

Total employees	1002
Insiders	72
Scenario 1	30
Scenario 2	30
Scenario 3	10
Scenario 4	1
Scenario 5	1

Table 4. Detection result by threshold

	0.9	1.0	1.1	1.2	1.3
Accuracy	0.997	0.998	0.998	0.998	0.999
Sensitivity	0.957	0.793	0.780	0.777	0.777
Specificity	0.998	0.998	0.999	0.999	0.999
Precision	0.351	0.434	0.489	0.519	0.679

비율을 나타낸다. Sensitivity는 민감도를 말하며, 이는 정확하게 탐지된 비정상행위의 비율을 의미한다. Specificity는 특이도 항목으로, 정확하게 탐지된 정상행위의 비율을 나타낸다. Precision은 정확률이며 이는 비정상행위라고 탐지한 것 중에 실제 비정상행위인 것의 비율을 나타낸다. Table 3.을 보면 threshold 값이 0.9에서 1.0으로 변할 때 Sensitivity 값은 0.164만큼 감소 하지만 Accuracy 0.001증가, Specificity 변동 없음, Precision 0.083증가 하는 것을 관찰할 수 있고 그 결과 변동 폭이 가장 큰 sensitivity를 기준으로 가장 좋은 성능을 보인 threshold는 0.9 구간으로 확인 할 수 있다. 우리의 실험에서 Precision이 낮게 나온 이유는 비정상행위로 탐지를 했지만 실제로는 정상행위이기 때문이다. 예를 들면, 한번도 USB를 사용하지 않던 사용자가 동료의 부탁으로 악의적 의도가 전혀 없이 USB를 사용했다면 우리의 모델에서는 이것을 비정상행위로 탐지를 한다. 실제로 이것은 정상행위지만 정황을 모른다면 비정상행위로 의심할 수 있는 행위이기 때문에 오탐이라고 할 수는 없다. 즉 우리의 실험에서는 Precision보다는 Sensitivity가 더 중요하다.

Threshold와 마찬가지로 penalty값을 정하는 것도 아주 중요한 과정이므로 마찬가지로 여기에 사용되는 임의의 변수 α값을 여러 값으로 변경하며 실험을 진행하였고 그 결과는 다음과 같다.

Table 5. 내용을 살펴보면 페널티가0.2에서 0.3으로 변할 때 Sensitivity 값은 0.400증가 하는 것을 관찰 할 수 있다. 이 외에 다른 값들의 최대 변경

Table 5. Detection result by penalty

	0.1	0.2	0.3	0.4	0.5
Accuracy	0.997	0.998	0.997	0.997	0.991
Sensitivity	0.298	0.557	0.957	0.957	0.957
Specificity	0.998	0.998	0.998	0.997	0.991
Precision	0.217	0.335	0.351	0.336	0.116

폭은 Accuracy와 Specificity가 0.006, Precision이 0.220으로 Sensitivity의 변동 폭보다 작다. 따라서 변동 폭이 가장 큰 sensitivity를 기준으로 가장 좋은 성능을 보인 penalty 0.3이 이 모델의 성능을 가장 좋게 하는 값이 된다.

앞선 실험을 종합하여 최종적으로 Threshold를 0.9로 하고 penalty를 0.3으로 하였을 때 시나리오별 탐지 결과는 다음과 같다. Detected는 실제 비정상행위가 탐지 된 수, Undetected는 비정상행위지만 탐지되지 않은 비정상행위의 수를 나타내는 값이다. Correct detection Rate는 존재하는 비정상 행위 중 탐지된 비정상 행위의 비율을 나타내는 항목이다.

Table 6.의 The day malicious act is performed는 악성 행위가 일어난 날을 의미한다. 각 시나리오마다 수 일에 걸쳐 악성행위를 수행하는 경우가 있기 때문에 Table 6.과 같은 값이 나오게 된다. 예를 들어 시나리오 1의 내부자는 30명 이지만 몇몇 내부자는 수 일에 걸쳐 위협 행위를 수행한다. 실험 결과 시나리오 1처럼 단순히 평소에 하지 않던 행위를 수행하여 위협행위를 하는 것에 대해서는 아주 높은 탐지율을 볼 수 있었고, 3번과 4번처럼 다른 사람 아이디를 이용하거나 다른 사람의 pc에 접속하여 수행하는 비정상 행위도 탐지해 낼 수 있었다. 시나리오 2번과 같은 경우는 특정 행위를 평소보다 더 많이 하여 위협행위를 하게 되는 시나리오인데, 이것 또한 penalty를 적용하여 탐지하는 것이 가능하다는 것을 알 수 있었다. 하지만 시나리오 5번의 경우 파일을 웹에 올리는 행위로 우리의 실험에서는 웹사이트에 접속하는 행위 외에 어떤 웹사이트에 접근하는지, 웹사이트에서 어떤 행위를 하

