Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.29, NO.2, Apr. 2019 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2019.29.2.321

제조공정 단말PC 작업자 접속 로그를 통한 이상 징후 탐지 모델 연구

안 종 성,† 이 경 호‡
고려대학교 정보보호대학원

A Study on Anomaly Detection Model using Worker Access Log in Manufacturing Terminal PC

Jong-seong Ahn,† Kyung-ho Lee‡
Graduate School of Information Security, Korea University

요 약

기업에서 내부자에 의한 기업 기밀 유출 방지는 기업의 생존을 위한 필수 과제이다. 내부자에 의한 정보유출 사고를 막기 위해 기업에서는 보안 솔류션을 도입하여 적용하고 있으나 접근 권한이 있는 내부자의 이상행위를 효과적으로 탐지하는 데에는 한계가 있다. 이번 연구에서는 기업의 제품 제조 이력, 품질 정보 등을 담고 있는 제조정보시스템의 작업자 작업화면 접근 로그 데이타를 기계학습 기법의 비지도학습 알고리즘을 활용하여 정상적인 접근 로그와 비정상적인 접근 로그를 효과적으로 군집화하는 방법을 연구하여 이상징후 탐지를 위한 최적화된 속성 선택 모델을 제시하고자 한다.

ABSTRACT

Prevention of corporate confidentiality leakage by insiders in enterprises is an essential task for the survival of enterprises. In order to prevent information leakage by insiders, companies have adopted security solutions, but there is a limit to effectively detect abnormal behavior of insiders with access privileges. In this study, we use the Unsupervised Learning algorithm of the machine learning technique to effectively and efficiently cluster the normal and abnormal access logs of the worker's work screen in the manufacturing information system, which includes the company's product manufacturing history and quality information. We propose an optimal feature selection model for anomaly detection by studying clustering methods.

Keywords: Machine Learning, Anomaly Detection, Feature Selection

I. 서 론

기업의 내부시스템에 접근 권한이 있는 내부자에 의한 기밀 유출은 기업의 존망을 결정할 수도 있는 중대한 사안이다. 2016년 중소기업청 기술통계조사 보고서에 따르면 기술 유출 사고 중 전·현직 임직원에 의한 기술 유출이 전체의 68%에 이른다고 하였으며[1] 미국 Vormetric사의 2015년 Insider Threat Report에 따르면 설문 응답자의 93%가 본인 조직이 내부자의 보안 위협에 취약하다고 응답하였고 그 중 59%가 접근 권한이 있는 사용자가 위협 요인이라고 했다[2].

내부자는 조직 내에서 조직의 핵심 정보가 무엇이고 어디에 있는지를 알고 있어 항상 정보유출의 가능

Received(10. 18. 2018), Modified(02. 18. 2019), Accepted(02. 18. 2019)
† 주저자, jsan6131@korea.ac.kr
‡ 교신저자, kevinlee@korea.ac.kr(Corresponding author)

성이 상존한다. 악의적인 목적을 가진 내부자에 의한 정보 유출은 기업에 막대한 손실을 줄 수 있어 기업에서는 문서중앙화, 보안파일서버 솔류션 도입 등 다양한 보안 해결책을 모색하고 있으나 접근권한이 있는 내부자의 이상행위 탐지에는 한계가 있다.

이번 연구에서는 사이버 보안 위협의 주요 포인트로 대두되고 있는 엔드포인트 중 제조현장 작업자 단말PC를 통해 발생할 수 있는 내부자에 의한 보안 위협에 대응하기 위해 제조공정 작업자의 작업 화면 접근 로그를 연구 데이터로 활용하여 내부 작업자에 의한 이상징후 접근로그를 탐지하고자 한다. 또한 이번 연구에서는 데이터 마이닝 기법의 기계학습 방법을 사용한다. 데이터 마이닝을 이용한 기계학습의 가장 큰 장점은 대량의 데이터 집합에서 손쉽게 일정한 패턴을 찾아내고 그 패턴속에서 이상징후 데이터를 쉽게 분류할 수 있다는 점이다. 작업자가 제조시스템에 로그인하여 기록하는 유저 ID, 작업그룹, 작업장소, 접근화면 ID, 접속 IP, 접속 시간 데이터를 기계학습의 비지도학습 기법에 의한 군집화 알고리즘을 활용하여 정상적인 로그 데이터 군집과 비정상적인 로그 데이터 군집을 데이타 전처리, 속성 선택(Feature Selection), 적정한 군집개수 선택 과정을 통해 효과적으로 분류하는 방안을 연구하여 제조공정 작업자 접근 로그에 대한 이상징후 탐지모델을 제시하고자 한다.

