연구보고서 2018-16

기계학습(Machine Learning) 기반 이상 탐지(Anomaly Detection) 기법 연구

- 보건사회 분야를 중심으로

오미애 · 박아연 · 김용대 · 진재현

KIHASA
한국보건사회연구원
Korea Institute for Health and Social Affairs

【책임연구자】
오미애 한국보건사회연구원 연구위원

【주요 저서】
기계학습(Machine Learning) 기반 사회보장 빅데이터 분석 및 예측모형 연구
한국보건사회연구원, 2017(공저)
보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안
한국보건사회연구원, 2014(공저)

【공동연구진】
박아연 한국보건사회연구원 부연구위원
김용대 서울대학교 통계학과 교수
진재현 한국보건사회연구원 전문연구원

연구보고서 2018-16

기계학습(Machine Learning) 기반
이상 탐지(Anomaly Detection) 기법 연구
- 보건사회분야를 중심으로

발 행 일 2018년 12월
저 자 오 미 애
발 행 인 조 흥 식
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1994년 7월 1일(제8-142호)
인 쇄 처 ㈜한디자인코퍼레이션

ISBN 978-89-6827-540-1 93510

발간사

현 정부는 대통령 직속의 '4차 산업혁명위원회'를 설치하여 4차 산업혁명 대응 계획을 발표하고 있고, AI 등 성장 동력 기술력 확보로 신산업 육성을 통해 저성장 극복 방안을 모색하고 있다.

기계학습(Machine Learning)은 AI의 한 분야로 데이터를 바탕으로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이며, 이미지 처리, 영상 인식, 음성 인식, 인터넷 검색 등의 다양한 분야의 핵심 기술로 예측(Prediction) 및 이상 탐지(anomaly detection)에 탁월한 성과를 나타낸다.

이상 탐지(anomaly detection)란 자료에서 예상과는 다른 패턴을 보이는 개체 또는 자료를 찾는 것을 일컫는데, 이상값은 신용카드 사기, 사이버 침입, 테러 행위 같은 악의적 행동이나 시스템의 고장, 비정상적인 상황 등과 같은 이유로 발생하기 때문에, 실생활에서 이러한 위협 또는 고장으로 발생하는 피해를 방지하기 위해 이상 탐지는 필수적으로 해결해야 할 문제이다. 보건사회 분야 역시, 복지 대상 발굴과 부정 수급 모두 이상 탐지 기법 적용이 가능하다.

이상 탐지 기법은 각 분야 및 업무 성격에 따라 다르게 정의되고 적용될 수 있기에 여러 한계점도 존재하는데, 이 연구에서는 이상 탐지 연구의 체계적이고 포괄적인 개요를 제공하였다. 이상 탐지를 목적으로 진행된 다양한 연구를 살펴보고, 이상 탐지 관련 이슈 및 보건복지 분야에 활용성을 높일 수 있는 방안을 모색해 보고자 하였다.

본 보고서의 결과는 우리 연구원의 공식적인 견해가 아니라 연구진의 의견임을 밝혀 둔다.

2018년 12월

한국보건사회연구원 원장

조 흥 식

목 차

표 목차

그림 목차

Abstract

A Study on anomaly detection based on Machine Learning

Project Head: Oh, Miae

Artificial intelligence (AI) and big data analysis are the core technologies underlying the Fourth Industrial Revolution, and the self-sustained evolution of algorithms, based upon machine learning and big data, is key to all related progress. Machine learning, which is a part of AI, refers to the technology with which computers learn and adapt on the basis of large quantities of accumulated data. Machine learning holds the key to analytical and anomaly detection tasks required in a variety of fields, including image processing, video and voice recognition, and Internet search.

In data mining, anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text.

In this paper, we define the concept of anomaly detection and discuss various applications of anomaly detection techni-

Co-Researchers: Park, Ah Yeon · Kim, Yongdai · Jin, Jae Hyun

ques using machine learning techniques. We introduce the anomaly detection technique and compare the disadvantages of each methodology. We also study the anomaly detection study using Deep Learning machine learning method which is the latest machine learning method. We conduct exploratory analysis by applying the methodology of anomaly detection technique using data of health field and welfare field respectively. Finally, we deal with issues related to the application of anomaly detection techniques and conclude with policy.

By using anomaly detection techniques based on machine learning techniques in combination with fraud detection social security and improving budget efficiency, we can get closer to predictable customized welfare.

* Keywords : Machine Learning, Exploratory Data Analysis, Anomaly Detection

요 약

1. 연구의 배경 및 목적

기계학습(Machine Learning)은 AI의 한 분야로 데이터를 바탕으로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이며, 이미지 처리, 영상 인식, 음성 인식, 인터넷 검색 등의 다양한 분야의 핵심 기술로 예측(prediction) 및 이상 탐지(anomaly detection)에 탁월한 성과를 나타낸다.

이상 탐지(anomaly detection)란 자료에서 예상과는 다른 패턴을 보이는 개체 또는 자료를 찾는 것을 일컫는다. 이러한 개체를 우리말로는 이상, 이상값, 극단값, 예외, 비정상 등으로 부르고, 영어로는 anomaly, outlier, discordant observation, exception, aberration, surprise, peculiarity, contaminant와 같은 표현을 쓴다. 이상 탐지는 사기 탐지, 침입 탐지, 안전 필수 시스템(safety critical system), 군사적 감시를 포함한 다양한 분야에 널리 활용되고 있다.

이상값은 신용카드 사기, 사이버 침입, 테러 행위 같은 악의적 행동이나 시스템의 고장, 비정상적인 상황 등과 같은 이유로 발생하기 때문에, 실생활에서 이러한 위협 또는 고장으로 발생하는 피해를 방지하기 위해 이상 탐지는 필수적으로 해결해야 할 문제이다.

현재까지 앞에서 언급한 문제들을 해결하기 위한 수많은 이상 탐지 기법들이 개발되었다. 본 연구에서는 이상 탐지 연구의 체계적이고 포괄적인 개요를 제공하고자 한다. 이상 탐지를 목적으로 진행된 다양한 연구를 살펴보고, 이상 탐지 관련 이슈 및 보건복지 분야에 활용성을 높일 수 있는 방안을 모색해 보고자 한다.

2. 주요 연구 결과

이 연구에서는 이상 탐지의 개념을 정의하고, 기계학습 기법이 사용된 이상 탐지 기법의 다양한 활용 사례를 살펴보았다. 이상 탐지 기법에 대한 소개와 각 방법론별 장·단점도 비교해 보았다. 최신 기계학습 기법이라고 할 수 있는 딥러닝(Deep Learning) 기법을 적용한 이상 탐지 연구도 함께 살펴보았다. 보건 분야와 복지 분야 데이터를 각각 활용하여 이상 탐지 기법의 방법론을 적용해 봄으로써 탐색적 분석을 실시하였다. 마지막으로, 이상 탐지 기법 적용과 관련된 이슈를 다루고 정책 제언으로 마무리하였다.

본 연구의 구성에 따라 주요 내용을 요약하면 다음과 같다.

2장에서는 이상 탐지 개념을 정의하고 국내·외 활용 사례를 살펴보았다. 이상(anomaly)은 정상(normal)의 반대 개념이며 개념 정의를 위해서는 'normal'에 대한 정의부터 내려야 한다. '정상'에 대한 개념은 각 분야 및 문제마다 다르게 정의될 수 있기 때문에 '이상'에 대한 개념 역시 다 다르게 정의될 수 있다. 본 연구에서의 Anomaly detection은 각 분야 및 문제에서 정의된 'normal'의 반대 개념으로 '이상'을 정의하고 이를 찾아낼 수 있는 모형을 구축하는 것으로 개념을 정의한다. 이상 탐지는 입력 자료의 성질, 이상의 종류, 자료의 라벨, 이상 탐지 모형의 출력값 등의 여러 가지 요소들을 고려해야 한다. 국내·외 활용 사례는 연구 분야별 이상 탐지 기법 활용 사례와 데이터 유형별 이상 탐지 기법 활용 사례로 나누어 살펴보았다.

3장에서는 기계학습 기반 이상 탐지 기법으로 분류 기반 이상 탐지 기법, NN(nearest neighbor) 기반 이상 탐지 기법, 군집화(clustering) 기반 이상 탐지 기법, 이상 탐지의 통계적 기법, 정보 이론 이상 탐지 기법,

스펙트럴 이상 탐지 기법, 맥락적 이상 탐지로 나누어 살펴보았다. 딥러닝을 활용한 이상 탐지 기법으로는 차원 축소와 이상 탐지 방법을 동시에 학습하는 방법론을 소개하였다.

4장에서는 보건사회 분야 자료를 활용하여 앞에서 서술한 이상 탐지 기법을 적용해 보고, 활용 가능성을 검토해 보는 탐색적 분석을 실시하였다. 우리나라는 현재 고령사회로 진입하였기 때문에 향후 노인들에 대한 지원 및 정책이 확대될 수밖에 없다. 고령화 및 노령인구의 증가는 치매 노인 증가, 노인 학대 문제 등 돌봄 수요와 밀접한 관련이 있다. 이러한 차원에서 노인에 초점을 두어 보건 분야에서는 치매 조기 진단을 위한 자료(ADNI의 FDG-PET), 복지 분야에서는 노인 학대 노출 특성 분석을 위한 자료(2017 노인실태조사)를 활용하여 분석하였다. 보건 및 복지 분야 분석 개념도는 다음과 같다.

〈요약표 1〉 보건 분야 분석 개념도

내용	세부 내용	설명
Data	사용 데이터	ADNI의 FDG-PET 영상 자료(이미지 자료로 변환된 자료) + scalar 자료
개념	anomaly 개념	NCI군이 알츠하이머병으로 전환된 환자
	normal 개념	NCI군이 알츠하이머병으로 전환되지 않은 환자
프로세스	이상 탐지 기법 적용을 위한 자료 속성 파악	입력 자료 성질: 수치형 자료 + 이미지 자료 이상의 종류: point anomaly 자료 라벨: 지도 이상 탐지 모형의 출력값: 정상/이상 라벨 부여
	분석 방법 (사용한 이상 탐지 기법)	1. 스펙트럴 이상 탐지 기법: PCA 2. 분류 기반 이상 탐지 기법: Lasso
	분석 결과 제시	Accuracy Sensitivity Specificity AUC
의미	활용성	수치형 자료만 사용하는 것보다 이미지 자료와 함께 분석 시, 치매 조기 진단 예측력을 높일 수 있음. 이는 이상 탐지 기법에 정형 데이터뿐만 아니라, 비정형 데이터 분석도 중요함을 시사

〈요약표 2〉 복지 분야 분석 개념도

내용	세부 내용	설명
Data	사용 데이터	2017 노인실태조사
개념	anomaly 개념	- 정상(학대 경험 없음)인데 비정상(학대 경험 있음)으로 예측된 대상자(오분류) - 비정상(학대 경험 있음)인데 정상(학대 경험 없음)으로 예측된 대상자(오분류)
	normal 개념	- 정상(학대 경험 없음)인데 정상(학대 경험 없음)으로 예측된 대상자(정분류) - 비정상(학대 경험 있음)인데 비정상(학대 경험 있음)으로 예측된 대상자(정분류)
프로세스	이상 탐지 기법 적용을 위한 자료 속성 파악	- 입력 자료 성질: 연속형 자료 + 범주형 자료 - 이상의 종류: 맥락적 이상(contextural anomaly) - 자료 라벨: (맥락적 이상 개념을 적용한) 지도 이상 탐지 - 모형의 출력값: 정상/이상 라벨 부여
	분석 방법 (사용한 이상 탐지 기법)	1. 이상 탐지의 통계적 기법: 혼합 모수적 방법 2. NN(nearest neighbor) 기반 이상 탐지 기법: LOF 3. 스펙트럴 이상 탐지 기법: t-SNE 4. 군집화 방법(clustering): DBSCAN 알고리즘
	분석 결과 제시	1. 4개의 class로 나누어 이상치가 어디에 속하는지 특성 분석 2. 학대 경험 없는 집단에서 이상값 상위 100명 plot 및 특성 분석/학대 경험 있는 집단에서 이상값 상위 100명 plot 및 특성 분석 3. plot에 anomaly 표시 및 분석 4. plot에 clustering 결과 표시 및 분석
의미	활용성	- 분석 방법에 따라 대분류->중분류->소분류로 정의할 수 있을 것임 - 대분류를 한다면 1번 -〉 4번 방법으로 top down 설명 가능 - 소분류를 한다면 4번 -〉 1번 방법으로 bottom up 설명 가능 - 이는 특성에 대한 대분류 또는 상세 분류(세부 속성)를 하기 위해서는 단계별로 이상 탐지 기법을 적용하는 것이 바람직하다는 것을 시사

5장에서는 이상 탐지 기법과 관련된 이슈 및 정부 정책에서의 활용 가능성을 가늠해 보았다.

3. 결론 및 시사점

이 연구에서는 이상 탐지 문제가 만들어지는 과정들을 논의했고, 다양한 기법에 대한 논문들을 개략적으로 설명하고자 했다. 각 (기반 이론) 분야에 대해 정상과 이상 자료에 대한 특수한 가정을 살펴보았다. 이 가정들은 적용한 기법의 효율성을 평가하는 지침이 될 수 있다. 현재의 연구들은 통일된 이상치의 개념이 없는 채로, 체계가 없이 이루어져 이상 탐지 문제에 대한 이론적인 이해를 하기가 매우 어렵다. 여러 기법의 이상값 개념 및 가정을 통계적 틀이나 기계학습의 틀 안에서 통합하는 것이 앞으로의 가능성 있는 과제를 위해서도 필요하기에 방법론에 대해 자세히 소개하였다.

이상 탐지 기법은 학습 자료를 기반으로 기존의 자료들과는 다른 특성을 갖는 자료를 찾는 모형을 만드는 방법으로, 대부분의 자료가 정상 분류이고 극소수의 자료가 비정상 자료인 경우 비정상 자료의 탐지를 위해 사용하는 방법이다. 하지만, 비정상 자료가 극소수가 아닐 수 있으며, 비정상의 개념을 문제에 따라 재정의할 수 있다. 보건사회 분야 자료의 이상 탐지 기법에 대한 탐색적 분석에서 살펴보았듯이, 정형 데이터뿐만 아니라 이미지 자료, 영상 자료의 비정형 데이터도 이상치를 탐지하는 데 정확도를 높일 수 있다. 또한, 여러 이상 탐지 기법을 사용하여 대분류부터 소분류까지 단계적으로 구분 지어 활용할 수 있다.

기계학습(Machine Learning)에 기반한 이상 탐지 기법 연구는 효과적인 정책 수립 및 집행으로 공공·행정 부문에서 효율성 증대가 가능하다. 아동의 권익 증진을 위하여 장기 결석, 건강검진 미실시 정보 등의 빅데이터를 활용하여 학대 등 위기 아동을 조기 발굴할 수 있는 데이터 분석에서 기계학습에 기반을 둔 이상 탐지 기법을 적용할 수 있다. 아동 학

대의 사례처럼 아동 1,000명 중 1~2명이 학대 경험이 있다면 분류(classification) 문제로 접근하기 힘들다. 이런 imbalanced data에서는 이상 탐지 기법이 하나의 대안이 될 수 있다.

보건사회 분야에서 이상 탐지 기법이 가장 잘 활용될 수 있는 부분은 부정 수급 탐지이다. 부정 수급은 정부에서 지원하는 복지 혜택이나 복지 시설, 의료기관 등의 보조금을 더 받기 위해 수급 자격을 속이거나 입소자를 늘리는 등의 부정한 방법으로 복지 예산을 낭비하는 사례를 의미[1] 하는데, 부정 수급을 탐지하기 위해서는 부정 수급의 주요 유형별로, 시간적 흐름에 따른 패턴을 파악하고, 다른 자료와의 연계를 통해 통합적으로 자료를 살펴볼 필요가 있다.

이상 탐지 기법은 개인정보 보호 기법과도 밀접한 연관이 있다. 개인정보는 민감 정보이기 때문에 개인정보 보호 기법 활용은 보건사회 분야에서 특히 중요하다. 개인정보 보호 기법은 특정한 개인이 드러나지 않도록 통계적으로 마스킹(masking) 등의 처리를 해 주어야 하는데, 특정한 개인을 찾을 수 있는 방법으로 이상 탐지 기법을 활용할 수 있다. 개인정보 보호를 위한 통계적 기법은 데이터 활용 수요가 증가할수록 다양한 방법들이 개발될 것이다. 이상 탐지 기법 역시 그 활용성이 증가할 것으로 생각된다.

이를 위해서는 이상 탐지 기법의 다양한 방법론을 보건사회 분야의 정책 목적에 맞게 개발해야 하며, 정책 집행에 직접적으로 활용한 사례를 만들어서 정책 입안자 및 연구자들에게 공유할 필요가 있다. 공유의 확산 속도를 증가시키기 위해서는 데이터 및 분석 방법의 공개가 필수적일 것이다.