Table 6. Detection result by Scenario

Threshold	0.9		penalty		0.3	
Scenario	S1	S2	S3	S4	S5	
The day malicious act is performed	68	207	20	9	1	305
Detected	68	200	17	7	0	292
Undetected	0	7	3	2	1	13
Correct detection Rate(%)	100	97	85	78	0	96

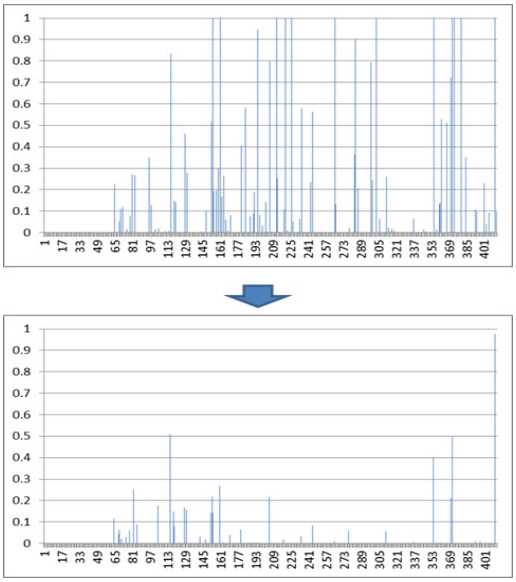

Fig. 11. Change of loss value after re-learning

는지 알 수 없기 때문에 탐지해 낼 수 없었다. 하지만 이는 더 자세한 지표를 이용한다면 충분히 탐지해 낼 수 있는 것으로 판단된다.

실험을 진행하면서 한번 정상 행위라고 판단된 비정상 행위를 다시 탐지해 내는 것은 성능상 문제가 될 수 있기 때문에 정상 행위라고 판단된 행위에 대해 일주일마다 다시 학습을 하는 방법으로 사용자의 최신의 행동 패턴을 학습 할 수 있도록 하였다. fig. 11.은 재학습 후 얻을 수 있는 loss 값의 변화를 나타낸 그래프이다. 일반 내부자의 경우 그림과 같이 loss 값이 평균적으로 줄어드는 것을 확인 할 수 있었다. fig 11.을 보면 0.9이상의 loss 값을 갖는 날이 재학습을 하지 않는 모델의 경우 14일인 반면 재학습을 수행하는 모델의 경우 1일인 것을 확인 할 수 있고, 악의적 내부자 탐지에는 거의 영향을 미치지 않는 것을 확인 할 수 있었다.

V. 결 론

본 논문에서는 지속적으로 발생하고 있는 유출사고의 주된 원인인 내부자 위협을 방지하기 위해 비정상행위를 탐지하는 방법에 대하여 연구를 진행 하였다. 기존의 HMM과 클러스터링에 의한 탐지 연구는 현재의 상태를 예측할 때 이전 상태에만 의존한 다는 모델 자체의 한계와 1주일 단위로 탐지를 진행하여 비정상행위 탐지가 낮다는 한계와 결과를 보고 직관적으로 비정상행위를 알아내기 어렵다는 한계가 있었다.

우리의 실험에서는 인공 신경망 모델 중 입력 값의 순서를 고려하여 학습을 하는 RNN의 속성을 이용한 Autoencoder를 구현하여 HMM보다 긴 의존성을 갖고 일단위로 내부자 위협을 탐지해내는 모델을 만들고, 모델을 통해 나온 값을 통해 쉽게 정상행위와 비정상행위를 구분할 수 있는 방법 대해 실험을 진행 하였다. 실험 데이터로는 내부자위협 연구에서 널리 사용되고 있는 CERT dataset을 이용하였다. Autoencoder의 입력 값과 출력 값이 같다는 성질을 이용해 사용자의 정상 행위 패턴을 학습한 후, 입력 값과 출력 값의 차이(loss)를 구하여 비정상 행위를 탐지해 냈다. 주 단위가 아닌 일 단위 탐지 결과로 95%이상의 민감도를 보였고, loss 값을 이용하여 직관적인 비정상행위의 분류에 성공하였다. 이는 기존의 연구와 비교하여 더 쉽고 빠르게 비정상행위를 탐지하여 조치를 취할 수 있다는 점에서 기존의 모델보다 좋은 성능을 갖고 있다고 말할 수 있다.

그러나 CERT dataset의 시나리오 5의 경우 파일을 웹에 올리는 행위로 우리의 실험에서는 웹사이트에 접속하는 행위 외에 어떤 웹사이트에 접근하는지, 웹사이트에서 어떤 행위를 하는지 알 수 없기 때문에 탐지해 낼 수 없었다. 향후에 본 논문에서 발생한 미탐과 과탐의 문제를 해결하기 위해 더욱 다양한 탐지 지표(파일의 중요도, 접속 pc, 누구에게 메일을 보냈는지, 어느 웹 사이트에 접속을 하였는지 등)를 이용하여 더 높은 성능을 나타내는 탐지 모델을 개발할 예정이며, 다른 기계학습 모델과의 결합을 고려하여 연구를 할 계획이다. 또한 특정 행위의 반복에 의한 악의적 행동에 대한 탐지를 penalty를 적용하는 방법 외에 학습 데이터 구성과 관련된 연구가 필요하며 반복적인 학습을 통해 과탐을 줄이는 방법 또한 고려해 볼 필요가 있다.