이번 연구의 구성은 다음과 같다. 2장에서는 관련 연구로 문헌연구, 기계학습, 군집화에 대해 학습한다. 3장에서는 연구 데이터에 대해 분류하고 접근로그의 유형에 따라 작업그룹별, 특정 User 및 IP 별로 기계학습에 의해 로그 패턴을 분류하고 데이터 전처리를 수행한다. 4장에서는 비지도학습에 의한 이상징후 탐지 모델을 세우고 선택된 속성별로 순차적으로 4개의 비지도학습 알고리즘을 수행하여 이상치 데이터에 대한 군집화 결과를 얻는다. 5장에서는 제안된 모델에 대한 검증을 실시한다. 6장에서는 연구 결과에 대한 결론을 도출한다.

II. 관련 연구

2.1 문헌 연구

장현송은 기업의 중요한 지적자산을 보호하기 위해 내부 인증된 사용자의 비정상 행위를 탐지할 수 있는 방법으로 사용자의 시스템 사용 패턴을 알 수 있는 주요 변수들을 파악하고 그에 따라 K-Means 및 SOM(Self-Organization Map) 알고리즘을 활용하여 데이터를 군집화하는 데이터 마이닝 기반의 비정상 행위 탐지 모델을 제안하였다[3]. 권영백은 사전에 등록된 패턴으로 탐지되지 않아 정상적인 이용으로 분류되는 행위 이벤트들을 분석하여 이상징후를 탐지하고 이상징후 분석은 CBR(Case-Based Reasoning)을 활용한 이상징후 탐지 모델 제시를 통해 침해시도 대응 방법을 제시하였다[4]. 김해동은 내부 구성원이 IT 시스템을 사용할 때 기록되는 로그 정보로부터 사용자의 행위를 일 단위로 인스턴스화하여 특정 기간 동안에 한 사용자의 여러 가지 행위를 발생 빈도로 요약하여 수치화된 벡터로 표현하는 사용자 행위 모델링 기법을 제시하였다[5]. 이진호는 비지도학습의 신뢰성과 정확도를 향상시키기 위해 SMOTE(Synthetic Minority Oversampling Technique)를 사용하여 다양한 데이터 셋을 학습하였으며 비지도학습이 대량의 데이터를 비용과 자원을 최소한으로 투자하여 가공하는데 효율적이라는 것을 기계학습을 통해 제시하였다[6]. Pallabi는 내부침입자 탐지를 위해 비지도학습 앙상블 기반의 학습을 수행하여 내부 침입정보를 담고 있는 데이터 스트림들의 분류 정확도를 향상시키는 방안을 제시하였다[7]. Eldardiry는 정상적인 소스와 비정상적인 소스가 함께 혼합되어 있는 실제 작업 데이터셋을 기계학습 알고리즘을 통해서 이상치를 확인하여 추적하였으며 자신이 속한 집단과 다른 행동을 보이는 이상행위에 대한 이상징후 탐지의 정확성과 유연적응성(robustness)을 향상시키는 결과를 제시하였다[8].

2.2 기계 학습

데이터 마이닝은 데이터베이스에 잠들어 있는 수많은 양의 데이터로부터 그 데이터의 잠재성을 밝혀 내어 유용한 의미를 가진 정보를 추출하는 것이며 기계학습은 데이터 마이닝의 기술적인 기반을 제공한다. 기계학습에는 지도학습(Supervised Learnig), 비지도학습(Unsupervised Learning), 강화학습(Reinforcement Learning)이 있다. 지도학습은 데이터가 무엇을 의미하는지 미리 알수 있게 학습 정보에 대한 라벨링이 되어 있어 분류된 데이터로 학습을 수행하고, 비지도학습은 라벨링이 되어있지 않아 데이터에 대한 아무런 정보가 없이 숨겨진 구도속에서 학습을 진행하며, 강화학습은 어떤 환경에서 특정

목표를 달성하기 위해 스스로 학습한다[9].

이번 연구에서는 비지도학습 군집화 알고리즘 4개를 선택하여 학습을 수행하여 정보시스템 접근로그 데이터의 사용자 행위에서 이상징후 데이터 탐지를 위한 최적화된 비지도 기계학습 속성 선택 모델을 제시한다.

2.3 군집화

군집화는 비지도학습의 모형으로 입력 데이터의 값들이 비슷한 성질이 있는 것들끼리 하나의 군집으로 묶어 주는 방법이다. 개체간의 거리나 유사성을 기준으로 군집화를 이루고 이루어진 군집들 간의 관계를 분석하여 특성을 파악하게 된다. 군집분석의 목적은 아무런 관계가 없는 자료에서 특정 유형의 군집을 찾아내는 것으로 군집의 형성 과정과 특성, 식별된 군집간의 관계 등을 효과적으로 분석하는 것이다[10].