빅데이터 시대에 데이터의 활용가치는 증대되는 만큼, 최신 기법인 기

1) 복지로 홈페이지(https://www.bokjiro.go.kr/wrsd/guide1.do)

계학습 기법에 기반한 이상 탐지 기법을 정책 대상 발굴이나 예산 효율성 제고에 접목시켜 활용한다면 예측 가능한 맞춤형 복지에 한층 가까이 다가설 수 있을 것이다.

*주요 용어: 기계학습, 이상 탐지 기법, 탐색적 분석

제 1 장 서론

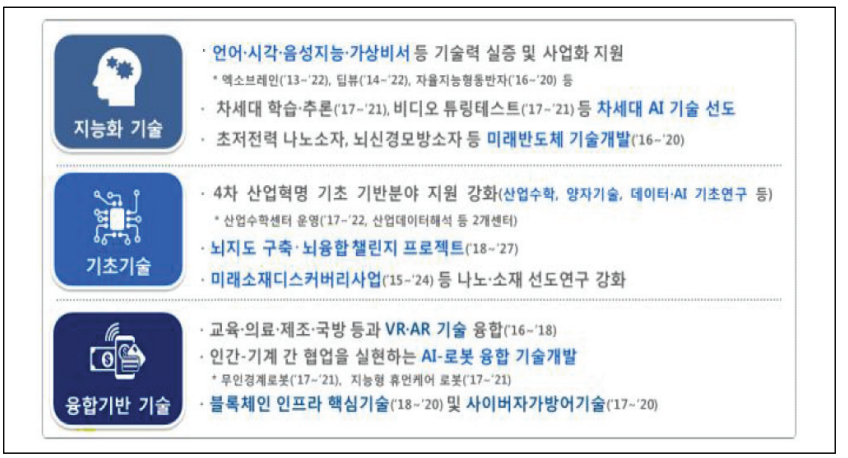
1 서론

제1절 연구 배경 및 목적

현 정부는 대통령 직속의 '4차 산업혁명위원회'를 설치하여 4차 산업혁명 대응 계획을 발표하고 있고, 인공지능 등 성장 동력 기술력 확보로 신산업 육성을 통해 저성장 극복 방안을 모색하고 있다.

4차 산업혁명 관련 기술은 지능화 기술, 기초 기술, 융합 기반 기술로 나눌 수 있으며 정부의 3대 중점 기술의 기술 개발 방향은 다음과 같다.

〔그림 1-1〕 정부의 3대 중점 기술의 기술 개발 방향



자료: 관계부처 합동(2017. 11). 4차 산업혁명 대응계획 p35.

기계학습(Machine Learning)은 AI의 한 분야로 데이터를 바탕으로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이며, 이미지 처리, 영상 인식, 음성 인식, 인터넷 검색 등의 다양한 분야의 핵심 기술로 예측(Prediction) 및 이상 탐지(anomaly detection)에 탁월한 성과를 나타낸다.

이상 탐지(anomaly detection)란 자료에서 예상과는 다른 패턴을 보이는 개체 또는 자료를 찾는 것을 일컫는다. 이러한 개체를 우리말로는 이상, 이상값, 극단값, 예외, 비정상 등으로 부르고, 영어로는 anomaly, outlier, discordant observation, exception, aberration, surprise, peculiarity, contaminant와 같은 표현을 쓴다. 이상 탐지는 사기 탐지, 침입 탐지, 안전 필수 시스템(safety critical system), 군사적 감시를 포함한 다양한 분야에 널리 활용되고 있다.

이상 탐지의 적용 분야는 사이버 보안, 의학 분야, 금융 분야, 행동 패턴 분야 등 다양하다(Kumar, 2005; Spence et al., 2001; Aleskerov et al., 1997; Liu et al., 2008). 사이버 보안의 경우, 컴퓨터 네트워크의 비정상적인 트래픽 패턴은 해킹당한 컴퓨터가 기밀 정보를 권한이 없는 경로로 전송함을 의미할 수 있다. 의학 분야에서는 MRI 이미지의 이상값은 악성 종양의 신호일 가능성이 있다. 금융 분야는 신용카드 거래 자료의 이상한 부분에서 신용카드 도난이나 신분 도용을 나타낼 수 있다. 복잡 시스템 관리 분야의 경우, 시스템 상태에 대한 시계열 자료 내 이상점은 시스템 내부의 고장이나 외부의 위협일 수 있다. 행동 패턴 분석 분야에서는 가정 내 또는 사업장의 CCTV 자료 내 이상 패턴이 도난이나 화재 같은 사고 발생을 의미할 수 있다.

이와 같이 이상값은 신용카드 사기, 사이버 침입, 테러 행위 같은 악의적 행동이나 시스템의 고장, 비정상적인 상황 등과 같은 이유로 발생하기

때문에, 실생활에서 이러한 위협 또는 고장으로 발생하는 피해를 방지하기 위해 이상 탐지는 필수적으로 해결해야 할 문제이다.

이를 보건사회 분야에서 생각해 본다면, 복지 대상 발굴과 부정 수급 모두 이상 탐지 기법 적용이 가능하다. 복지 서비스가 필요한 사람들을 발굴한다는 측면에서는 학대 등의 위기 아동을 발굴할 때 위기 아동의 비정상적인 상황을 정의하고 일반 아동과의 특성을 비교 분석할 때 이상 탐지 기법이 필요하다. 부정 수급 관련해서는 기초연금 수급, 어린이집 보육료 지원 수급, 장애인복지 수급, 실업급여 수급 등의 부정 수급을 방지하기 위해 각 부처마다 부정 수급 신고 사이트 및 부정 수급 방지 업무 처리 규정 등을 두고 있다. 부처 간 행정 자료 연계로 부정 수급을 탐지하기도 하지만, 이상 탐지 기법을 적용한다면 일반적인 패턴 분석 이상의 심층 분석이 가능할 것이다.

현재까지 앞에서 언급한 문제들을 해결하기 위한 수많은 이상 탐지 기법들이 개발되었다. 이상 탐지 기법은 각 분야 및 업무 성격에 따라 다르게 정의되고 적용될 수 있기에 여러 한계점도 존재하는데, 본 연구에서는 이상 탐지 연구의 체계적이고 포괄적인 개요를 제공하고자 한다. 이상 탐지를 목적으로 진행된 다양한 연구를 살펴보고, 이상 탐지 관련 이슈 및 보건복지 분야에 활용성을 높일 수 있는 방안을 모색해 보고자 한다.

제2절 연구 내용 및 방법

이 연구에서는 이상 탐지의 개념을 정의하고, 기계학습 기법이 사용된 이상 탐지 기법의 다양한 활용 사례를 살펴본다. 이상 탐지 기법에 대한 방법론 소개와 각 방법론별 장·단점도 비교해 본다. 최신 기계학습 기법

이라고 할 수 있는 딥러닝(Deep Learning) 기법을 적용한 이상 탐지 연구도 함께 살펴본다. 보건 분야와 복지 분야 데이터를 각각 활용하여 이상 탐지 기법의 방법론을 적용해 봄으로써 탐색적 분석을 실시한다. 마지막으로, 이상 탐지 기법 적용과 관련된 이슈를 다루고 정책 제언으로 마무리하고자 한다. 이 보고서는 모두 6개의 장으로 구성되어 있다.

이 보고서 작성을 위해, 국내·외 문헌 연구, 해외 사례 연구, 자료 분석, 국제 학회 자료집을 통한 해외 동향 파악 등 다양한 방법이 활용되었다.

데이터 탐색적 분석에서는 보건 분야의 경우 Alzheimer's Disease Neuroimaging Initiative(ADNI)의 FDG-PET(Fludeoxyglucose-Positron Emission Tomography), 복지 분야의 경우 한국보건사회연구원·보건복지부의 2017년 노인실태조사를 사용하였고, 기계학습 기반 이상 탐지 방법론을 적용하기 위하여 R 프로그램과 매트랩(Matlab)을 사용하였다.

제 2 장

이상 탐지 개념 정의 및 국내·외 사례 연구

2 이상 탐지 개념 정의 및 국내·외 사례 연구

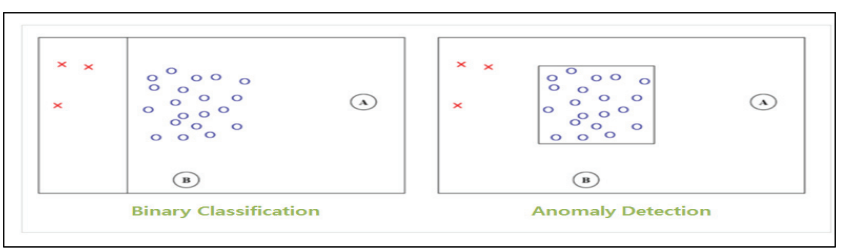
제1절 이상 탐지(Anomaly detection) 개념 정의[2)]

1. 이상 탐지(Anomaly detection) 개념 및 특성

'anomaly'는 정상(normal)의 반대 개념이며 개념 정의를 위해서는 'normal'에 대한 정의부터 내려야 한다. '정상'에 대한 개념은 각 분야 및 문제마다 다르게 정의될 수 있기 때문에 '이상'에 대한 개념 역시 다 다르게 정의될 수 있다.

본 연구에서의 Anomaly detection은 각 분야 및 문제에서 정의된 'normal'의 반대 개념으로 '이상'을 정의하고 이를 찾아낼 수 있는 모형을 구축하는 것으로 개념을 정의한다.

[그림 2-1] Classification과 Anomaly Detection 차이



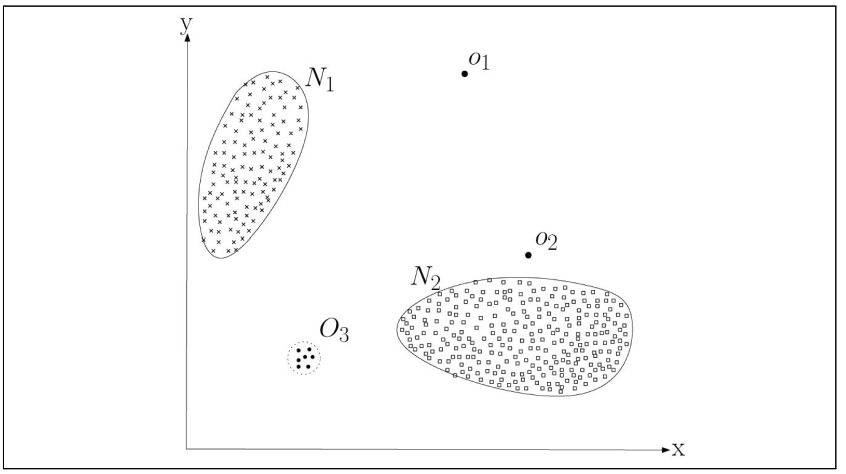
자료: 정재윤. (2017). Novelty Detection-Overvier. https://jayhey.github.io/novelty%20detection/2017/10/18/Novelty_detection_overview/〉에서 2018년 11월 인출

2) 이상탐지 개념의 예시 및 특성은 [Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15] 논문 내용을 번역하여 보고서 구성에 맞게 재구성하였음

classification은 두 범주를 구분할 수 있는 경계면을 찾는 것이라고 하면, Anomaly detection은 다수 범주를 고려하며 이상치가 아닌 데이터들의 sector를 구분 짓는 것이라고 볼 수 있다.

이상값은 자료 내 잘 정의된 정상적인 패턴을 따르지 않는 개체를 말한다. 〔그림 2-2〕는 2차원 자료의 이상값을 보여 준다. 먼저 자료 내 대부분의 점을 포함하는 두 정상 영역 N_1, N_2가 있고 정상 영역에서 멀리 떨어진 점 o_1, o_2와 영역 O_3 내 점들이 이상값이다.

〔그림 2-2〕 2차원 자료에서의 이상치 예시



자료: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15. 2page

잡음(noise) 제거(Teng et al., 1990)와 잡음 조정(Rousseeuw & Leroy, 1987)은 이상 탐지와 관련은 있지만, 분명한 차이점이 있다. 잡음은 분석가들이 관심을 가지는 대상이 아니며, 단지 분석에 방해가 될 뿐이다. 잡음 제거는 분석하기 이전 단계에서 불필요한 개체 또는 성분을 없애는 것이 목적이며, 잡음 조정은 통계적 모형 추정에서 이상값의 영향

을 줄이기 위한 것이다(Huber, 2011).

다른 연관 분야로 신상 탐지(novelty detection)(Markou & Singh, 2003a, 2003b; Saunders & Gero, 2000)가 있는데, 이는 지금까지 발견되지 않았던 새로운 패턴을 찾아내는 것을 목표로 한다. 신상 탐지는 이상 탐지와 달리 찾아낸 패턴을 정상에 포함한다. 잡음 제거, 조정 또는 신상 탐지 분야에서 주목할 점은 이러한 연관 주제에서 나온 해결책을 이상 탐지 분야에서 이용하고, 그 반대도 마찬가지라는 점인데, 이와 관련해서는 추후 논의하겠다.

이상 탐지는 시간의 특성을 가지는 시간 자료 분야에서도 연구되었다. 시간을 맥락적 변수(contextual variable)로 보았을 때, 시간 자료 내 이상 탐지 문제의 특수한 성질은 다음과 같다.

- 시간 자료에서는 시간의 연속성이 존재하여 특정 시점이 그 시점 전, 후의 값에 크게 영향을 받는다. 일반적으로 시간의 작은 창(window)을 적절히 선택하여 분석을 진행한다. 반면에 일반적인 자료에서는 시간적 특성에 영향을 받지 않기 때문에 개체들의 독립성을 가정하고 데이터마이닝/머신러닝 방법을 사용하여 이상값을 탐지한다.
- 시간 자료에서 비정상적인 시점을 찾는 것을 목표로 하느냐, 비정상정인 변화의 패턴을 찾는 것을 목표로 하느냐에 따라 분류된다.
- 시간 자료라도 연속형, 이산형, 고차원 스트림 혹은 네트워크와 같은 자료 성질에 따라 서로 다른 분석 기법이 필요하다.
- 과거 자료의 이상값에 대한 라벨이 이용 가능한지의 여부에 따라 비지도 vs. 지도 방법으로 분류된다.

2. 이상 탐지의 여러 요소

가. 입력 자료의 성질

입력 자료는 자료 개체(instance)들의 모임이고(Tan et al., 2005), 각 개체는 하나 이상의 속성 또는 변수(attribute)들로 표현된다. 자료 내 변수들의 성질에 따라 적용할 수 있는 이상 탐지 기법 또한 달라진다. 예를 들어, 대부분 통계 모형은 연속형이나 범주형 자료에만 적용할 수 있고, k-NN 기반 기법들은 자료 개체 간 거리의 정의가 추가적으로 필요하다. 특히 자료의 실제 값 대신 거리 행렬이나 유사도 행렬과 같은 형태로 자료 개체 간 거리만 주어지는 경우에는 원래의 자료 값이 필요한 대부분의 통계적 기법들을 적용하기 어렵다. 또한 자료 내 개체들이 서로 상관관계가 있는지에 따라 입력 자료의 특성이 분류된다(Tan et al., 2005). 대부분 방법론은 자료 개체들 사이에 관련이 없다고 여겨지는 레코드 자료(record data)나 점 자료(point data)를 대상으로 하고 있지만, 순차 자료(sequence/sequential data), 공간 자료(spatial data), 그래프 자료(graph data)와 같이 자료 개체가 상호 관계를 갖는 자료 형태에 대해서도 연구되어 왔다.

나. 이상의 종류

1) 점 이상(point anomaly)

자료 내 하나의 개체가 나머지에 대해 이상하다고 판단되는 경우를 일컫는다. [그림 2-2]의 점 o_1, o_2와 영역 O_3 내 점들처럼 일반적인 영역에 포함되지 않는 점들이 그 예시이다.

2) 맥락적 이상(contextual anomaly)

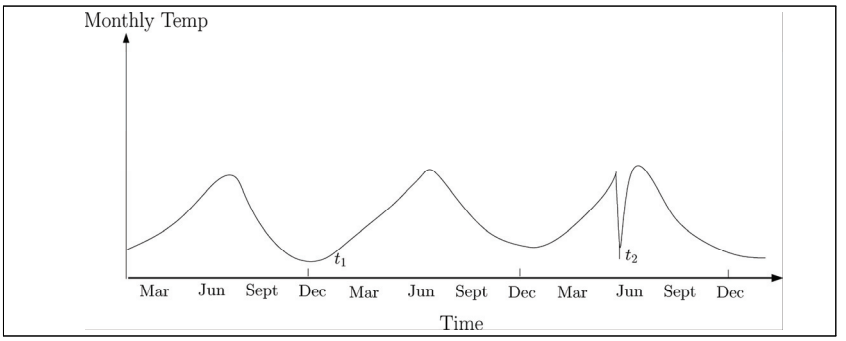
자료 내 개체가 특정 맥락에서 이상하다고 판단되는 경우를 말한다. '맥락'의 개념을 명확히 정의하기 위해, 맥락 변수와 행동 변수를 다음과 같이 정의한다.