References

[1] M.B.Salem, S.Hershkop, and S.J.Stolfo. "A Survey of Insider Attack Detection research," Advances in Information Security, vol.39 pp 69-90, Aug. 2007

[2] Financial Security Institute, 2017 Top 10 issue report in Financial IT Security. Jan. 2017

[3] T.Rashid, I.Agrafiotis, and J.R.C. Nurse,

"A New Take on Detecting Insider Threats: Exploring the Use of Hidden Markov Models," Proceedings of the 8th ACM CCS International Workshop on Managing Insider Security Threats, pp. 47-56, Oct. 2016

〔4〕 P.Parveen and B.Thuraisingham, "Unsupervised incremental sequence learning for insider threat detection," Proceedings of 2012 IEEE International Conference on Intelligence and Security Informatics, pp.141-143, Jun. 2012

〔5〕 Hyun-Song Jang, "Data-mining Based Anomaly Detection in Document Management System," Journal of the Knowledge Information Technology and Systems(JKITS), 10(4), pp. 465-473, Aug. 2015

〔6〕 Jun-hong Kime, Min-sik, Hae-dong Kim, Su-hyun Cho, Phil-sung Kang, Dae-woo Lee, Kyung-ah Yang, and Ki-hun Kim, "Methodology about Insider Threat Detect Technic Using Anomaly Detection," Proceedings of the Korean Institute Of Industrial Engineers(KIIE) Fall Conference, pp 1217-1249, Nov. 2016

〔7〕 O.Bradiczka, J.Liu, B.Price, J.Shen, A.Patil, R.Chow, E.Bart, and N.Ducheneaut, "Proactive Insider Threat Detection through Graph Learning and Psychological Context," Proceedings of the 2012 IEEE Symposium on Security and Privacy Workshops, pp.142-149, May, 2012

〔8〕 Department of Homland Security(DHS), "Combating the Insider Threat," May. 2014

〔9〕 Young-Hwan Lim, Jun-Suk Hong, Kwang Ho Kook, and Won Hyung Park, "A Study on Insider Behavior Scoring System to Prevent Data Leaks," Jouranl of the Information and Security, 15(5), pp.77-86, Sep. 2015

〔10〕 Insider Threat Tools - The CERT Divisio n. [Online]. Available: "https://www.cer t.org/insider-threat/tools/"

〔11〕 Bong-Goo Park, "Anomaly Detection Performance Analysis of Neural Networks using Soundex Algorithm and N-gram Techniques based on System Calls," Journal of the Internet Computing and Services(JICS), 6(5) pp. 45-56, Oct. 2005

〔12〕 M.Goldstein and S.Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data," PLOS ONE vol.11 no.4 http://dx.do-i.org/10.1371/journal.pone.0152173 Apr. 2016

〔13〕 H.Kaur, G.Singh, and J.Minhas, " A Review of Machine Learning based Anomaly Detection Techniques," Journal of Computer Applications Technology and Research vol.2-issue 2, pp 185-187, Jul. 2013

〔14〕 X.Xu, Machine Learning for Sequential Behavior Modeling and Prediction. Machine Learning, Abdelhamid Mellouk and Abdennacer Chebira (Ed.), InTech. Jan. 2009

〈저자소개〉

하 동 욱 (Dong-wook Ha) 학생회원
2016년 2월: 명지대학교 컴퓨터공학과 졸업
2016년 3월~현재: 명지대학교 보안경영공학과 석사 과정
〈관심분야〉 시스템보안, 기계학습, 내부자위협, 정보보호

강 기 태 (Ki-tae Kang) 학생회원
2016년 2월: 명지대학교 컴퓨터공학과 졸업
2016년 9월~현재: 명지대학교 보안경영공학과 석사 과정
〈관심분야〉 시스템보안, 빅데이터, 내부자위협

류 연 승 (Yeonseung Ryu) 종신회원
1990년 2월: 서울대학교 계산통계학과 학사
1992년 2월: 서울대학교 계산통계학과 전산과학 석사
1996년 8월: 서울대학교 계산통계학과 전산과학 박사
1996년 9월~2000년 8월: 삼성전자 선임연구원
2003년 3월~현재: 명지대학교 컴퓨터공학과 교수
2014년 9월~현재: 명지대학교 대학원 융합보안학과 교수
2015년 3월~현재: 명지대학교 대학원 보안경영공학과 교수
〈관심분야〉 시스템보안, 방산보안, 국방SW