군집화 알고리즘은 중심을 기반으로 하는 알고리즘과 밀도를 기반으로 하는 알고리즘으로 나눌 수 있다. 중심 기반 알고리즘은 '동일한 그룹에 속하는 데이터는 어떠한 중심을 기준으로 분포될 것이다'라는 가정을 기반으로 하는 방법이고 밀도 기반 알고리즘은 '동일한 그룹에 속하는 데이터는 서로 가깝게 분포할 것이다'라는 가정을 두고 동작한다[11].

III. 연구 데이터 및 전처리

3.1 연구 데이터

연구에 사용되는 데이터는 원자력연료를 설계·제조하는 K공공기관의 MTS(Material Tracking System, 튜브제조정보시스템)시스템의 접속 로그 데이터이다. MTS시스템은 K공공기관의 주요 정보시스템인 ERP(Enterprise Resource Planning), MES(Manufacturing Execution System) 시스템과 데이터 연계시스템을 통해 필요한 생산정보를 주고 받으며 원자력연료 핵심 부품중의 하나인 튜브를 제조하는 시스템이다. 이번 연구에 사용되는 MTS시스템 접속 로그 데이터는 1년 6개월(2017.1월~ 2018.6월)동안 작업자가 작업PC에서 Log-On하여 Log-Off까지 접속한 작업 화면ID를 기록한 로그이다.

USER_ID	EMP_GROUP	WORK_SITE	LOG_TYPE	WIN_ID	IP_ADDRESS	C_NAME	C_DATE
S*B*N	Technician	TSA_현장	LOGIN	SYS_LOG_010	1*2.1*.10*.5*	세정열처리_3	Jan-02-2017 06:42:04
S*B*N	Technician	TSA_현장	OPEN	SYS_BUL_010	1*2.1*.10*.5*	세정열처리_3	Jan-02-2017 06:42:05
S*B*N	Technician	TSA_현장	OPEN	SYS_BUL_010	1*2.1*.10*.5*	세정열처리_3	Jan-02-2017 06:42:05
S*B*N	Technician	TSA_현장	START	mnu	1*2.1*.10*.5*	세정열처리_3	Jan-02-2017 06:42:06
S*B*N	Technician	TSA_현장	OPEN	MPC_PROCESS1	1*2.1*.10*.5*	세정열처리_3	Jan-02-2017 06:42:27
HJ**OI	Technician	TSA_현장	LOGIN	SYS_LOG_010	1*2.1*.10*.6*	세정열처리_2	Jan-02-2017 07:00:15
HJ**OI	Technician	TSA_현장	OPEN	SYS_BUL_010	1*2.1*.10*.6*	세정열처리_2	Jan-02-2017 07:00:17

Fig. 1. MTS Access Log Data

3.2 연구 데이터 분류

2017년 1월 부터 2018년 6월 기간동안 K공공기관의 MTS시스템에 접근한 작업자 ID는 125개이다. 작업자 그룹은 기술직(Engineer)과 생산기술직(Technician)으로 구분하며, 기술직은 생산지시를 발행하고 생산현황을 확인하는 생산관리 업무 수행과 품질관리 업무 수행을 위해서 MTS시스템에 접근하며, 생산기술직은 작업현장에서 생산업무를 수행하면서 생산지시에 따른 생산실적, 자재 사용실적, 품질작업 수행 데이타를 입력하기 위해 MTS시스템에 접근한다. 로그 타입은 LOGIN, EXIT 등 4가지로 구분되며, 작업장소는 본사, 공장 현장 등 5가지로 구분된다. 작업자가 접근하는 작업화면은 124개이며, 작업 PC는 103대이다.

연구대상 데이터는 총 192,675건이며 기계학습에 의한 학습 전 분류 알고리즘에 의해 다음과 같이 데

Table 1. Source Data Classification

Column	Description	etc.
user_id	worker account	125ea
emp_group	worker group	2ea
log_type	access type	4ea
work_site	work place	5ea
win_id	access window id	124ea
ip_address	access ip address	103ea
c_name	computer name	
c_date	creation date	

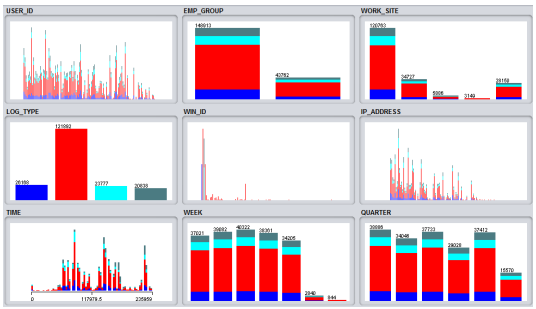

Fig. 2. Source Data Classification

이터가 분류되었다.