- 맥락적 속성 또는 맥락 변수(contextual attribute)는 맥락 또는 그 근방(neighborhood)을 결정한다. 예를 들면 공간 자료의 위도나 경도, 시간 자료의 시간 등이 맥락 변수이다.
- 행동적 속성 또는 행동 변수(behavioral attribute)는 맥락적이지 않은 특성(characteristic)을 나타낸다. 예로 평균 강우량 자료가 있을 때, 각 위치에서의 강우량이 이에 해당한다.

이상 여부는 특정한 맥락에서 행동 변수의 값으로 판단한다. 맥락적 이상은 시간 자료(Weigend et al., 1995; Salvador & Chan, 2003)와 공간 자료(Kou et al., 2006; Shekhar et al., 2001)에서 가장 흔하게 찾을 수 있으며, 〔그림 2-3〕이 그 예시이다. 겨울(t_1)에 기온이 화씨 35도(섭씨 약 1.7도)인 것은 그럴 만하지만, 여름(t_2)에는 매우 이상한 상황일 것이다.

대상이 되는 영역에서 맥락적 이상이 얼마나 의미가 있는지와 맥락적 속성을 쉽게 구분할 수 있는지에 따라 맥락적 이상 탐지 기법 적용 여부가 결정된다.

[그림 2-3] 기온에 대한 시계열 자료에 대한 예



자료: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15. 8page

주: t_2는 맥락적 이상치이고, t_1과 t_2의 기온은 같지만 다른 맥락에서 나오고, t_1에 대한 맥락에서는 정상이기 때문에 t_1은 이상으로 간주하지 않는다.

다. 자료 라벨

라벨은 자료 개체의 이상 여부를 나타낸다. 전체 학습(training) 자료에 정확히 이상 여부에 대한 라벨을 부여하는 것(분류)은 엄청난 노력과 비용이 발생하며, 특히 모든 가능한 종류의 이상을 분류하는 것은 훨씬 더 어렵다. 게다가 이상이 매우 드물게 나타나거나 새로운 종류의 이상이 등장했을 경우, 거기에 분류된 자료 개체를 구하는 것은 불가능에 가깝다. 따라서 라벨에 대한 정보가 없는 자료를 다룰 필요가 있고, 그 정도에 따라 세 가지로 나눌 수 있다.

- 지도 이상 탐지(supervised anomaly detection)

학습 자료 내 모든 개체에 라벨 정보가 있을 때 쓰이는 방법으로, 정상 또는 이상을 판단하는 분류 모형을 학습시키는 것이 가장 일반적인 접근 방법이다. 보통 자료가 정상에 비해 이상의 비율이 매우 작은 불균형(imbalanced)한 상태에 있고, 앞서 언급했듯이 정확한 분류가 어렵다는

것이 특징인데, 이러한 점들을 제외하면 예측 모형을 세우는 과정과 비슷하기 때문에 이 연구에서는 더 다루지 않는다.

- 준지도 이상 탐지(semi-supervised anomaly detection)
  학습 자료 중 정상 개체에만 라벨 정보가 있고 라벨 정보가 없는 자료에 대해 정상/비정상 여부를 알 수 없는 경우 사용하는 기법으로 지도 이상 탐지보다 넓은 범위에 적용할 수 있다. 일반적으로 정상 자료만을 사용하여 모형을 학습시킨 뒤 시험 자료에 적용하는 방식을 이용한다.
- 비지도 이상 탐지(unsupervised anomaly detection)
  라벨이 없는 자료에서 사용하는 이상 탐지 방법으로, 가장 널리 쓰일 수 있는 기법이다. 주로 자료 내 개체들 간의 거리를 기반으로 하여 이상값을 탐지한다. 일반적으로 정상의 비율이 압도적으로 크다는 가정을 하는데, 이 가정이 틀리면 높은 오경보율(false alarm rate)과 같은 문제가 발생한다.

라. 이상 탐지 모형의 출력값

이상 탐지 모형의 출력값을 이용해 각 개체의 이상 여부를 판단한다. 주로 사용되는 방법으로는 각 개체마다 크기 순서가 있는 이상 점수(outlier score 또는 anomaly score)를 계산하는 것과 정상 또는 이상의 라벨을 부여하는 것이 있다. 이상 점수를 이용하면 정상과 이상을 구분하는 특정한 경계를 분석자가 설정해야 한다. 반면 라벨을 붙이는 방식에서는 분석자가 모형의 모수 또는 초모수(hyper parameter) 값의 조정을 통해 간접적으로 비정상과 정상의 경계를 바꿀 수 있다.

제2절 이상 탐지 국내·외 사례 연구

ICT 기술의 발전에 따라 금융 등 각종 산업 시스템의 이상 탐지 방식은 점차 컴퓨터를 기반으로 한 자동 학습 방식으로 변화하고 있다. 기존 임계값 설정에 의한 이상 탐지 기법은 각종 시스템에 대한 해석 대상이 방대하고 감시 대상을 설명하는 매개변수가 너무 많아 데이터 분석에 어려움이 있었다. 최근 세계적인 ICT 기업들은 컴퓨터 기반 시스템 개발 역량과 빅데이터 분석 경험을 활용하여 정상 시의 데이터를 학습하고 평소와 다른 상태를 자동으로 탐지하는 이상 징후 감지 솔루션 개발에 한창이다. 이 장에서는 이상 탐지 기법 적용 국내·외 사례를 연구 분야별, 데이터 유형별로 나누어 소개하고자 한다.

1. 연구 분야별 이상 탐지 기법 활용 사례

가. 침입(intrusion) 탐지

침입이란 컴퓨터 시스템에서의 악의적인 움직임을 나타낸다. 침입 탐지를 위해서는 대용량의 자료를 다루기 위해 효율적인 계산이 필요하며, 오경보율(false alarm rate)이 조금만 높아져도 분석에 큰 부담이 될 수 있다. 또한 자료가 순간마다 생겨나기 때문에 실시간 분석이 요구된다. 여기서는 정상값에 대한 라벨은 쉽게 얻을 수 있지만, 이상값은 그렇지 못해 준지도나 비지도 기법이 선호된다. Denning(1987)은 침입 탐지 시스템을 네트워크 기반과 호스트 기반으로 분류하였다.

국내 대표적인 보안관제 전문기업인 이글루시큐리티는 지난 2018년 1월 대구 AI 기반 지능형 보안관제 체계(D-Security) 구축을 완료했다. AI 시스템은 대구시의 정보자산 시스템에 대해 악의적인 사이버 위협을 가하는 공격자의 특성이 담긴 데이터를 학습하여 이상 탐지 알고리즘에 적용한다. 이글루시큐리티 이득춘 대표는 "AI 기술은 기하급수적으로 증가하는 고도화된 보안 위협에 대한 대응력을 한 단계 높이기 위한 핵심 요소로 주목을 받고 있다. 위협 정보에 대한 학습을 통해 공격을 탐지·예측하는 기계학습 기반 지능형 보안관제 체계 구축을 통해, 고객이 사이버 침해 분석·대응 능력을 강화하고, 선제적 예방 체계를 마련하며, 정보 자산 운영의 효율성을 높일 수 있도록 지속적으로 지원할 계획이다."라고 밝혔다.

[그림 2-4] 이글루시큐리티의 대구 AI 기반 지능형 보안관제 체계

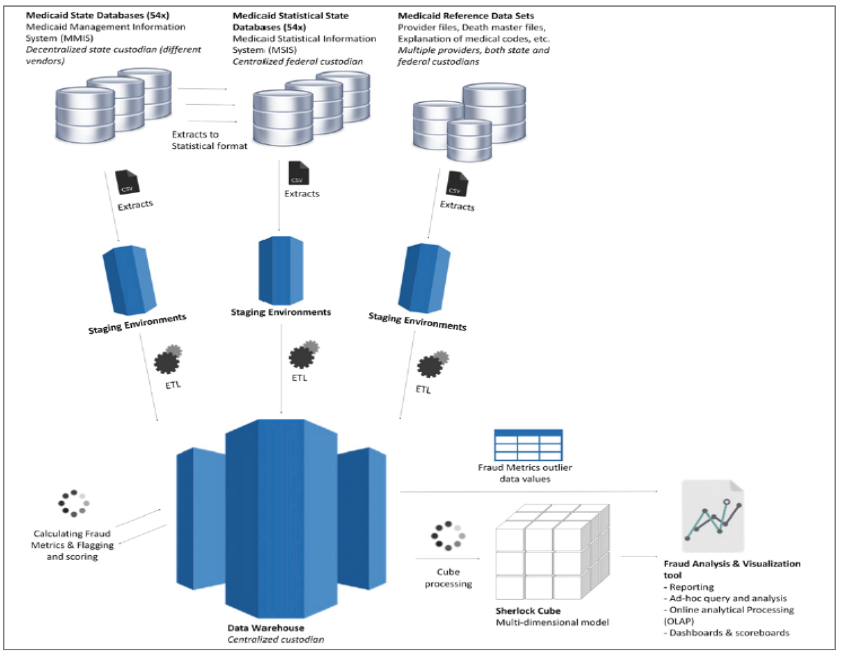
자료: SECU N CCTV News. (2018. 2. 5.). 보안 관제, 인공지능으로 효과적인 대응방안 구현해 나갈 것. Retrieved from http://www.cctvnews.co.kr/news/articleView.html?idxno=78306 2018. 6. 7.

글로벌 사이버 보안 분야를 선도하는 기업인 시만텍(Symantec)은 전 세계적으로 주목받고 있는 커넥티드 차량(connected car)에 적용할 임베디드 시스템을 개발하였다. 커넥티드 차량은 일반적인 자동차에 ICT를 융합해 언제 어디서든 네트워크에 연결되며, 크고 안전하며 타고 다닐 수 있는 스마트폰으로도 불린다. 이와 관련하여 차량과 교통 인프라 간 정보 공유 기술인 V2X(Vehicle to Everything), 차량 간 연결 기술인 V2V(Vehicle to Vehicle), 차량과 보행자 간 연결 기술인 V2P(Vehicle to Pedestrian), 차량과 인프라 간 통신 네트워크 기술인 V2I(Vehicle to Infrastructure)의 보안 솔루션이 필요한 상황이다. 이에 시만텍은 실시간으로 외부 공격자로부터 보안 위협 및 이상 징후를 탐지하여 차량의 보안 상태를 실시간으로 알려 주는 수동 침입 탐지 보안 소프트웨어를 개발(Symantec, 2018)한 바 있다.

나. 사기(fraud) 탐지

사기 탐지는 은행, 신용카드사, 보험대리점, 휴대전화 회사, 주식시장 등 산업기관에서 일어나는 범죄 행위를 잡아내는 것을 의미한다. 사기 행위의 주체는 실제 고객일 수도 있고, (신분 도용 등으로) 고객으로 위장했을 가능성도 있다. 사기는 산업기관이 제공하는 재화나 서비스를 허락되지 않은 방식으로 소비했을 때 발생하며, 산업기관은 금전적 손실을 막기 위해 이러한 사건을 즉각적으로 탐지하기를 원한다. Fawcett & Provost(1999)는 사기 탐지의 일반적인 접근을 활동 모니터링(activity monitoring)이라는 용어로 표현하였다. 이는 각 고객의 사용명세를 관리하고 거기에서 벗어난 부분 또는 패턴이 있는지를 찾아내는 방식을 말한다. 사기 탐지의 세부 분야로 신용카드, 휴대전화, 보험 청구 사기 탐지 및 내부자 거래 탐지 등이 있다.

〔그림 2-5〕 미국 메디케이드 사기 방지를 위한 이상 탐지 기법 적용 구조



자료: Capelleveen, V. G., Poel, M., Mueller, R. M., Thornton, D., Hillegersberf, J. (2016). Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. International Journal of Accounting Information Systems 21 (2016) p. 23.

Capelleveen, G. 외(2016)는 미국 메디케이드와 관련된 보험금 청구 사기 문제를 해결하기 위한 연구를 수행하였다. 의료 서비스의 공급자가 환자 진료 후 보험금을 청구하는 과정에서 유효하지 않은 서비스 청구, 상향 코드 작성, 보험 청구 중복, 과도한 진료 제공 등의 사기가 발생할 수 있다. 하루에도 50억 건씩 발생되는 많은 양의 보험금 청구에 대한 사기 패턴을 탐지하기 위해, 사후 보험금 지불 단계에서 이상 탐지 기법을 적용하였다. 연구에서는 다변량 분석(선형회귀분석, 군집분석)에서의 표준편차를 활용하여 반복적으로 이상치를 측정하여, 이상치로 볼 수 있는 특정 기준을 설정하였다.

의료급여와 관련된 부정 수급 탐지로, 미국 뉴욕주의 사례가 있다(차경엽, 오창석, 2015 재인용). 뉴욕주는 400만 명의 저소득층에게 의료급여를 지급하고 있으나, 수급자 관리에 대한 지적이 지속적으로 제기되었다. 예전에는 부정 수급 심사가 표본 추출의 개념으로 임의의 수급자를 조사하는 수준이었는데, 의료급여 관리에 대한 모니터링 체계를 구축하면서 부정 수급 예측 시스템인 SURS(Surveillance and Utilization Review Subsystem)를 개발하였다. SURS는 이상(anomaly)이라고 할 수 있는 부정 수급자의 특성을 분석하고 이를 모형화하여, 이상 징후가 발견될 경우 해당 수급자를 대상으로 심층 조사하는 방식으로 운영하고 있다. SURS 시스템의 도입 효과를 자체적으로 분석했을 때, 2008년 2억 1000만 달러, 2009년 3억 2000만 달러, 2010년 4억 3000만 달러, 2011년 6억 4000만 달러 규모의 부정 수급을 적발하는 성과를 거둔 것으로 조사되었다.

주택보조금에서도 미국 주택도시부의 부정 수급 예측 모형이 있다(차경엽, 오창석, 2015 재인용). 미국 주택도시부의 감찰관실은 주택보조금의 부정 수급 주원인을 집행상의 오류 및 거주자의 소득 은닉으로 파악하고 있는데, 이를 해결하기 위해 주택보조금 탈세를 방지하기 위한 'Operation FEDRent' 프로그램을 개발하였다. 이 프로그램 중 하나로 데이터마이닝 기법을 적용하여 주택보조금 부정 수급 예측 모형을 구축하였다. 행정 자료는 주택도시부의 세입자 정보와 외부기관인 인사관리처의 인사 정보를 활용하였다. 업무 적용 절차로, 소득 적격 여부를 심사한 뒤 부정 징후가 높은 대상에 대해 관련 서류를 확인하고 실태 조사를 실시하는 방식으로 운영하고 있다.

LA 카운티의 Department of Public Social Services(DPSS)에서도 분석 시스템을 통해 부정 수급을 탐지하고 있다(Losangeles DPSS,

2018). DPSS는 한시적 재정 지원, 고용 안정 서비스, 식료품 혜택, 노인 및 장애인을 위한 재가복지, 기타 재정 지원 등 다양한 프로그램을 제공하고 있는데, child care program에서 잠재적 사기 행위를 확인하고 부정 수급을 탐지하기 위해 분석 솔루션을 도입하였다. 보육 서비스 부당 활용 사례를 탐지하기 위해 예측 모델 및 집단 분석을 통해 고위험 등급제를 개발하였고, 사회 연결망 분석을 통해 보육 사기 및 보육 프로그램을 대상으로 한 사기 네트워크 결탁 가능성을 파악하는 데 사용하고 있다. 이러한 이상 탐지 사례들은 보고서로 저장되어 사기 사례에 대한 정보를 모니터링하고 공유하는 데 사용하고, 조사관들은 사기 행위의 패턴을 확인하여 사기 개연성이 높은 사례에 집중할 수 있도록 하고 있다. 우리나라에서도 객관적인 근거에 기반한 실업급여 부정 수급 적발 방안을 마련하고자 하였다(행정안전부, 2017) 실업급여 부정 수급 적발을 위해 행정자치부는 고용노동부, 한국고용정보원과 함께 공공 빅데이터를 활용하였다. 부정 수급 조사관들의 심층 인터뷰로 적발 노하우 및 개선 사항을 분석하여 새로운 유형의 부정 수급 패턴을 발굴하였다. 이를 통해 실업급여 신청자 및 사업장의 위험 점수를 측정하여 조사관들에게 부정 수급 우선순위 리스트를 제공하였다. 그리고, 네트워크 다이어그램 기법을 적용하여 실업급여 신청 현황과 부정 수급자 사업장의 분포를 시각화하여 조사관들에게 유용한 정보를 제공해 주고 있다.