연구대상 데이터의 작업자 그룹별, 요일별, 분기별 데이터 분류는 아래 table 2, table 3, table 4와 같다.

table 2. 작업그룹별 데이터 분류에서 기술직(Engineer) User 계정은 37개이며 생산기술직(Technician) User 계정은 88개이다. table 3. 요일별 데이터 분류에서 토요일과 일요일의 접근 로그 수 평일보다 현저히 적음을 알고 있다. table 4. 분기별 데이터에서 2018년 2분기 접근 로그 수가 다른 분기에 비해 적은 것은 원자력발전소의 가동률 저하로 인해 원자력연료 생산공정의 가동율도 낮아졌기 때문이다.

Table 2. Work Group Unit Data

Work Group	Data Qty
Engineer	43,762
Technician	148,913

Table 3. Weekly Data

Week	Qty	Week	Qty
Sun.	844	Thu.	38,361
Mon.	37,021	Fri.	34,205
Tue.	39,082	Sat.	2,840
Wed	40,322	-	

Table 4. Quarterly Data

Quarter	Qty	Quarter	Qty
2017.1Q	38,886	2017.4Q	29,028
2017.2Q	34,046	2018.1Q	37,412
2017.3Q	37,733	2018.2Q	15,570

3.3 데이터 전처리

기계학습을 하기 위해서는 데이터 분석이 필요하다. 데이터 분석을 통해 어떠한 방향으로 데이터 전처리를 수행해야 기계학습의 효과를 최대화 할 수 있을지 먼저 연구되어야 한다. 데이터 전처리는 불필요한 정보를 제거하고 데이터를 기계학습에 사용할 수 있는 형태로 바꾸는 과정이다. 이번 연구 데이터는 기계학습 알고리즘이 이해할 수 있도록 데이터 전처리가 수행하였으며, 범주형 데이터를 이진 특성의 수치형 데이터로 변환해 주는 방식으로 해당하는 데이터만 1로 변경해 주고 나머지는 0으로 채워주는 방법으로 데이터 전처리를 수행하였다. 이번 연구의 데이터 전처리내용은 다음과 같다.

Table 5. EMP_GROUP

EMP_GROUP	Value
Engineer	0
Technician	1

작업 장소는 대전지역 A공장, 논산지역 B공장, 대전본사의 사무실과 현장으로 구분된다.

Table 6. WORK_SITE

Work_site	Variable
A_Factory_office	TSA1
A_Factory_site	TSA2
B_Factory_office	NSA1
B_Factory_site	NSA2
Head_office	HO

로그 유형은 LOGIN, OPEN, START, EXIT 4가지로 구분된다.

Table 7. LOG_TYPE

LOG_TYPE	Variable
LOGIN	LT1
OPEN	LT2
START	LT3
EXIT	LT4

IP Address는 IP 대역에 따라 근무장소가 구분되는 점을 활용하여 A공장 제조현장과 사무실, B공장, 본사 사무실로 구분된다.

Table 8. IP_ADDRESS

IP GROUP	Variable	Work_site
XXX.XXX.101.XX XXX.XXX.110.XX	IP101	A_Factory_site
XXX.XXX.102.XX	IP102	A_Factory_office
XXX.XXX.111.XX	IP111	B_Factory
Others IP	IPXX	Head_office et al.

데이터 생성일은 근무시간에 따라 table 9와 같다. 주간근무자인 기술직(Engineer)이 심야 시간인 3근 근무시간에 생성한 데이터 값을 이상치 데이터로 판단하기 위해 1로 처리하였으며 나머지는 정상 근무 데이터로 판단하여 모두 0으로 채워 주었다.

Table 9. Creation Date

Division	Time table	Value
Engineer Day	08:30~17:30	0 or 1
Technician 1	07:00~15:00	0
Technician 2	15:00~23:00	0
Technician 3	23:00~07:00	0

IV. 비지도학습에 의한 이상징후 탐지 모델

접근권한이 있는 내부자의 접근로그 데이터에 대해 이상치로 추정되는 유의미한 데이터를 군집화하기 위해 데이터 전처리를 통해 가변수(dummy variables)를 추가하고 Feature Selection을 통해 선택된 변수를 바탕으로 EM, K-Means, Canopy, FarthestFirst 총 4가지 비지도학습 군집화 알고리즘을 학습하여 내부자에 의한 이상행위 데이터 탐지 성능을 평가하였다.