다. 작업환경 이상 탐지

제조업이 활성화된 국내 산업현장에는 다양한 유해·위험 요인이 존재한다. 유해화학물질, 소음, 위험기계 및 설비 등은 노동자의 건강과 생명을 위협하는 요인으로 작용하고 있다. 특히 유해화학물질은 눈에 보이지

않는 입자, 증기 등의 형태로 인간의 체내에 흡수됨으로써 직업성 질병을 유발하기도 한다. 고용노동부에서는 이를 관리하기 위하여 산업안전보건법 제42조의 「작업환경측정제도」와 고용노동부 고시 제2018-62호 「화학물질 및 물리적 인자의 노출기준」을 제정·운영하고 있다. 즉, 정부에서는 화학물질에 대한 공기 중 노출기준을 제시하고, 사업주로 하여금 작업환경 측정을 정기적으로 시행토록 함으로써 화학물질로 인한 유해·위험요인을 관리하고 있는 것이다.

대부분의 제조사에서는 좋은 품질의 제품 생산을 유지하기 위해 화학물질, 기계설비, 작업 내용, 온도 및 습도 등의 환경 등을 일정하게 유지시키고 있으며, 이에 따라 작업환경 측정 결과도 매년 유사하게 나타나고 있다. 그럼에도 일부 산업현장에서는 작업환경 측정 결과의 전·후 수치가 2배 이상 차이 나는 이상 현상이 나타나기도 한다. 이와 같은 이상치 등을 관리하기 위하여 고용노동부 및 안전보건공단에서는 작업환경 측정 결과에 대한 평가를 시행하고 있다(안전보건공단, 2012). 이에 더해, 유해화학물질을 보유하고 있는 제조사의 물질 특성, 설비 및 작업 내용 등을 데이터베이스화하여 작업환경 측정 결과 차이에 대한 원인을 추정하기 위해 노력하고 있다. 또한, 측정 결과 차이로 인한 영향을 예측함으로써 노동자 및 동종 사업장에 대한 관리를 좀 더 정확하고 정교하게 할 수 있다. 전문가의 지식에 기반한 논리 시스템에 학습(learning) 기능을 추가한 예방정비 시스템의 활용은 사람에 대한 의존도를 줄이면서도 효과적으로 고장 예측과 진단이 가능하다는 점에서 현재 대두되고 있는 스마트 팩토리의 트렌드에 부합한다.

라. 기타 분야

세계 최대 숙박 공유 서비스를 제공하는 에어비앤비(Airbnb)는 숙박 제공자와 고객이 각자 원하는 지불 방법을 택할 수 있도록 결제 플랫폼을 구축하였다(Lu, J., 2015). 전 세계 190개국의 통화를 지원하면서 발생하는 특정 통화에 대한 결제를 처리할 수 없거나 특정 지불 게이트웨이(인터넷 사이트, 어플)에 접근하지 못하는 문제를 최대한 빨리 해결하고자, 문제가 발생하면 이를 실시간으로 감지하고 분석하기 위해 이상 탐지 기법을 적용한 시스템을 구축한 바 있다.

또한 스마트폰을 기반으로 교통 서비스를 제공하는 우버(Uber)는 사기성 계정을 식별하여, 효율적으로 승차 정보를 공유하고 최적의 배달 음식 추천 및 대기 시간 예측을 위해 실시간 이상 탐지 솔루션을 개발한 바 있다(Jin, J., 2018).

2. 데이터 유형별 이상 탐지 기법 활용 사례

가. 의약, 공중보건(medical and public health) 이상 탐지

이 분야의 이상 탐지는 주로 환자의 기록을 바탕으로 진행한다. 이상의 원인은 환자 상태의 변화, 계측 오차, 기록 오차 등으로 다양하다. 몇몇 기법은 특정한 부위에서 질병이 생겼는지 알아내는 데 중점을 둔다(Wong et al., 2003). 의학 분야 내 이상 탐지는 이상값을 정상으로 판정했을 때 발생하는 대가가 아주 크기 때문에 높은 정확도가 요구된다. 자료는 연속형, 범주형, 시간적, 공간적 등 다양한 형태로 구성될 수 있다. 대부분 기법은 비정상적인 기록(점 이상)을 찾는 것을 목표로 한다. 주로

건강한 경우 라벨이 확보되어 준지도 접근이 큰 비중을 차지한다. 이외에 심전도나 뇌전도와 같은 시계열 자료를 다루기도 한다.

국내 사례로는 한국정보화진흥원과 건강보험심사평가원이 2016년 구축한 환자 안전 조기 이상 감지 시스템이 있다. 3~5년간 의료기관에서 청구하여 건강보험심사평가원에서 보유하고 있는 행정 데이터를 분석하여, 감염병별 의약품 처방 패턴 구축 프로세스, 감염병 발생 예측 및 조기 이상 감지 프로세스, 의약품안전사용서비스(DUR, Drug Utilization Review)와 연계한 실시간 감시 체계 프로세스를 개발하였다. 이를 통해 실시간에 가까운 의료 정보를 활용하여 감염병 발생을 감지할 수 있는 시스템이 구축되었다(한국정보화진흥원, 건강보험심사평가원, 2016).

나. 영상 처리(image processing)에서의 이상 탐지

영상을 다루는 이상 탐지 기법은 시간에 따른 이미지의 변화(움직임 탐지)나 정적인 이미지에서 이상한 부분을 찾는 것이 목적이다. 위성 영상(Augusteijn & Folkert, 2002; Byers & Raftery, 1998; Moya et al., 1993; Torr & Murray, 1993; Theiler & Cai, 2003), 숫자 인식(LeCun et al., 1990), 분광학(Chen et al., 2005; Davy & Godsill, 2002; Hazel, 2000; Scarth et al., 1995), 유방 X선 분석(Spence et al., 2001; Tarassenko, 1995), 비디오 감시(Diehl & Hampshire, 2002; Singh & Markou, 2004; Pokrajac et al., 2007)가 활용 예시이다. 이상값은 보통 움직임, 이질적인 물체나 계측 오차에 의해 나타난다. 자료는 시간적, 공간적 특성을 모두 지니며, 각 개체는 색, 밝기, 질감과 같은 몇 개의 연속적 속성을 가진다. 입력 자료의 크기가 매우 크기 때문에 자료를 저장하거나 이상 탐지를 위해 계산을 할 때 추가적인 어려움이 있다.

또한 비디오 자료를 다룰 때에는 실시간에 처리하는 기법이 필수적이다.

다. 문자 자료(text data) 이상 탐지

인터넷을 통해 발생하는 데이터가 기하급수적으로 늘어나면서 자연어 형태(natural language)의 비정형 자료(unstructured data)의 양도 증가하고 있다. 문자 자료 이상 탐지는 문서나 뉴스 기사 모음에서 새로운 주제나 사건을 찾는 것이 주된 목적이다. 즉, 그 새로운 사건이나 이례적인 주제가 이상값을 만든다. 자료는 보통 고차원이고 밀도가 드물며(sparse), 문서가 시간이 지나면서 누적되기 때문에 시간적인 성질도 가진다.

라. 센서망(sensor network)

센서망은 최근에 떠오르는 주제로, 여러 무선 센서에서 모은 자료의 독특한 특징 때문에 자료 분석 관점에서 특히 중요하다. 센서망 자료의 이상값은 센서가 비정상적 사건을 제대로 잡았거나, 아니면 센서에 문제가 있다는 것을 의미한다. 따라서 센서망 이상 탐지는 센서 오작동과 침입을 모두 다룬다. 센서가 수집하는 자료 종류는 이진, 이산적, 연속적, 음성, 비디오 등 천차만별이며, 자료가 지속적으로 생성된다. 또한, 센서가 설치된 환경에 따라 자료에 잡음이 끼거나 결측값이 있을 수도 있다.

센서망 이상 탐지에는 여러 난관이 있다. 먼저 실시간으로 작동해야 하며, 센서가 여러 장소에 설치되어 있기 때문에 분석에 있어 분산 데이터 마이닝(distributed data mining) 접근이 요구된다. 추가적으로 잡음, 결측값을 이상값과 구별해야 한다는 점도 큰 걸림돌이다.

마. 기타 분야

음성 인식(Albrecht et al., 2000; Emamian et al., 2000), 로봇의 행동 관찰(Crook & Hayes, 2001; Crook et al., 2002; Marsland et al., 1999, 2000b, 2000a), 교통 모니터링(Shekhar et al., 2001), 부정 클릭 방지(click-through protection)(Ihler et al., 2006), 웹 애플리케이션(Ide & Kashima, 2004; Sun et al., 2005), 생물학(Kadota et al., 2003; Sun et al., 2006; Gwadera et al., 2005; MacDonald & Ghosh, 2007; Tomlins et al., 2005; Tibshirani & Hastie, 2007), 인구 조사(Lu et al., 2003), 범죄 연계(Lin & Brown, 2003), 고객 관계 관리[Customer Relationship Management, CRM(He et al., 2004b)], 천문학(Dutta et al., 2007; Escalante, 2005; Protopapas et al., 2006), 생태계 파괴(Blender et al., 1997; Kou et al., 2006; Sun & Chawla, 2004) 등을 비롯한 이상 탐지를 활용하는 수많은 분야가 있다.

제 3 장

데이터 사이언스 기반 이상 탐지 기법 연구

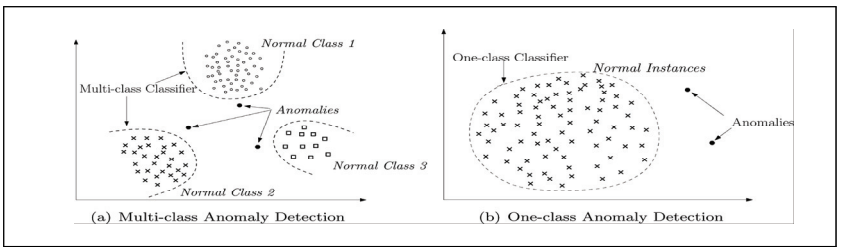
3 데이터 사이언스 기반 이상 탐지 기법 연구

제1절 기계학습 기반 이상 탐지 기법[3)]

1. 분류 기반 이상 탐지 기법

분류[classification(Tan et al., 2005; Duda et al., 2000)]는 각 개체에 어느 클래스에 속하는지에 대한 라벨이 붙어 있는 자료로 분류기(classifier)를 학습(training)한 뒤, 학습된 모형으로 새로운 개체에 대해 각 클래스에 속할 확률을 예측하는 방법이다. 분류 기반 이상 탐지 기법도 거의 비슷한 과정을 거친다. 이 기법을 적용할 때는 **'분류기를 주어진 특성 공간(feature space)에서 학습시킬 수 있다'**고 가정한다. 자료의 라벨 개수에 따라 다집단(multi-class)과 일집단(one-class)으로 나눌 수 있다.

[그림 3-1] 분류 기반 이상 탐지



자료: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15. 21page

3) 기계학습 기반 이상 탐지 기법은 [Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15]논문 내용을 번역하여 정리하였음

다집단 기법은 자료가 여러 클래스로 이루어졌다고 가정한다(Stefano et al., 2000; Barbara et al., 2001b). 이상 탐지를 위해서 분류기가 각 정상 클래스와 나머지를 구분하도록 학습시키고, 어느 클래스에도 포함되지 않는 개체를 이상값으로 처리한다(〔그림 3-1(a)〕 참고). 일부 기법은 개체에 신뢰 점수(confidence score)를 부여해 개체를 확실하게 정상으로 분류하는 분류기가 없으면 이상값이라고 판정한다.

일집단 기법은 학습 자료가 모두 정상이라고 가정한다. 〔그림 3-1(b)〕과 같이 일집단 SVM(Schölkopf et al., 2001), 일집단 커널 Fisher 판별식(Roth, 2004, 2006) 등의 일집단 분류 알고리즘을 이용해 정상 개체들을 두르는 결정 경계(decision boundary)를 학습한다. 테스트 개체가 학습된 경계 밖에 있으면 이상값이 된다.

분류기를 만드는 알고리즘별로 나누어 살펴보면 다음과 같다.

가. 신경망(Neural network) 기반

신경망은 다집단과 일집단 문제에 모두 적용할 수 있다. 다집단의 경우, 기본 과정은 신경망이 여러 정상 클래스를 학습한 후에 테스트 자료를 그 신경망에 입력값으로 넣는 방식이다. 테스트 결과는 신경망의 출력값으로 이상 여부를 판단한다(Stefano et al., 2000; Odin & Addison, 2000).

일집단 문제에는 복제 신경망(replicator neural network)이 쓰여 왔다(Hawkins et al., 2002; Williams et al., 2002). 복제 신경망은 입력 자료를 저차원으로 압축하기 위해 제안된 신경망으로, 입력층(input layer)과 출력층(output layer)의 노드 수가 같고, 자료 x에 대해 하나 이상의 은닉층(hidden layer)으로 구성된 인코더를 통해 입력 자료를 압

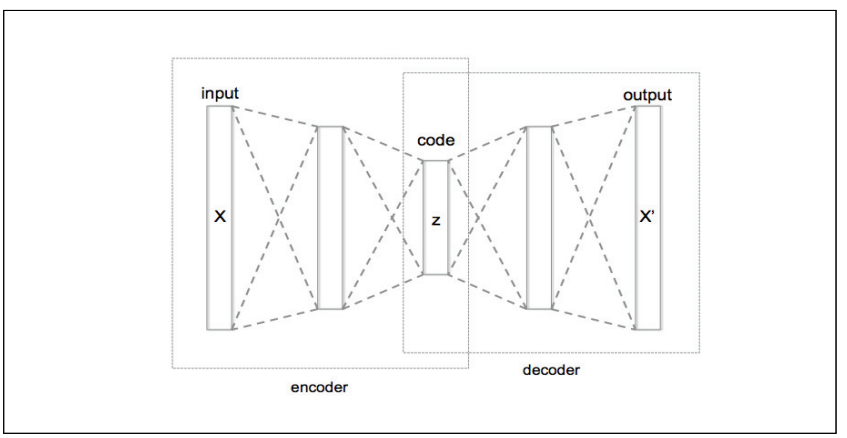
축하고, 디코더를 사용하여 개체를 복원하여 출력값 x'를 구한다(〔그림 3-2〕 참고). 복제 신경망의 모수는 복원 오차를 최소화하는 방향으로 학습시킨다.

일집단 기반 이상 탐지를 위해서는 먼저 정상 자료를 사용하여 복제 신경망의 모수를 학습시킨 뒤, 테스트 때 입력 개체를 학습된 신경망으로 압축한 뒤 복원하였을 때 발생하는 복원 오차가 크면 클수록 이상 개체라고 판단한다. 즉, 시험 입력 자료 x_i를 학습된 신경망으로 압축시킨 뒤, 복원해 출력값 x_i'를 계산한다. 시험 자료 개체 x_i의 복원 오차 δ_i는 $i = 1, \cdots, n$이 각 특성을 의미할 때

$$\delta_i = \frac{1}{p}\sum_{j=1}^{p}(x_{ij} - x_{ij}')^2$$

로 주어지며, 이 값을 그대로 이상 점수로 쓸 수 있다. 여기서 p는 입력 자료의 차원 수를 의미한다.

〔그림 3-2〕 복제 신경망 구조의 예



자료: 위키피디아(2018. 9. 1), https://en.wikipedia.org/wiki/Autoencoder

나. 베이지안 네트워크(Bayesian network) 기반

베이지안 네트워크는 다집단 문제에 이용된다. 분류 문제일 때 나이브 베이지안 네트워크로 테스트 개체의 각 정상 클래스와 이상에 대한 사후 확률을 추정해 가장 높은 확률에 해당하는 클래스로 지정한다. 이때 각 클래스의 사전확률과 조건부확률은 학습 자료를 사용하여 추정한다. 확률이 0이 나오는 경우 라플라스 스무딩(Laplace smoothing)으로 0 대신 적절한 양수 값을 부여한다.

네트워크 침입 탐지(Barbara et al., 2001b; Sebyala et al., 2002; Valdes & Skinner, 2000), 비디오 감시(Diehl & Hampshire, 2002), 문자 자료(Baker et al., 1999), 질병 발생 탐지(Wong et al., 2002, 2003)에 이를 응용한 기법이 제안되었다. 한편 나이브 베이즈 분류는 변수들이 독립이라는 조건이 필요한데, 더 복잡한 베이지안 네트워크를 사용해 몇 변수가 조건부 종속일 때를 고려하는 기법들이 연구되었다(Siaterlis & Maglaris, 2004; Janakiram et al., 2006; Das & Schneider, 2007).