4.1 비지도학습에 의한 이상징후 데이터 군집화 모델

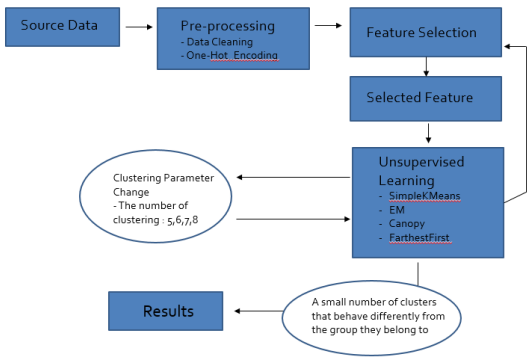

Fig. 3. Anomaly Detection Clustering Model for Unsupervised Learning

4.2 비지도학습 알고리즘

4.2.1 EM

기대치 최대화(Expectation maximization) 알고리즘으로 관측되지 않은 잠복변수에 의존하는 통계 모델에서 최대우도 또는 최대 사후 확률을 갖는 모수의 추정 값을 반복적으로 찾는 방법이다. 무작위 초기 모델을 생성한 다음 최적화 모델을 생성하기 위해 반복적으로 개선하는 프로세스를 수행한다[9].

4.2.2 K-Means

주어진 데이터를 k개의 군집으로 묶는 알고리즘으로, 각 군집과 거리 차이의 분산을 최소화하는 방식으로 동작한다. 각 개체와 K개 클러스터의 중심(Centroid)과의 거리를 산출하고, 개체와 중심간 거리가 가장 가까운 클러스터에 각 개체를 할당하며 일정한 조건을 만족할 때까지 새로 정해진 중심에 따라 할당이 반복적으로 이루어지는 알고리즘이다[9].

4.2.3 Canopy

모든 객체를 하나 이상의 캐노피에 할당하는 알고리즘으로 캐노피의 형성은 입력 변수 T1(캐노피와 캐노피 사이의 거리) 및 T2(캐노피 중심과 캐노피의 작은 원 사이의 거리)에 의해 결정된다. T2의 크기를 크게 설정하면 캐노피 총 수가 줄어들고 나중에 실행되는 군집화 알고리즘에 짧은 실행시간을 기대하기 어렵다. 또한 T2의 크기를 작게 설정하면 너무 많은 캐토피를 만들 수 있고 군집 결과를 변경할 수 있다. T1이 너무 크거나 작으면 복제된 캐노피의 수가 크거나 없으므로 결과에 영향을 준다. 따라서 적절한 T1과 T2의 값을 선택하는 것이 중요하다[9].

4.2.4 FarthestFirst

최장 거리 우선 탐색을 위한 군집화 알고리즘으로 기존 군집 중심에서 가장 먼 지점에서 각 군집 중심을 교대로 배치한 K 평균의 변형이다. 특성 값 및 빈도에 대한 정보를 기반으로 특성 수와 동일한 해시 테이블 구조를 구성하는 첫 번째 스캔과 해당 해시 테이블의 속성 값이 예상 시간 및 빈도에 따라 결정하는 두 번째 스캔을 하는 군집화 알고리즘이다[6].

4.3 군집화 실험 학습

이번 연구에 사용되는 기계학습 도구로 Weka를 활용하였다. 각 비지도학습 알고리즘에 의한 학습 데

이터 군집화 수행은 MTS시스템의 소스 데이타를 전처리한 후 의미있는 군집을 확인하기 위하여 군집의 개수를 5~8개까지 적용하여 군집화를 시도하였으며 그 결과 유의미한 군집을 확인 할 수 있는 7개의 군집으로 정하여 수행하였다.

4.3.1 속성 선택

속성 선택은 모두 7개의 유형으로 만들어 실험하였다. 전체 속성이 선택된 S1부터 로그 유형과 같이 이상치에 영향이 적은 변수를 차례로 제거하는 방식으로 S7까지 속성을 선택하여 비지도학습 알고리즘을 수행한다.