다. SVM(Support vector machine) 기반

SVM(Vapnik, 1995)은 일집단 이상 탐지 문제를 다룬다. SVM을 이용한 일집단 학습 기법(Ratsch et al., 2002)을 이용해 학습 집합을 포함하는 영역(또는 그 경계)을 학습한다. 영역 기반 기법은 경계에만 초점을 맞추고 경계의 내·외부에서의 분포에서는 관심을 갖지 않기 때문에, 분포에 둔감하고 자료의 샘플링이 어떻게 이루어졌는지와 무관한 결과가 나온다. 영역의 구분이 복잡하면 RBF(radial basis function)와 같은 커널

함수를 사용하기도 한다. 테스트 개체가 학습된 영역에 들어가면 정상으로, 아니면 이상으로 판별한다.

SVM 기법은 음성 신호 자료(Davy & Godsill, 2002), 발전소 이상 탐지(King et al., 2002), 시스템 호출(Eskin et al., 2002; Heller et al., 2003; Lazarevic et al., 2003)을 포함해 시간적 순서 자료[temporal sequence(Ma & Perkins, 2003a, 2003b)]에도 이용되었다.

라. 결정 규칙(Decision rule) 기반

결정 규칙 기반 기법은 시스템에서 정상 자료를 판단하는 규칙들을 학습하고, 어떠한 규칙에도 해당하지 않는 개체를 이상으로 취급한다. 다집단, 일집단 문제 모두에 이러한 기법을 적용할 수 있다.

기초적인 학습 구조는 다음과 같다. 먼저 학습 집합에 RIPPER, 의사결정나무 등의 결정 규칙 학습 알고리즘을 이용해 결정 규칙을 학습한다. 각 규칙에는 규칙이 올바르게 분류한 학습 개체의 개수와 전체 학습 자료 수의 비율에 비례하는 신뢰도(confidence)값이 부여된다. 그 다음 각 개체에 대해 해당 개체를 가장 잘 잡아내는(capture) 규칙을 찾고, 그 규칙의 신뢰도의 역수를 이상 점수로 한다. 이것을 약간 변형한 기법들이 여럿 제안되었다(Fan et al., 2001; Helmer et al., 1998; Lee et al., 1997; Salvador & Chan, 2003; Teng et al., 1990).

연관 규칙 마이닝[association rule mining(Agrawal & Srikant, 1995)]이 일집단 문제에 대한 비지도 결정 규칙 학습에 사용되어 왔다. 결정 규칙은 범주형 자료에서 생성된다. 규칙들이 강한 패턴만을 보여 주게 하도록 지지도(support)가 일정 수준을 넘지 못하는 규칙은 제거한다(Tan et al., 2005). 연관 규칙 마이닝 기반 기법은 네트워크 침입

(Mahoney & Chan, 2002, 2003; Mahoney et al., 2003; Tandon & Chan, 2007; Barbara et al., 2001a; Otey et al., 2003), 시스템 호출 침입(Lee et al., 2000; Lee & Stolfo, 1998; Qin & Hwang, 2004), 신용카드 사기(Brause et al., 1999), 우주선 살림[spacecraft housekeeping (Yairi et al., 2001)] 등에 활용되었다. 연관 규칙 마이닝 과정에서 빈발 항목 집합(frequent item set)이 얻어지는데, He et al.(2004a)은 이상 점수를 개체가 포함되는 빈발 항목의 개수로 하는 알고리즘을 제안하였다.

마. 계산 복잡도 및 장단점 비교

(참고 1) 계산 복잡도

분류 기반 기법의 계산 복잡도는 사용한 분류 알고리즘에 따라 다르다. 더 자세한 논의는 Kearns(1990)에서 볼 수 있다. 일반적으로 의사결정 나무 학습이 빠른 편이고 2차 최적화가 필요한 SVM이 느린 편이지만, Joachims(2006)가 학습을 선형 시간에 완료하는 방법을 제안하였다. 분류 모형 기반 이상 탐지 방법들은 학습 과정에서 이미 학습된 모형을 테스트 과정에 사용하기 때문에 테스트 과정은 매우 빠르다.

(참고 2) 분류 기반 이상 탐지 기법의 장단점

- 여러 강력한 알고리즘들을 이용할 수 있다.
- 이미 학습된 모형에 대해 예측만 하면 되므로 테스트 과정이 매우 빠르다.
- 다집단 분류에서 각 정상 개체 종류에 대한 라벨을 구하기 어려울 수 있다.
- 신경망 모형의 경우 학습을 위해서 뛰어난 성능의 기계가 필요하다는 점이 문제 될 수 있고, SVM을 사용할 때 어떠한 커널을 사용해야 할

지 결정해야 한다.

- 각 개체에 대한 이상 점수가 필요한 경우에는 라벨만 지정하는 기법들을 활용하기 어렵다. 이것을 해결하기 위해 몇 기법은 분류기의 출력에서 확률적인 예측 점수를 얻어 낸다(Platt, 2000).

2. NN(Nearest neighbor) 기반 이상 탐지 기법

이 분야의 기법에서는 **'정상값들은 어떤 근방(들)(neighbor)에 밀집되어 있고, 이상값은 각 근방에서 멀리 떨어져 있다'**고 가정한다. Nearest neighbor(NN) 기반 기법을 쓰려면 두 개체 사이 거리의 개념이 정의되어야 한다. 거리는 여러 가지 방법으로 정의할 수 있다. 연속형 변수에 대해서는 유클리드 거리가 일반적인 선택이고, 다른 척도를 사용할 수도 있다(Tan et al., 2005, Chapter 2). 범주형 변수에 대해서는 단순 일치 계수(simple matching coefficient)가 자주 쓰이며 마찬가지로 더 복잡한 척도가 존재한다(Boriah et al., 2008; Chandola et al., 2008). 자료가 다변량이면 각 변수에 대한 거리를 결합한다(Tan et al., 2005, Chapter 2).

NN 기반 기법에서 이상 점수를 구하는 방법은 k번째로 가까운 개체와의 거리를 이용하거나, 상대 밀도(relative density)를 이용하는 방법이다. 이외에 거리를 다른 방식으로 활용하는 몇몇 기법을 추후에 간략하게 살펴보겠다.

가. k번째로 가까운 개체와의 거리 이용

이상 점수를 k번째로 가까운 개체와의 거리로 정의한다. 위성 사진에서 지뢰를 찾거나(Byers & Raftery, 1998) 최근접 이웃과의 거리를 이용하여 대형 동기 터빈발전기(large synchronous turbine-generator)의 직류 계자권선의 이상 탐지(Guttormsson et al., 1999)에 응용되었다. 일반적으로 이상 점수의 경곗값(threshold)을 설정하지만, Ramaswamy et al.(2000)은 이상 점수를 기준으로 정렬시킨 뒤 이상 점수가 가장 큰 m개를 이상값으로 보았고, 여기서 m은 분석자가 선택하는 상수이다.

기본적인 방법에서 다음과 같이 세 가지 방향으로 확장해 나갔는데, 첫 번째는 이상 점수의 정의를 바꾸었고, 두 번째는 연속형이 아닌 자료 처리를 위해 다른 거리 척도를 도입했고, 세 번째는 가까운 개체를 얻기 위해 자료 수의 제곱 시간이 소요되는 기본 기법의 계산 효율을 늘렸다.

Eskin et al.(2002), Angiulli & Pizzuti(2002), Zhang & Wang (2006)은 가장 가까운 k개 개체와의 거리의 합을 이상 점수로 주었다. 이와 비슷한 기법인 동료 집단 분석(peer group analysis)을 Bolton & Hand(1999)가 신용카드 사기 탐지에 활용하였다.

이상 점수를 구하는 다른 방법으로 한 개체에서 일정 거리(d) 이내에 있는 개체의 수(n)를 세는 방법이다(Knorr & Ng, 1997, 1998, 1999; Knorr et al., 2000). 즉 반지름 d인 초구(hypersphere) 안의 개체 개수를 세는 것이므로, 이 방법을 전역 밀도(global density) 추정으로 생각할 수도 있다. 예를 들면 2차원 자료에서 밀도는 $n/\pi d^2$이 된다. 이상 점수는 구한 밀도의 역수로 놓을 수 있지만, 실제 밀도를 쓰는 대신 어떤 기법들은 d를 고정하고 $1/n$을 이상 점수로, 다른 기법들은 n을 고정하고 $1/d$을 이상 점수로 사용한다.

위에 언급한 내용은 대부분 연속형 변수에 관한 기법이었지만, 범주형 변수의 자료를 다루는 여러 기법이 존재한다. HOT(Wei et al., 2003)이라 불리는 초그래프(hypergraph) 기반 기법은 범주형 값들을 초그래프로 모형화한 뒤 그 그래프에서의 연결 관계를 통해 거리를 계산한다. Otey et al.(2006)은 범주형, 연속형 변수가 모두 포함되어 있는 자료에 대한 거리 척도를 제안하였다. 범주형, 연속형에 대한 거리를 따로 구한 후에 더하는 방식을 쓰는데, 범주형 속성에 대해서는 두 개체가 일치하는 값의 개수를, 연속형에 대해서는 종속(dependence)을 찾기 위해 공분산 행렬을 이용한다. Palshikar(2005)는 연속형 순차 자료에 Knorr & Ng(1999)의 기법을 적용하였고, Kou et al.(2006)은 Ramaswamy et al.(2000)의 기법을 공간 자료로 확장하였다.

그리고 효율성 개선을 위해 다양한 변형이 제시되었다. 그 중 몇 가지는 이상값이 될 수 없는 개체를 무시하거나, 가장 이상값이 될 만한 개체에만 집중해 탐색 공간을 축소하였다. Bay & Schwabacher(2003)는 자료가 충분히 랜덤화되면 간단한 가지치기(pruning)만으로 평균 시간을 거의 선형으로 줄일 수 있음을 보였다. 개체에 대해 가까운 이웃들을 구했으면 알고리즘은 지금까지 찾아낸 이상값의 이상 점수 중 최소를 경곗값으로 정한다.

한편 Ramaswamy et al.(2000)은 분할 기반 기법을 소개하였다. 우선 자료를 분할적 군집화(partitional clustering)한 뒤, 각 분할에 대해 그 안에서 각 개체의 k번째로 가까운 이웃과의 거리를 구해 그것의 최소, 최댓값으로 가장 이상한 k개의 개체를 포함하지 않을 만한 군집은 제거하고 나머지에서만 이상값을 찾아낸다. Eskin et al.(2002), McCallum et al.(2000), Ghoting et al.(2006), Tao et al.(2006)이 비슷한 군집화 기반 가지치기 기법을 제안하였다.

Wu & Jermaine(2006)은 각 개체에 대해 정해진 작은 수 M의 표본을 뽑아 그 안에서 NN을 구한다. 따라서 계산 복잡도는 M이 표본의 크기일 때 $O(MN)$이 되고, 여기서 N은 자료 수이다.

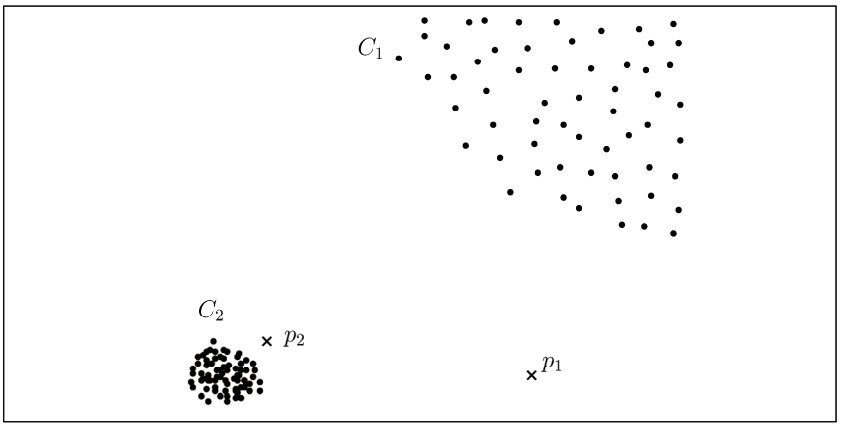
또 다른 방법은 변수 공간을 고정된 크기의 초입방체(hypercube)로 나누고, 특정 초입방체에 속한 개체가 많으면, 포함된 개체들은 정상일 것이라는 직관에서 기인한다. 마찬가지로 한 개체가 속한 초입방체 및 그와 인접한 것들에 개체가 거의 없으면 이상값일 가능성이 클 것이다(Knorr & Ng, 1998). Angiulli & Pizzuti(2002)는 d차원 자료 공간을 $[0,1]^d$로 변환한 뒤 공간 채움 곡선(space filling curve)을 이용해 $I=[0,1]$로 선형화하고 I에서 NN 기법을 적용하였다.

나. 상대 밀도 이용

밀도 기반 기법은 각 개체 근방의 밀도를 추정한다. 근방의 밀도가 낮은 개체는 이상값이라 판단한다. 앞에서 언급한 k번째로 가까운 개체와의 거리를 구하는 것은 k개를 포함하는 구의 반지름을 구하는 것과 같으므로, 밀도 기반 기법의 한 예시로 볼 수 있다.

이러한 기법은 영역에 따라 밀도가 다를 때 취약하다. 〔그림 3-3〕의 2차원 자료를 살펴보면 C_1 군집의 밀도가 낮아, C_1 군집의 많은 개체가 가장 가까운 개체와의 거리가 p_2와 C_2 군집 사이의 거리보다 크다. 따라서 밀도 기반 기본 방법은 이상값 p_1은 쉽게 발견하겠지만, p_2는 잡아내지 못할 것이다. 이러한 문제에 대처하려면 근방에 대한 상대적인 밀도를 활용할 필요가 있다.

〔그림 3-3〕 국소적 밀도의 전역 밀도에 대한 이점



자료: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15. 28page

Breunig et al.(1999, 2000)은 LOF(local outlier factor)라는 이상 점수를 제안하였다. 개체의 LOF 점수는 가장 가까운 k개 점들의 국소적 밀도(local density)의 평균과 자기 자신의 국소적 밀도의 비율로 정의된다. 여기서 국소적 밀도는 k/(k개의 이웃을 포함하는 가장 작은 구의 부피)이다. 정상값은 조밀한 영역에 위치해 국소적 밀도가 그 이웃들과 비슷한 반면, 이상값은 NN에 비해 상대적으로 국소적 밀도가 매우 낮기 때문에 큰 LOF 점수를 얻는다. 〔그림 3-3〕에 LOF를 적용하면 p_1, p_2를 이상값이라고 판단한다.

LOF는 'core distance'와 'reachability distance'의 개념이 사용되는데, $k-distance(A)$는 A와 k번째 근접 이웃과의 거리(kNN)라고 하면, $N_k(A)$는 k번째 근접 이웃 데이터 포인트와 A와의 거리 안에 포함되는 데이터 포인트들의 수(집합)라고 정의한다.

'reachability distance'는

$$reachability-distance_k(A,B) = \max\{k-distance(B), d(A,B)\}$$

이다.

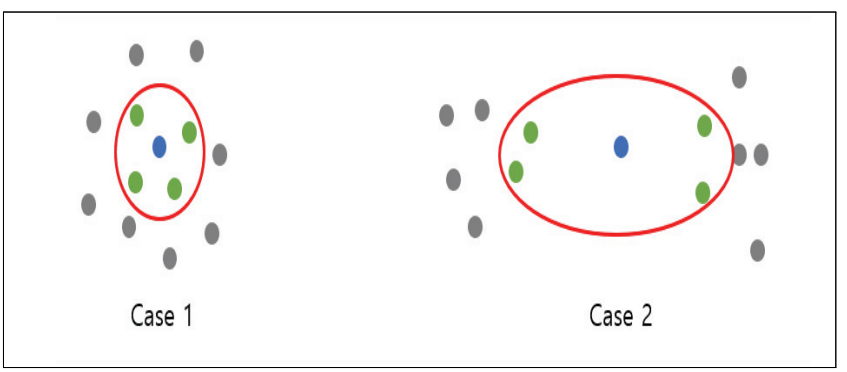
여기에서의 distance 개념은 symmetric하지 않기 때문에 수학적 정의의 거리 개념은 아니다. $reachability-distance_k(A,B)$와 $reachability-distance_k(B,A)$는 다를 수 있기 때문에 symmetric하지 않다.

object A의 local reachability density

$$lrd(A) = 1/\left(\frac{\sum_{B\in N_k(A)} reachability-distance_k(A,B)}{|N_k(A)|}\right)$$

lrd(A)는 A에 속한 B의 reachability distance 평균의 역수값이다.

[그림 3-4] LOF 밀도 비교 1



자료: 정재윤. (2017). 로컬 아웃라이어 팩터. https://jayhey.github.io/novelty%20detection/2017/11/10/Novelty_detection_LOF/) 2018. 11. 28. 인출

위 그림에서 보면 lrd(Case 1) 〉 lrd(Case 2)이다. 즉, A(파란색)가 밀도가 높은 곳에 있는 경우, 밀도가 낮은 곳에 있는 경우보다 lrd값이 높다.

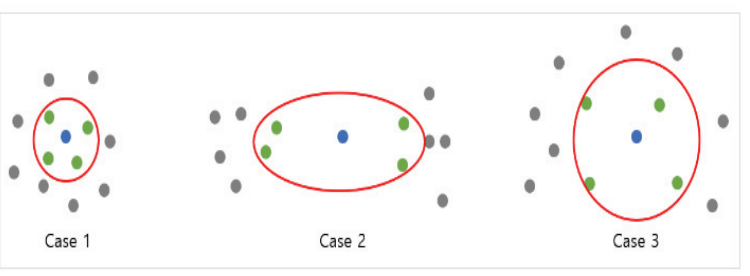
그렇다면, A가 그 주위의 이웃들과 비교했을 때의 이상치 정도를 계산한 것이 Local Outlier Factor이다.

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} lrd(B)}{|N_k(A)|} / lrd(A)$$

LOF값이 1에 비슷한 값이면 클러스터 안에 있는 데이터 포인트라고 해석할 수 있다.

주어진 데이터 포인트와 그 이웃이 동질적인 밀도(density) 영역에 있다고 볼 수 있다. 반면에 LOF값이 1보다 크면 이상값이라고 판단할 수 있다.

[그림 3-5] LOF 밀도 비교 2

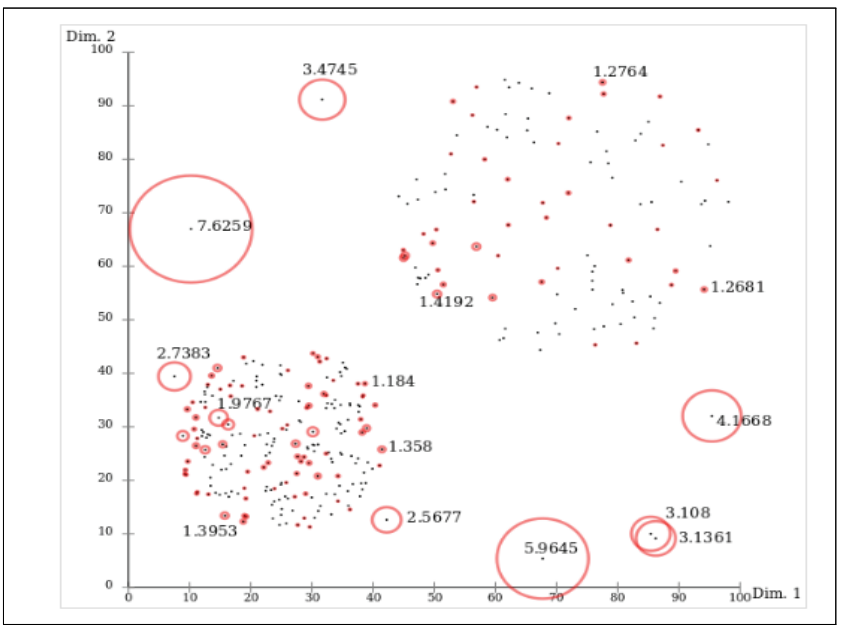


Case	$ldr_k(A)$	$ldr_k(B)$	$LOF_k(A)$
Case 1	Large	Large	Small
Case 2	Small	Large	Large
Case 3	Small	Small	Small

자료: 정재윤. (2017). 로컬 아웃라이어 팩터. https://jayhey.github.io/novelty%20detection/2017/11/10/Novelty_detection_LOF/) 2018. 11. 28. 인출

A(파란색)가 밀도가 낮은 곳에 있을수록, B(녹색)가 밀도가 높은 곳에 있을수록 A의 LOF값은 커진다.

〔그림 3-6〕 LOF 밀도 비교 3



자료: 정재윤. (2017). 로컬 아웃라이어 팩터. https://jayhey.github.io/novelty%20detection/2017/11/10/Novelty_detection_LOF/) 2018. 11. 28. 인출

그림 안의 숫자들은 위 수식으로 구한 LOF값을 나타내는데, 밀도가 높은 곳에서 가까운 이상값에 상대적으로 높은 LOF값이 계산됨을 알 수 있다.

LOF는 밀집된 클러스터에서 조금만 벗어나 있어도 이상값으로 판단할 수 있다는 장점이 있지만, 이상치의 기준값을 LOF값으로 판단해야 한다는 어려움이 있다. 또한, k의 선택 역시 고려해야 할 부분이다.

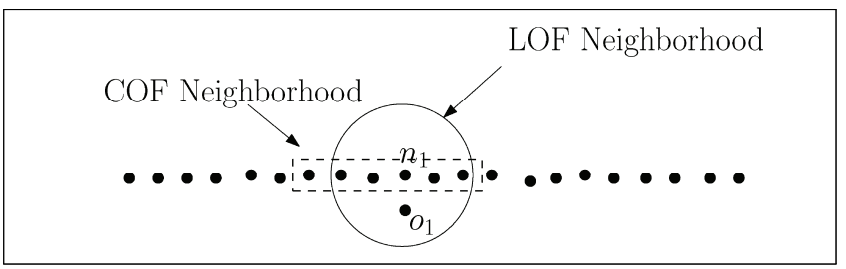
LOF의 변형이 많은 연구에서 제시되었다. 일부는 국소적 밀도를 다른 방식으로 계산하였고, 다른 일부는 LOF를 복잡한 형태의 자료에 적용하였으며, LOF 기법의 복잡도를 줄이는 방법도 제안하였다.

Tang et al.(2002)은 LOF의 변형인 COF(connectivity-based outlier factor)를 제안하였다. 기존 LOF와의 차이는 근방을 구하는 방

식인데, COF에서는 귀납적으로 근방과의 거리(근방에 포함된 개체와의 거리들 중 최솟값)가 가장 작은 점을 추가해 k개가 포함될 때까지 근방을 만들어 나간다. 근방이 구해진 이후의 이상값 계산 과정은 LOF와 같다. COF는 [그림 3-7]와 같은 선형 영역에 유리하다.

Hautamaki et al.(2004)은 LOF의 간단한 버전으로 ODIN(outlier detection using in-degree number)이라는 방법을 제안하였다. 개체 x의 ODIN값은 x의 가장 가까운 k개 개체 중에 그 개체에서 가장 가까운 k개 안에 x를 포함하는 개수이다. ODIN의 역수를 이상 점수로 쓸 수 있다. 비슷한 기법이 Brito et al.(1997)에 의해 소개되었다.

[그림 3-7] LOF와 COF에서의 근방 차이



자료: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15. 28page

Papadimitriou et al.(2002)에서도 LOF의 변형인 MDEF (multi-granularity deviation factor)를 제안하였다. 각 개체의 MDEF 값은 가장 가까운 이웃들의 국소적 밀도의 표준편차이다. 이상 점수는 MDEF의 역수로 놓는다.

다른 형태의 자료에 LOF를 응용한 사례도 있다. Sun & Chawla(2004, 2006)가 기상 자료에서 공간적 이상 탐지를 목적으로 한 척도를 제안했고, Yu et al.(2006)은 범주형 자료에 유사한 척도를 사용하였다. 또한

Pokrajac et al.(2007)이 비디오 센서 자료에 LOF를 확장한 기법을 적용하였다.

몇몇 LOF의 변형은 효율 개선을 목표로 하였다. Jin et al.(2001)은 모든 개체의 LOF 점수를 구하는 대신 가장 이상한 n개의 개체만을 찾아내는 방법을 제안하였다. 이 방법은 자료에서 극소군집들을 찾아내고, 각 군집에서 LOF값의 최소, 최대를 구하는 과정을 포함한다. Chiu & Chee Fu(2003)는 세 가지의 접근법을 내놓았는데, 모두 가장 이상한 n개를 가지지 않을 군집을 제거하고 남은 군집들에 대해서만 추가 계산을 통하여 자세히 분석한다.

다. 계산 복잡도 및 장단점 비교

(참고 1) 계산 복잡도

기본 NN 기법과 LOF 기법의 단점은 자료 수에 대해 제곱 시간이 소요되는 알고리즘이라는 점이다. 이 기법들이 각 개체에 대해 근방을 찾는다는 것을 고려하면, $k-d$ 트리(Bentley, 1975)나 R-트리(Roussopoulos et al., 1995)를 활용해 효율을 높일 수 있다. 하지만 이들 역시 변수의 수가 늘어나면 그 효력이 떨어진다. 어떤 기법은 가장 이상한 몇 개의 값에만 관심을 가지는데, 각 개체의 이상 점수가 필요한 경우에는 사용할 수 없다. 변수 공간을 초입방체로 나누는 방식도 자료 크기에는 선형이지만 속성의 개수에는 지수적이어서 변수가 많은 자료에는 부적절하다. 표본을 뽑는 방법은 시간을 줄이는 데에는 효과적이지만 표본 크기가 매우 작으면 실제와 큰 괴리가 있는 결과를 가져올 수 있다.

(참고 2) NN 기반 이상 탐지 기법의 장단점

- 비지도 학습 기반이며, 자료에 대해 어떠한 가정도 필요하지 않다.
- 이상값이 학습 집합에서 가까운 근방을 형성할 확률이 매우 낮아 준지도 기법이 비지도 기법보다 이상값을 잘 찾아내는 면에서 더 우수하다.
- 거리만 잘 정의되어 있으면 자료의 형태에 구애받지 않는다.
- 정상값이 가까운 이웃이 없다거나, 이상값이 가까운 이웃이 있는 상황에서 비지도 기법은 이상을 정상으로 보는 오류를 범한다.
- 준지도 기법을 적용할 때 학습 자료에서 없었던 패턴의 정상값이 테스트 집합에서 나타나면 이상으로 처리할 가능성이 높다.
- 테스트 과정에서 각각의 근방을 구해야 하므로 오랜 시간이 걸린다.
- 성능이 거리 척도의 영향을 많이 받으며, 그래프, 순차 자료같이 형태가 복잡하면 거리를 정의하기 까다로울 수 있다.
  - 일반적인 유클리드 거리를 사용할 경우, 각 변수별로 척도의 차이가 존재하기 때문에 왜곡이 일어날 수 있으며, 표준화를 통해 이를 바로잡더라도 범주형 변수가 존재하는 경우에는 가변수를 생성하는 등 다른 방식으로 문제를 해결해야 한다.
- 자료의 차원이 높아질수록 거리를 계산하는 데 있어 어려움이 많이 따른다.

3. 군집화(clustering) 기반 이상 탐지 기법

가. 근본 가정에 따른 군집화 기반 이상 탐지 기법

군집화(Jain & Dubes, 1988; Tan et al., 2005)는 비슷한 개체들을 군집으로 형성하여 탐색적 자료 분석과 자료 시각화를 위해 사용된다. 군집화는 원래 비지도 기법이지만 준지도 군집화(Basu et al., 2004)도 최근 연구되고 있다. 군집화와 이상 탐지는 서로 완전히 다른 것처럼 보이지만, 지금까지 많은 군집화 기반 이상 탐지 기법이 개발되었다. 이 분야의 기법들은 근본 가정에 따라 세 종류로 나누어진다.

첫 번째 그룹은 **'정상값들은 하나 또는 몇 개의 군집에 모여 있고, 이상값은 군집에 속하지 않는다'**고 가정한다. 이 가정을 바탕으로 한 기법은 모든 개체를 군집에 넣지 않아도 되는 DBSCAN(Ester et al., 1996), ROCK(Guha et al., 2000), SNN 군집화(Ertöz et al., 2003) 등의 알고리즘을 활용한다. FindOut 알고리즘(Yu et al., 2002)은 WaveCluster 알고리즘(Sheikholeslami et al., 1998)의 확장으로, 자료에서 군집을 찾아낸 뒤 제거하고 나머지를 이상값으로 처리한다. 이 기법들은 군집을 찾아내는 것이 주된 목적이기 때문에 이상 탐지에 최적화되어 있지 않다는 단점이 있다.

두 번째 그룹은 **'군집의 중심(centroid) 중 가장 가까운 것과의 거리가 짧으면 정상값, 길면 이상값이다'라**는 생각을 바탕으로 한다. 먼저 군집화를 하고 개체가 포함된 군집의 중심과 개체 사이의 거리를 이상 점수로 놓는 것이 기본 과정이다. Smith et al.(2002)은 자기 조직화 지도(self-organizing map, SOM), k-평균 군집화, EM 알고리즘을 군집화에 이용하였다. 특히 자기 조직화 지도(Kohonen, 1997)는 준지도 방식

으로 침입 탐지(Labib & Vemuri, 2002; Smith et al., 2002; Ramadas et al., 2003), 오작동 탐지(Harris, 1993; Ypma & Duin, 1998; Emamian et al., 2000), 사기 탐지(Brockett et al., 1998) 등 분야에 널리 이용되었다.

또한 이 그룹의 기법들은 학습 자료를 군집화하고 테스트 개체를 군집과 비교해 이상 점수를 얻는 식으로 준지도 기법이 될 수 있다 (Marchette, 1999; Wu & Zhang, 2003; Vinueza & Grudic, 2004; Allan et al., 1998). 학습 자료가 여러 클래스로 이루어진 경우는 준지도 군집화를 쓸 수 있다. He et al.(2002)은 비지도적 군집화 기반 이상 탐지 기법(He et al., 2003)에 의미 이상값 지수(semantic anomaly factor)를 도입해 라벨을 활용하였다. 의미 이상값 지수는 어떤 개체의 클래스 라벨이 그 개체가 속한 군집에서 다수를 차지하는 클래스와 다를 때 큰 값을 보인다. 이와 같은 기법은 이상값들이 군집을 이룰 때 매우 취약한데, 마지막 그룹의 기법으로 이 문제를 해결할 수 있다.

'정상값은 크거나 조밀한 군집에, 이상값은 작거나 한산한(sparse) 군집에 속한다'는 가정을 기반으로 한 기법이 마지막 그룹에 포함된다. 개체가 속한 군집의 크기나 밀도가 이상 여부를 판단하는 기준이 된다. 여러 응용 기법이 제안되었는데(Pires & Santos-Pereira, 2005; Otey et al., 2003; Eskin et al., 2002; Mahoney et al., 2003; Jiang et al., 2001; He et al., 2003), He et al.(2003)의 기법 FindCBLOF에서는 개체가 속한 군집의 크기 및 개체와 그 군집의 중심 사이의 거리를 반영하는 CBLOF(cluster-based local outlier factor)라는 이상 점수를 각 개체에 부여한다.

개발된 기법들의 효율을 늘리는 방안도 여럿 제시되었다. 먼저 선형 시간($O(Nd)$) 근사 알고리즘인 고정 폭(fixed width) 군집화가 여러 기법

에 활용되었다(Eskin et al., 2002; Portnoy et al., 2001; Mahoney et al., 2003; He et al., 2003). 어떤 군집의 중심이 새로운 개체에서 미리 정해진 거리 이내에 있으면 개체를 그 군집에 포함하고, 그러한 군집이 없으면 그 개체를 중심으로 하는 새로운 군집을 만든다. 군집화가 끝나면 어떤 군집의 개체들이 이상값인지를 군집의 밀도와 다른 군집과의 거리를 바탕으로 정한다. 고정된 폭은 일반적으로 분석자가 정한다(Eskin et al., 2002; Portnoy et al., 2001). Chaudhary et al.(2002)은 $k-d$ 트리로 분할을 선형 시간에 끝내는 기법을 제안해 계산 효율이 매우 중요한 천문학 자료에 활용하였다. Sun et al.(2004)은 CD-트리로 자료를 효율적으로 분할한 뒤, 한산한 군집에 속한 개체를 이상값으로 보았다.

여러 군집화 기반 이상 탐지 기법은 한 쌍의 관측치들 간의 거리 계산이 필요하다. 따라서, 거리 계산이 필요하다는 점에서 NN 기반 이상 탐지 기법과 유사하다. 거리 측정의 선택은 기법의 성능에 중요하기 때문에 거리 측정에 관한 이슈는 군집화 기반 이상 탐지 기법에도 적용된다. 그러나 두 기법 간의 주요 차이점은 군집화 기반 이상 탐지 기법이 자신이 속한 클러스터와 관련하여 각 관측치를 평가하는 반면, NN 기반 이상 탐지 기법은 가까이 존재하는 로컬 이웃에 대해 각 관측치를 분석한다는 것이다.

나. 계산 복잡도 및 장단점 비교

(참고 1) 계산 복잡도

학습 과정의 복잡도는 사용하는 군집화 알고리즘에 의존한다. 각 개체 쌍의 거리를 모두 구해야 하는 제곱 시간 알고리즘도 존재하는 반면, k-평균 군집화(Hartigan & Wong, 1979)나 근사 군집화(Eskin et al.,

2002)처럼 경험적 기법(heuristic based)은 선형 시간에 해결한다. 테스트 과정은 적은 수의 군집의 중심들과 비교하면 충분하기 때문에 빠르게 진행된다.