Table 10. Selected Feature

Division	S1	S2	S3	S4	S5	S6	S7
User ID	○	○	○	○	○	○	○
EMP_GROUP	○	○	○	○	x	x	○
TSA1	○	○	○	○	○	○	○
TSA2	○	○	○	○	○	○	○
NSA1	○	○	○	○	○	○	○
NSA2	○	○	○	○	○	○	○
HO	○	x	x	○	○	○	○
LT1	○	○	○	○	x	x	x
LT2	○	○	○	x	x	x	x
LT3	○	○	○	x	x	x	x
LT4	○	○	○	x	x	x	x
IP101	○	○	○	○	○	○	○
IP102	○	○	○	○	○	○	○
IP111	○	○	○	○	○	○	○
IPXX	○	x	x	○	○	○	○
w_time	○	○	x	x	○	x	○

4.3.2 군집화 학습 결과

Table 10의 S1의 속성 선택에 의한 비지도학습 알고리즘을 수행한 결과 각 군집별 데이터셋의 갯수와 군집율은 다음과 같다.

Table 11. Selection Feature S1 Clustering

Div.	K-Means		EM		Canopy		FarthestFirst	
	Qty	Per.	Qty	Per.	Qty	Per.	Qty	Per.
C0	24065	12.49%	15278	7.93%	120634	62.61%	120827	62.71%
C1	30537	15.85%	16927	8.79%	32537	16.89%	6684	3.47%
C2	75000	38.93%	23548	12.22%	28279	14.68%	5363	2.78%
C3	15614	8.10%	88560	45.96%	5190	2.69%	28478	14.78%
C4	8521	4.42%	28069	14.57%	3152	1.64%	2615	1.36%
C5	34508	17.91%	9001	4.67%	2190	1.14%	5843	3.03%
C6	4430	2.30%	11292	5.86%	693	0.36%	22865	11.87%

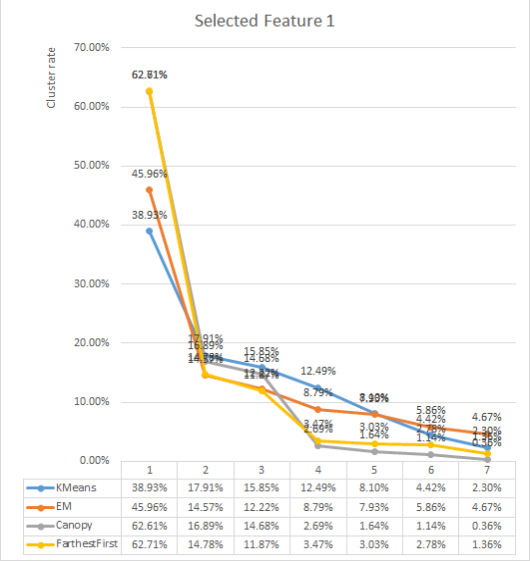

Fig. 4. Selected Feature S1 Result

Fig. 4는 S1의 속성 선택에 의한 비지도학습 알고리즘을 수행한 결과를 그래프로 나타낸 것이다. 4개의 비지도학습 알고리즘에 의해 생성된 군집 데이터의 군집율에 차이가 나타남을 알수 있다. 또한 table 10의 S1의 속성 선택에 의한 K-Means 알고리즘에 의한 군집별 대표 데이터 셋은 table 12와 같으나 유의미한 비정상 로그 데이터 셋을 찾을 수는 없었다. 이는 속성 15개 모두를 선택하여 최적화되지 않은 속성 선택이었기 때문에 나타난 결과로 판단된다.

Table 12. Dataset of K-Means about Selection Feature S1

Cluster 0: 1,0,0,0,1,0,0,0,1,0,0,0,1,0,0
Cluster 1: 1,0,1,0,0,0,0,0,0,1,1,0,0,0,0
Cluster 2: 1,0,1,0,0,0,0,1,0,0,1,0,0,0,0
Cluster 3: 1,0,1,0,0,0,0,0,1,0,1,0,0,0,0
Cluster 4: 0,0,0,0,0,1,0,1,0,0,0,0,0,1,0
Cluster 5: 0,1,0,0,0,0,0,1,0,0,0,1,0,0,0
Cluster 6: 1,0,0,0,1,0,1,0,0,0,0,0,1,0,0

Fig. 5. Selected Feature S2~S7 Result

Fig. 5는 table 10의 S2~S7의 속성 선택에 의한 비지도학습 알고리즘을 수행한 결과이다. 4개의 비지도학습 알고리즘에 의해 생성된 군집 데이터의 군집율이 S5~S7에서 거의 일치하여 나타남을 알 수 있다.

이는 최적화된 속성 선택에 의해 4개의 비지도학습 알고리즘의 이상치 데이터 탐지성능이 비슷하다는 것을 알 수 있다.