(참고 2) 군집화 기반 이상 탐지 기법의 장단점

- 비지도 방식으로 작동된다.
- 적당한 군집화 알고리즘만 있으면 복잡한 자료도 다룰 수 있다.
- 테스트 과정이 빠르다.
- 성능이 군집화 알고리즘이 정상 개체의 군집을 얼마나 잘 잡아내는지에 달려 있다.
- 많은 방법이 군집화의 부산물로 이상 탐지를 진행해, 이상 탐지에 최적화되어 있지 않다.
- 다수의 군집화 알고리즘이 모든 개체에 군집을 지정하기 때문에 이상값이 큰 군집에 들어가 정상값으로 판단될 수 있다.
- 이상값이 군집을 이루는 경우 극도로 취약해지는 기법들이 있다.
- 특히 $O(N^2d)$ 알고리즘을 쓰는 경우 군집화 과정에 오랜 시간을 소모한다.

4. 이상 탐지의 통계적 기법

통계적 기법의 근본적 원칙은 **'이상값은 가정된 확률분포에서 생성되지 않아 부분적으로, 또는 완전히 동떨어졌다고 여겨지는 관측값이다'**(Anscombe & Guttman, 1960)라는 것이다. 그리고 **'이상값은 확률분포에서 낮은 영역에 나타난다'**고 가정한다.

통계적 기법은 주어진 자료로 (보통 정상값의) 모형을 적합한 뒤 통계

적 추론을 통해 새로운 개체가 그 모형을 따르는지를 판단한다. 검정 통계량을 바탕으로 학습된 모형으로부터 생성되었을 확률이 낮은 개체를 이상값으로 본다. 모수적, 비모수적 기법 모두 적용할 수 있다. 모수적 기법은 분포의 꼴을 미리 알고 있다고 가정하고 모수를 추정하는 반면 (Eskin, 2000), 비모수적 기법은 일반적으로 분포에 대한 가정이 없다 (Desforges et al., 1998).

가. 모수적 기법

먼저, 모수적 기법에서는 모수가 Θ인 확률밀도함수 $f(\cdot, \Theta)$에서 정상 자료가 생성되었다고 가정한다. 테스트 개체 $\boldsymbol{x}$의 이상 점수는 확률밀도함수의 값 $f(\boldsymbol{x}, \hat{\Theta})$의 역수로 주어지고, 여기서 $\hat{\Theta}$은 주어진 학습 자료로 추정된 모수이다. 이를 대신해 가설 검정을 이용할 수 있다. 테스트 개체가 추정된 분포에서 생성되었다는 것을 귀무가설로 한다. 이때 가설 검정에 사용한 검정 통계량을 이상 점수에 활용할 수 있다.

모수적 기법은 분포의 종류에 따라 다시 나눌 수 있다.

1) 정규모형 기반

자료가 정규모형에서 생성된 것으로 생각하고, 모수는 최대가능도 추정량(maximum likelihood estimator, MLE)을 사용한다. 각 개체와 추정된 평균 사이의 거리가 이상 점수가 되고, 이상 점수의 경계를 정해 이상값 여부를 결정한다. 거리의 정의와 경계에 대해 다양한 방법들이 제안되었다. 이 중 간단한 방법인 상자 그림 규칙(box plot rule)은 의학 (Laurikkala et al., 2000; Horn et al., 2001; Solberg & Lahti,

2005), 터빈 로터 자료(Guttormsson et al., 1999)의 이상값을 찾는 데 활용되었다. 상자 그림은 최소 정상값(min), 1사분위수(Q_1), 중앙값(median), 3사분위수(Q_3), 최대 정상값(max)을 나타낸다. 3사분위수와 1사분위수의 차($Q_3 - Q_1$)를 사분위수 범위(inter quartile range, IQR)라 하는데, 범위$[Q_1 - 1.5IQR, Q_3 + 1.5\,IQR]$ 밖에 위치한 개체들을 이상값으로 판단한다.

Grubbs 검정은 정규분포를 가정한 일변량 자료의 이상 탐지에 이용되었다(Grubbs, 1969; Stefansky, 1972; Anscombe & Guttman, 1960). 표본의 평균과 표준편차를 $\overline{\boldsymbol{x}}$, s라 할 때, 테스트 개체 $\boldsymbol{x}$의 z 점수는

$$z = \frac{|\boldsymbol{x} - \overline{\boldsymbol{x}}|}{s}$$

로 주어지고,

$$z > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

이면 이상값으로 처리한다. 여기서 N은 자료의 크기, $t_{\delta,n}$은 자유도 n인 t 분포의 δ분위수이다. α는 신뢰 수준을 나타내며 이상값의 빈도를 간접적으로 조정하는 역할을 한다.

Laurikkala et al.(2000)은 다변량 자료에 대한 Grubbs 검정의 변형을 제시하였는데, 테스트 개체 $\boldsymbol{x}$의 표본평균 $\overline{\boldsymbol{x}}$에 대한 Mahalanobis 거리로 자료를 일변량으로 줄이는 방법이다. 즉, 표본분산행렬 S가 역행렬이 존재할 때

$$y^2 = (\boldsymbol{x} - \overline{\boldsymbol{x}})' S^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}})$$

을 구하고, y들에 실시한 Grubbs 검정 결과를 $\boldsymbol{x}$에 그대로 가져오는 방식이다. 다른 Grubbs 검정의 응용이 다변량(Aggarwal & Yu, 2001, 2008; Laurikkala et al., 2000), 그래프(Shekhar et al., 2001), OLAP(online analytical processing) 큐브(Sarawagi et al., 1998) 자료 처리를 목적으로 발표되었다.

이 외에도 Student t 검정(Surace & Worden, 1998; Surace et al., 1997), Hotelling t 검정(Liu & Weng, 1991), 카이제곱 검정(Ye & Chen, 2001), Rosner 검정(Rosner, 1983), Dixon 검정(Gibbons, 1994), 미끄러짐 검출(slippage detection) 검정(Hawkins, 1980) 등이 활용되었다.

2) 회귀모형 기반

회귀모형 기반 이상 탐지는 시계열 자료에 대해 널리 연구되었다(Abraham & Chuang, 1989; Abraham & Box, 1979; Fox, 1972). 기본적 기법은 자료에 회귀모형을 적합한 뒤 그 모형에 대한 테스트 개체의 잔차(residual)로 이상 점수를 구한다. 잔차란 회귀모형으로 설명되지 않는 부분을 뜻한다. 잔차의 크기를 그대로 이상 점수로 할 수도 있지만, 신뢰도로 이상값을 정하는 여러 통계적 검정들이 있다(Anscombe & Guttman, 1960; Beckman & Cook, 1983; Hawkins, 1980; Torr & Murray, 1993).

하지만 학습 자료에 이상값이 있으면 회귀 모수 및 회귀모형에 영향을 줄 수 있다. 이러한 문제에 주로 쓰이는 방안이 로버스트(robust) 회귀(Rousseeuw & Leroy, 1987)이고, 이상값이 로버스트한 적합에 대해 큰 잔차를 가지는 경향이 있어, 로버스트 회귀를 통해 이상값을 가려냄과

동시에 발견할 수 있다. ARIMA(autoregressive integrated moving average) 모형에 이와 비슷한 기법이 적용되었다(Bianco et al., 2001; Chen et al., 2005).

다변량 시계열 자료에도 이러한 기법의 변형이 이용되었는데, Tsay et al.(2000)은 일변량에 비해 다변량 시계열 자료가 가지는 복잡함을 논의하고 다변량 ARIMA 모형의 이상 탐지에 쓰일 수 있는 통계량을 제시하였다. 이것은 Fox(1972)에서 제시된 통계량의 일반화이기도 하다.

또한 Galeano et al.(2006)이 다변량 ARMA(autoregressive moving average) 자료에 대한 기법을 제안하였다. 변수들에 대해 선형결합을 진행하여 자료를 일변량 자료로 바꾸는데, 선형결합은 첨도(kurtosis)를 최대화하는 사영 추적(projection pursuit) 기법(Huber, 1985)으로 얻는다. 이 일변량 자료의 이상 탐지는 Fox(1972)에서 제시된 검정 통계량으로 시행된다.

3) 혼합 모수적 모형 기반

해당 범주의 기법은 자료에 모수적 분포들을 혼합한 모형을 이용한다. 여기서 다시 두 부분으로 나눌 수 있는데, 하나는 정상값과 이상값에 서로 다른 분포를 주는 방법이고, 다른 하나는 정상값에만 혼합 분포를 주는 방법이다.

전자의 기법들은 테스트 과정에서 개체가 정상값 분포와 이상값 분포 중 어디에 속하는지를 결정한다. Abraham & Box(1979)는 두 집단 모두 평균이 같은 정규분포에서 생성되었지만, 이상값 분포의 분산이 더 크다고 가정하였다. 테스트 개체에는 두 분포에 대한 Grubbs 검정을 해서 판단을 내린다. Lauer(2001), Eskin(2000), Abraham & Box(1979), Box & Tiao(1968), Agarwal(2005)이 이와 유사한 기법을 선보였다.

Eskin(2000)은 이상값, 정상값의 사전확률을 각각 λ, $1-\lambda$로 놓고 EM(expectation maximization) 알고리즘으로 두 집단에 대한 모형을 구하였다. 따라서 D, A, M을 각각 전체 자료, 이상값, 정상값의 분포라 하면 $D=\lambda A+(1-\lambda)M$이 성립한다. M은 분포 추정 기법으로 구하지만 A는 균등분포라 가정한다. 처음에는 모든 개체를 정상으로 놓고, 한 점을 제거했을 때 분포가 얼마나 달라지는지로 이상 점수를 정한다.

또한 후자의 기법들은 정상값의 혼합 모형을 구하고 어떠한 모형도 따르지 않는 개체를 이상으로 판단한다. 혼합 정규모형이 널리 쓰이는데(Agarwal, 2006), 기체 변형 감지(Hickinbotham & Austin, 2000a; Hollier & Austin, 2002), 유방 X선 이미지 분석(Spence et al., 2001; Tarassenko, 1995), 네트워크 침입 탐지(Yamanishi & ichi Takeuchi, 2001; Yamanishi et al., 2004) 등에 이용되었다. 또 유사한 기법이 생체신호(biomedical signal) 자료에 적용되었는데(Roberts & Tarassenko, 1994; Roberts, 1999, 2002), 테스트 과정에서 극단값 통계(extreme value statistics)가 활용되었다.

나. 비모수적 기법

비모수적 기법은 모형의 구조를 사전에 정하지 않고 자료를 통해 구한다. 몇몇 기법은 밀도함수의 매끄러움(smoothness)과 같은 가정이 있기는 하지만, 모수적 기법보다는 일반적으로 가정이 적다.

1) 히스토그램 기반

히스토그램은 정상 자료의 개요(profile)를 살피는 가장 간단한 기법이라 할 수 있고, 특히 침입 탐지(Eskin, 2000; Eskin et al., 2001; Denning,

1987)나 사기 탐지(Fawcett & Provost, 1999)와 같이 (고객, 소프트웨어 또는 시스템의) 프로필이 자료의 행동적 속성을 제한하는 분야에서 유용하다.

학습 자료로 히스토그램을 그린 뒤 테스트 개체가 유의미한 구간(bin)에 포함되면 정상으로, 그렇지 않으면 이상으로 하는 것이 일반적 방법이다. 이때 개체가 포함되는 구간의 빈도(frequency)를 바탕으로 이상 점수를 구하기도 한다.

여기서 구간의 폭(size)을 적절하게 정하는 것이 매우 중요한데, 너무 작으면 정상값이 빈 구간에 들어가 이상으로 판단되어 오경보율이 높아지고, 너무 크면 이상값도 의미 있는 구간에 포함되어 가음성률(false negative rate)이 높아진다.

히스토그램을 그리는 데에는 정상 자료가 필요하다(Anderson et al., 1994; Javitz & Valdes, 1991; Helman & Bhangoo, 1997). 이상값 라벨이 있는 경우 이상값의 히스토그램을 그리기도 한다(Dasgupta & Nino, 2000).

자료가 다변량이면 각 속성에 대한 히스토그램으로 이상 점수를 구한 뒤 그것들을 모아 총 이상 점수를 구하는 방법이 기본적이다. 이 기법은 시스템 호출 침입(Endler, 1998), 네트워크 침입(Ho et al., 1999; Yamanishi & ichi Takeuchi, 2001; Yamanishi et al., 2004), 사기(Fawcett & Provost, 1999), 구조물 손상(Manson, 2002; Manson et al., 2001, 2000), 웹 기반 공격(Kruegel & Vigna, 2003; Kruegel et al., 2002), 문자 자료의 새로운 주제(Allan et al., 1998)를 탐지하는 데 활용되었다. 이 기법의 변형으로 Mahoney & Chan(2002)이 제안한 PHAD(packet header anomaly detection)와 ALAD(application layer anomaly detection)가 있으며, 네트워크 침입 탐지에 적용되었다.

SRI International 사의 실시간 네트워크 침입 탐지 시스템(network intrusion detection system, NIDES)(Anderson et al., 1994; Anderson et al., 1995; Porras & Neumann, 1997)은 컴퓨터 시스템의 정상 행동을 잡아내는 장기 통계적 프로필을 관리하는 하부조직(subsystem)을 보유하고 있다(Javitz & Valdes, 1991). 해당 연구의 저자는 장기 프로필과 단기 프로필의 비교에 Q 통계량을 사용한다. Q 통계량은 또 다른 S 통계량의 계산에 쓰이는데, S 통계량은 어떤 특성이 과거의 프로필에 대비해 비정상인 정도를 나타낸다. 각 특성의 S 통계량을 모아 하나의 IS 통계량 값을 얻어 이것으로 이상 여부를 정한다. Sargor (1998)는 이를 변형해 링크 상태 라우팅 프로토콜(link-state routing protocol)의 이상 탐지에 이용하였다.

2) 커널 함수 기반

비모수적 밀도함수 추정 기법으로 커널 함수를 이용한 Parzen 창(window)(Parzen, 1962)이 있다. 커널 함수 기반 기법은 모수적 기법과 매우 비슷하고, 유일한 차이는 사용하는 밀도함수 추정 기법에 있다. Desforges et al.(1998)은 정상값의 밀도함수를 커널 함수로써 추정하는 준지도 통계적 기법을 제시하였다.

Parzen 창의 응용 기법이 네트워크 침입(Chow & Yeung, 2002), 기름의 유동(oil flow) 자료(Bishop, 1994), 유방 X선 이미지(Tarassenko, 1995)의 이상 탐지에 활용되었다.

다. 계산 복잡도 및 장단점 비교

(참고 1) 계산 복잡도

사용하고자 하는 모형의 종류에 따라 복잡도가 달라진다. 정규, 포아송, 다항처럼 간단한 분포를 쓰면 보통 자료 크기에 선형이며, 혼합 모형이나 은닉 마코프 모형(hidden Markov model) 같은 복잡한 모형은 추정에 반복적(iterative)인 계산을 요구해 수렴 속도와 기준에 따라서 오랜 시간이 걸릴 수 있다. 커널 기반 기법은 자료 크기에 대해 잠재적으로 제곱 시간 알고리즘이다.

(참고 2) 이상 탐지의 통계적 기법의 장단점

- 분포에 대한 가정이 맞으면 통계적으로 적합한 모형을 확보할 수 있다.
- 이상 점수를 구하는 과정에서 신뢰 구간과 같이 의사 결정에 도움이 될 추가 정보를 얻는다.
- 분포 추정이 이상값에 대해 로버스트하면 비지도 학습이 가능하다.
- 자료가 특정한 분포를 따른다는 가정이 성립하지 않을 때가 많고, 자료가 고차원일 때 특히 그러하다.
- 적당한 가설 검정 통계량을 정하는 일이 까다롭고(Motulsky, 1995), 분포가 복잡해지면 가설을 세우기조차 쉽지 않다.
- 히스토그램으로는 변수 사이의 교호작용을 볼 수 없다.
  - 어떤 개체가 두 속성 각각에서 흔한 값을 가지면서도 그 조합은 매우 드문 경우, 히스토그램 기반 기법은 이것을 이상값으로 잡아내지 못할 것이다.

5. 정보 이론 이상 탐지 기법

가. 정보 이론 기법

이 분야의 기법은 Kolmogorov 복잡도, 엔트로피, 상대 엔트로피와 같은 척도로 구한 자료의 정보량(information content)을 분석한다. **'이상값은 정보량의 불규칙을 유발한다'**고 가정하고 아래의 흐름을 따른다.

주어진 자료 집합 D의 복잡도를 $C(D)$라 한다. $C(D)-C(D-I)$을 크게 하면서 충분히 작은 $I \subset D$를 찾아 I에 속한 개체들을 이상값으로 본다. 즉 복잡도와 I의 크기, 두 방향의 최적화가 필요한 Pareto 최적화 문제를 풀어야 한다.