Fig 6은 속성 선택 S7에 의한 K-Means 시각화 그림이다. x축은 군집이고 y축은 user_id이다. 군

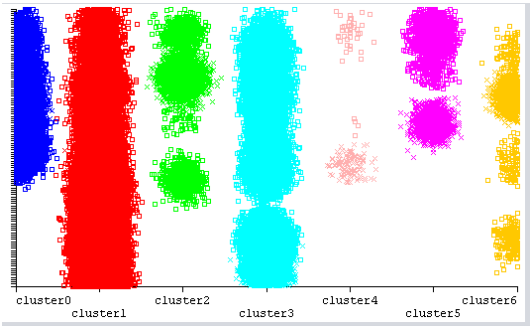

Fig. 6. K-Means Visual of Selected Feature S7

집 4가 최소 군집이며 데이터 군집률은 0.09%이다. 그 다음 최소 군집인 군집 6의 데이터 군집률은 1.13%이다.

4.4 이상징후 탐지 실험 학습 결과

Weka에서 제공하는 비지도학습 알고리즘에 의한 군집화 결과 소수 군집화를 이루는 군집의 특징은 자신이 평소에 근무하는 장소가 아닌 다른 IP 대역에서 접속한 소수의 접근 로그 기록들로 파악되었다. 이는 자신이 소속된 집단과 자신의 일상적인 행위와 다른 행위를 했다는 점에서 이상징후 데이터 군집으로 분류되었다.

Fig. 7의 그래프에 표시된 군집들은 7개의 선택된 속성으로 수행한 결과 중 가장 최소 군집을 이루는 군집들을 모아 순차적으로 놓은 것으로 비정상 데이터의 군집으로 판단되는 군집들이다.

4개의 군집화 알고리즘에 의한 군집화 결과 table 10의 S5, S6과 S7 선택 속성의 군집화 결과가 모

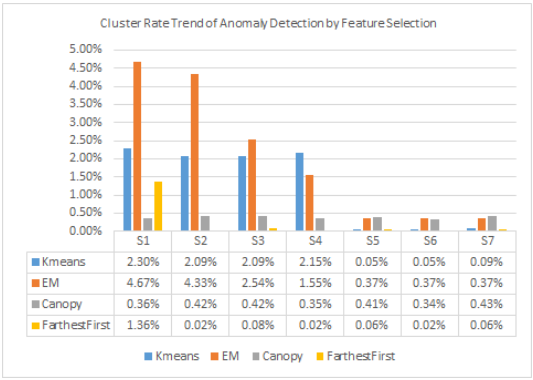

Fig. 7. Cluster Rate Trend of Anomaly Detection by Feature Selection

든 군집 알고리즘에서 비슷한 결과 값을 가져왔으며 1%에 근접하는 소수 군집 데이터들이 데이터 전처리에 의해 분류되는 정상 데이터셋에서 벗어난 이상치 데이터 군집으로 분석되어 가장 최적화된 선택 속성 모델로 판명되었다.

V. 검 증

제안한 이상징후 탐지모델을 검증하기 위해서 전체 학습 데이터 18개월분 중 임의로 선정한 3개 개월 데이터를 각각 분리하여 검증 데이터로 하였다. 검증방법은 table 10의 S7과 동일한 속성 선택으로 4개의 비지도학습 알고리즘을 수행하였으며 TP(True Positive) rate를 측정하였다. TP rate는 실제로 이상징후로 판단되는 데이타셋이 금번 연구 모델을 통해서도 이상징후로 판단되어진 데이터셋으로 옳게 예측한 비율이다.

검증 결과 EM 알고리즘은 2017.6월 79.0%, 2017.12월 92.9%, 2018.6월 77.5%의 이상징후 탐지 결과를 보여주어 이번 이상징후 탐지를 위한 최적화된 속성 선택 모델 검증에서 일정하게 좋은 탐지 결과를 얻었음을 알 수 있다.

Table 13. TP Rate

Div.	2017.6.	2017.12.	2018.6.
K-Means	59.9	77.7	50.7
EM	79.0	92.9	77.5
Canopy	77.8	85.9	77.5
FartherFirst	61.4	77.7	42.3

VI. 결 론

이번 연구에서는 K공공기관의 MTS(Material Tracking System, 튜브제조정보관리시스템) 시스템 접속로그 데이터를 이용하여 학습하였다. 기업의 제조공정 데이터는 내부자에 의해서든 외부 침입자에 의해서든 어떤 경우에도 외부에 유출되어서는 안되는 중요한 핵심 자산이다. 기업의 입장에서는 약 20여만건의 MTS 데이터 전체가 정상적인 데이터라 확신할 것이지만 정보보안을 학습하는 연구자 입장에서는 데이터를 여러 각도로 분석하여 이상징후 데이터를 가려낼 모델을 찾아 내는 것이 그 역할일 것이다.

이번 연구에서는 MTS 시스템 접속로그를 분석하

여 데이터 전처리 작업을 수행하였다. 비지도 학습에 의한 기계학습 알로리즘을 적용하기 위해 작업장소를 5개로 분류하고, 작업 IP 대역을 4개로 분류하였다. 데이터 전처리를 수행한 후 비지도 학습 알고리즘인 K-Means, EM, Canopy, FarthestFirst를 수행하였다. 이상치 데이터 군집화를 위하여 군집화 개수를 5~8개 까지 수행하였으며 군집의 개수를 7개로 했을 때 가장 적합한 이상치 데이터 군집화를 확인할 수 있었다. 이상치로 분류된 군집화 데이터를 분석한 결과 평소의 작업장소나 작업IP 대역을 벗어난 접속 로그이다. 제시된 이상징후 탐지모델로 3개 개월의 검증 데이터로 모델 검증 테스트를 수행하여 EM 알고리즘이 높은 탐지율을 보여 주었고 일정하게 좋은 탐지 결과를 얻었음을 확인하였다. 이에 기계학습의 비지도학습 알고리즘을 이용하여 MTS 시스템 작업자의 일상적인 사용자 행위 패턴을 벗어난 이상치 로그 기록을 군집 알고리즘으로 탐지하는 고도화된 이상징후 탐지모델을 제시하였다.

이번 연구에서는 실제 제조 현장의 접근로그 데이터를 이용하여 학습을 진행하였지만 연구의 한계도 있었다. 비지도학습 알고리즘에 의해 이상징후 데이터의 군집화를 확인할 수 있었으나 탐지율을 높이기 위해 기계학습에서 제공하는 유용한 알고리즘을 충분히 활용하지 못한 것이 아쉬운 점으로 다양한 관점에서 제조 공정 로그 데이터의 이상징후 탐지에 관한 연구가 향후 추가적으로 이루어져야 할 것이다.

References

[1] Small and Medium Business Administration, “2016 Technical statistics survey report for small and Medium businesses”, 2016.

[2] Vormetric, “Insider threat Repoert”, 2015

[3] Hyun-Song Jang, “Data-mining Based Anomaly Detection in Document Management System”, ISSN 1975-7700, 2015

[4] Young-baek Kwon, In-seok Kim, “A Study on Anomaly Signal Detection and Management Model Uing Big Data”, JIIBC, Vol.16, No. 6, pp.287-294, Dec. 2016

[5] Haedong Kim, “Insider Threat Detection based on User Behavior Model and Novelty Detection Algorithms”, Korea University, 2017

[6] Ho Jin Lee “Feature Selection Practice for Unsupervised Learning of Credit Card Fraud Detection”, Korea University, Feb. 2017.

[7] Pallabi Parveen, Nate McDanial, Varun S. Hariharan, “Unsupervised Ensemble based Learning for Insider Threat Detection“, IEEE 2012.

[8] Eldardiry, H., Sricharn,k.,Liu, j., Hanley,J., Price,B., Brdiczka, O., & Bart,E(2014). “Multi-source fusion for anomaly detection: using across-domain and across-time peer-group consistency checks“. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications* (JoWUA), 5(2),39-58

[9] Tae-ho Kim, “Feature Selection Optimization in Unsupervised Learning for Insider threat Detection“, Korea University, 2018

[10] Youn-Im Choi, “A Study on Improvement of K-means Clustering With Bisecting”, Chung-Ang University, Aug. 2011

[11] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining(KDD-96), AAAI Press. pp. 226-231, 1996

〈저자소개〉

안 종 성 (Jong-seong Ahn) 정회원
1996년 2월: 조선대학교 제어계측공학 학사
1996년 3월~현재: 한전원자력연료 ICT보안실 근무
2019년 2월: 고려대학교 정보보호대학원 사이버보안학 석사
〈관심분야〉 데이터 마이닝, 네트워크 보안, 서버 보안

이 경 호 (Kyung-ho Lee) 종신회원
1989년 2월: 서강대학교 수학 학사
1997년 2월: 서강대학교 정보통신학 석사
2009년 2월: 고려대학교 대학원 공학 박사
2011년: 고려대학교 정보보호대학원 조교수
2013년~현재: 고려대학교 정보보호대학원 부교수
2017년: 고려대학교 정보전산처장
〈관심분야〉 정보보호 정책, 개인정보보호 정책, 위험관리, 머신러닝, 블록체인