복잡도 함수(C)에는 여러 선택지가 있다. 먼저 Kolmogorov 복잡도(Li & Vitanyi, 1993)가 몇몇 기법(Arning et al., 1996; Keogh et al., 2004)에 쓰였다. Arning et al.(1996)은 정규 표현식(regular expression)의 길이를, Keogh et al.(2004)은 압축된 파일의 크기를 Kolmogorov 복잡도 계산에 이용하였다. 또한 엔트로피나 상대 불확정도 같은 척도가 범주형 자료의 복잡도를 구하는 데 쓰였다(Lee & Xiang, 2001; He et al., 2005, 2006; Ando, 2007).

앞에서 언급했듯이 기본 기법에서 복잡도와 I의 크기에 대한 최적화를 한꺼번에 실시해야 한다. 하지만 모든 부분집합을 고려하는 것은 지수적 시간이 필요해 사실상 불가능하므로 적당한 일부만을 살펴 근사 해를 찾는 전략이 필요하다. He et al.(2006)은 엔트로피를 척도로 LSA(local search algorithm)(He et al., 2005)로써 선형 시간에 근사 최적해를 구하였다. Ando(2007)는 정보 병목 방법(information bottleneck method)을 써서 비슷한 기법을 제안하였다.

순차 자료나 공간 자료처럼 자연스럽게 순서가 부여되는 자료에도 정보 이론 기법이 이용되었다. 이 경우에는 자료를 부분구조(substructure)들로 분해하고 $C(D) - C(D - I)$을 최대화하는 부분구조 I를 구한다. 이 기법은 순열(Lin et al., 2005; Chakrabarti et al., 1998; Arning et al., 1996), 그래프(Noble & Cook, 2003), 공간 자료(Lin & Brown, 2003)에 활용되었다. 이 기법에서 까다로운 부분은 부분구조의 가장 적절한 크기를 찾는 것이다.

나. 계산 복잡도 및 장단점 비교

(참고 1) 계산 복잡도

앞에서도 언급했듯이 기본 기법은 지수적 시간이 걸리지만, 근사 기법을 쓰면 선형 시간까지 줄일 수 있다.

(참고 2) 정보 이론 이상 탐지 기법의 장단점

- 비지도 방식이 가능하다.
- 자료의 분포에 대한 가정이 필요하지 않다.
- 성능이 정보량 척도에 크게 의존하며, 이상값이 많이 있어야 검출이 가능할 때가 많다.
- 순차, 공간 자료에 대한 기법의 성능은 부분구조의 크기에 영향을 받는데, 최적의 크기를 구하기 쉽지 않다.
- 각 개체의 이상 점수를 구하기 까다롭다.

6. 스펙트럴 이상 탐지 기법

가. 스펙트럴 기법

스펙트럴 기법은 변수들의 조합으로 자료의 변동(variability)을 대부분 설명하도록 자료를 근사한다. 이러한 기법에서는 **'자료를 더 낮은 차원의 부분공간으로 보낸(embed) 뒤, 그 공간에서는 정상과 이상이 확연히 구분된다'**고 가정한다. 이상값을 쉽게 찾아낼 만한 부분공간을 정하는 것이 일반적인 접근이다(Agovic et al., 2007). 스펙트럴 기법은 비지도나 준지도 설정에서 작동한다.

많은 기법이 주성분 분석(principal component analysis, PCA) (Jolliffe, 2002)을 써서 자료를 저차원 공간으로 사영시킨다. 그중 하나(Parra et al., 1996)는 분산이 낮은 주성분들로 구성된 공간을 활용한다. 자료의 상관 구조(correlation structure)를 만족하는 정상 개체는 사영값이 낮을 것이고, 상관 구조에서 벗어난 이상 개체는 높은 값을 보일 것이다. Dutta et al.(2007)은 천체 목록의 이상 탐지에 이러한 방식으로 접근하였다.

Ide & Kashima(2004)는 그래프의 시계열에 대한 스펙트럴 기법을 제안하였다. 각 그래프는 인접 행렬(adjacency matrix)을 가지는데, 인접 행렬의 첫 번째 주성분을 그래프의 활성 벡터(activity vector)로 한다. 그러면 활성 벡터들을 차례대로 모아 행렬 하나를 얻는데, 그 행렬의 첫 번째 왼쪽 특이 벡터(left singular vector) $\boldsymbol{v}$가 시계열의 정상성을 보여 준다고 생각한다. 테스트 그래프를 받으면 그것의 활성 벡터와 $\boldsymbol{v}$가 이루는 각도를 이상 점수에 활용한다. 비슷한 접근으로, Sun et al.(2007)은 각 그래프의 인접 행렬을 콤팩트 행렬 분해(compact matrix de-

composition, CMD)로 근사한 뒤, 그 오차들로 이루어진 시계열을 얻어 이상 탐지를 하고 이상한 오차에 해당하는 그래프를 이상값으로 취급한다.

Shyu et al.(2003)은 로버스트 PCA(Huber, 2011)로 정상 자료의 공분산 행렬에서 주성분의 근삿값을 구한다. 테스트 과정은 개체가 주성분 방향으로 얼마나 멀리 위치했는지를 살핀다. 즉, 개체 x를 고유값 λ_1, λ_2, $\cdots$, λ_p에 해당하는 주성분에 사영한 값이 $y_1, y_2, \cdots, y_p$이면 $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i}$, $q \leq p$는 카이제곱 분포를 따른다(Hawkins, 1974). 따라서 유의수준을 α로 두면

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > \chi_q^2(\alpha)$$

일 때 x가 이상값이 된다. 만약 $q = p$라면 $\sum_{i=1}^{p} \frac{y_i^2}{\lambda_i}$은 x와 표본평균의 Mahalanobis 거리가 된다. 따라서 이 경우 로버스트 PCA 기반 기법은 이상 탐지의 통계적 기법 중 모수적 기법인 정규모형 기반에서 논의한 방법과 같다고 볼 수 있다.

로버스트 PCA 기반 기법은 네트워크 침입(Shyu et al., 2003; Lakhina et al., 2005; Thottan & Ji, 2003)과 우주선 부품 이상(Fujimaki et al., 2005) 탐지에 활용되었다.

나. 계산 복잡도 및 장단점 비교

(참고 1) 계산 복잡도

주성분 분석 기반 기법은 일반적으로 자료 크기에는 선형 시간이고 차원에는 제곱 시간이다. 비선형 기법을 쓰면 차원에 선형이 되도록 줄일 수 있지만, 대신 주성분 개수에 대해 다항식 시간이 된다(Gunter et al., 2007). 특이값 분해를 사용하는 기법은 자료 크기에 제곱 시간이다.

(참고 2) 스펙트럴 이상 탐지 기법의 장단점

- 고차원 자료 처리에 적합하며, 다른 기법의 적용을 위한 전처리처럼 사용할 수도 있다.
- 비지도 학습이 가능하다.
- 자료를 보낼 저차원 공간에서 정상, 이상값이 제대로 분리되어야만 유용하다.
- 보통 계산 복잡도가 크다.

7. 맥락적 이상 탐지

지금까지 소개된 기법들은 특정 개체 이상 즉, 점 이상에만 초점을 맞추었지만, 여기에서는 맥락적 이상을 다루는 방법을 논의하고자 한다.

앞에서 논의하였듯이, 맥락적 이상은 자료가 (맥락을 정의하는) 맥락적 속성과 (맥락 안에서 이상값을 찾기 위한) 행동적 속성을 가져야만 의미가 있다. Song et al.(2007)은 같은 의미로 환경적(environmental) 속성, 지표(indicator) 속성이라는 용어를 사용하였다. 다음은 맥락적 속성이 정의되는 몇 가지 과정이다.

1) 공간 자료

위치와 근방을 정하는 공간적 속성을 가진다. 공간 자료에 대한 많은 맥락적 이상 탐지 방법들이 제안되었다(Lu et al., 2003; Shekhar et al., 2001; Kou et al., 2006; Sun & Chawla, 2004).

2) 그래프 자료

해당 점과 변으로 이어진 점들을 근방으로 볼 수 있다. Sun et al.(2005)이 그래프 자료에 대한 맥락적 이상 탐지 기법을 발표하였다.

3) 순차 자료

개체의 위치가 맥락적 속성이 된다. 특히 시계열 자료에 대해 많은 연구가 진행되었다(Abraham & Chuang, 1989; Abraham & Box, 1979; Rousseeuw & Leroy, 1987; Bianco et al., 2001; Fox, 1972; Salvador & Chan, 2003; Tsay et al., 2000; Galeano et al., 2006; Zeevi et al., 1997). 순차 자료의 다른 예시로는 시스템 호출 자료나 웹 자료(Ilgun et al., 1995; Vilalta & Ma, 2002; Weiss & Hirsh, 1998; Smyth, 1994)와 같이 시간이 기록된 사건(event) 자료이며, 시계열 자료와 달리 사건 사이의 간격이 일정하지 않을 수 있다.

점 이상 탐지 기법에 대한 논문은 매우 많은 데 반해 맥락적 이상 탐지 연구 자료는 아직 한정되어 있다. 맥락적 이상 탐지 기법은 맥락적 이상을 점 이상 문제로 축소시키는 방법과 자료의 구조를 모형화하는 방법으로 나눌 수 있다.

가. 점 이상 문제로 축소

맥락적 이상도 점 이상처럼 개별적인 개체이고, 맥락에 대해 이상할 뿐이므로 맥락 내 점 이상 탐지 기법을 이용하는 방식으로 접근한다. 각 개체에 대한 맥락을 찾은 뒤, 그 맥락 안에서 점 이상 탐지 기법으로 해당 개체가 이상인지를 판정하는 것이 포괄적으로 사용할 수 있는 축소 기반 기법이다.

한 예시로 맥락을 잡아내기 힘들 때 유용한 Song et al.(2007)의 기법이 있다. 먼저 맥락적 속성과 행동적 속성은 사전에 구분되어 있다고 가정하면, 각 개체 d를 $[x,y]$로 표현할 수 있고 여기서 x는 맥락적 속성을, y는 행동적 속성을 의미한다. 그리고 맥락적 부분과 행동적 부분은 각각 서로 다른 혼합 정규모형 U, V를 따른다고 가정한다. 이때 맥락적 부분 x가 U의 구성요소 U_i에서 생성되었을 때 행동적 부분 y가 V의 구성요소 V_i에서 생성되는 조건부확률 $p(V_j|U_i)$를 같이 학습한다. 그리고 $d=[x,y]$의 이상 점수를

$$\sum_i \sum_j p(x \in U_i)p(y \in V_j)p(V_j|U_i)$$

로 놓는다. $p(x \in U_i)$, $p(y \in V_j)$는 각각 x가 U_i에서, y가 V_j에서 생성되었을 확률을 뜻한다.

또 다른 기법은 휴대전화 사기 탐지(Fawcett & Provost, 1999)에 적용되었다. 자료는 휴대전화 사용 기록이고, 사용자를 맥락적 속성으로 본다. 다른 속성들로 이상값을 탐지하기 위해 각 사용자의 활동을 살핀다. 사용자의 ID와 사용 시각을 맥락적 속성으로 하는 비슷한 기법이 컴퓨터 보안에 활용되었다(Teng et al., 1990). 여기서는 나머지 속성을 정상값 결정 규칙과 비교해 이상 여부를 결정하였다.

동료 집단 분석(Bolton & Hand, 1999) 역시 유사한 기법으로, 사용자를 동료 집단으로 묶은 뒤 그 집단에서 분석하는 방식이다. He et al.(2004b)은 클래스 이상 탐지(class anomaly detection)라는 개념을 도입하였다. 라벨을 이용해 자료를 세분화(segmenting)한 뒤 그 안에서 알려진 군집화 이상 탐지 기법(He et al., 2002)을 적용하였다.

공간 자료는 위치 좌표로 쉽게 근방을 구할 수 있다. 그래프 기반 이상 탐지 기법(Shekhar et al., 2001; Lu et al., 2003; Kou et al., 2006)은 Grubbs 검정(Grubbs, 1969)이나 다른 통계적 점 이상 탐지 방법으로 근방 안에서 이상값을 찾는다. Sun & Chawla(2004)는 척도 SLOM[Spatial Local Outlier Measure(Sun & Chawla, 2006)]을 제시하였다.

Basu & Meckesheimer(2007)는 시계열 자료에서 각 개체와 그 근방의 중간값을 비교하는 기법을 제안하였다. 위상 공간(phase space)을 이용해 시계열 자료를 변환하는 기법(Ma & perkins, 2003b)은 먼저 시간 지연 삽입(time delay embedding)으로 자료를 벡터 집합으로 바꾼다. 이때 각 개체의 시간적 관계는 해당하는 위상 벡터에 옮겨진다. 그리고 변환된 집합에서 SVM으로 이상을 탐지한다.

나. 자료 구조 활용

많은 경우 자료를 맥락들로 나누기가 쉽지 않다. 시계열 자료나 사건열 자료가 이에 해당한다. 자료들을 맥락으로 나누기 어려운 경우에는 모형화 기법을 확장해 이상 탐지에 쓸 수 있다. 과정은 다음과 같다. 학습 자료에서 모형을 학습해 주어진 맥락에서의 행동을 예측한다. 구한 예측값과 관측값의 차이가 유의미하면 이상값으로 본다. 회귀분석이 간단한 예시로, 맥락적 속성이 행동적 속성 예측에 이용될 수 있다.

로버스트 회귀(Rousseeuw & Leroy, 1987), 자기회귀 모형(Fox, 1972), ARMA(Abraham & Chuang, 1989; Abraham & Box, 1979; Galeano et al., 2006; Zeevi et al., 1997), ARIMA(Bianco et al., 2001; Tsay et al., 2000) 모형에 대한 회귀 기반 기법이 이상 탐지에 활용되었다. 공진화열(coevolving sequence)에 대한 회귀모형 및 수열 사이의 상관관계 모형을 적용한 기법도 연구되었다(Yi et al., 2000).

시계열 자료 이상 탐지에서 가장 초기의 연구 중 하나는 정적(stationary) 자기회귀 모형을 적용한 Fox(1972)의 방법이다. 각 개체를 자기회귀 과정의 공분산 행렬과 비교한 뒤 이상 여부를 정한다. 이 방법의 확장으로 SVM으로 회귀모수를 추정하는 기법이 있다(Ma & Perkins, 2003a).

Keogh et al.(2004)은 순차 자료에서 하나의 이상값을 찾는 기법을 제안했는데, 수열을 둘로 나눈 뒤 Kolmogorov 복잡도가 더 높은 부분을 골라 한 개체만 남을 때까지 이 과정을 반복한다.

Weiss & Hirsh(1998)는 순차 자료에서 희귀한 사건을 찾는 기법을 제시했는데, 특정 시각의 예측값을 그 이전에 일어났던 사건들로 구해 그것이 관측값과 다르면 해당 사건을 희귀한 것으로 판단한다. 여기서 조건부 확률을 구할 때 빈발 항목 마이닝(Vilalta & Ma, 2002), FSA(finite state automation)(Ilgun et al., 1995; Salvador & Chan, 2003), 마코프 모형(Smyth, 1994)을 이용하기도 하였다. Marceau(2000)는 FSA로 이전 n개 사건들을 바탕으로 다음 사건을 예측하는 기법을 시스템 호출 침입 탐지에 활용하였다. Hollmen & Tresp(1999)은 은닉 마코프 모형을 휴대전화 사기 탐지에 이용하였는데, 사용자의 활동을 계층적 국면 전환 모형(hierarchical regime switching model)으로 모형화한 뒤 사기 확률을 예측하였고, EM 알고리즘으로 모수를 추정하였다.

Scott(2001)와 Ihler et al.(2006)은 전화망 침입 탐지와 웹 클릭 자료에 대한 모형을 각각 제안하였다. 두 논문 모두 정상값은 비정적(nonstationary), 이상값은 동질적(homogeneous) 포아송 과정에서 생성되었다고 가정한 기법을 따랐다. 정상, 이상 사이의 전이(transition)는 마코프 과정으로 모형화하였고, 모수 추정에는 MCMC(Markov chain Monte Carlo)를 이용하였다.

P2P 망의 이분 그래프(bipartite graph) 구조가 그래프에서 각 점의 근방을 찾는 데 이용되었고(Sun et al., 2005), 그 후에 근방 안에서 그 점의 관련성을 구하고 관련성 점수가 낮은 점을 이상값으로 보았다. 또한 그래프를 METIS(Karypis & Kumar, 1998)와 같은 알고리즘으로 겹치지 않는 부분 그래프로 분할한 뒤 그 안에서 근방을 찾는 기법도 연구되었다.

다. 계산 복잡도 및 장단점 비교

(참고 1) 계산 복잡도

축소 기반 기법에서 학습 과정의 복잡도는 축소 알고리즘과 각 맥락에서 쓰는 점 이상 탐지 기법에 의존한다. 세분화나 분할 기법은 축소 과정이 빠르게 진행되며, 군집화 또는 혼합 모형 추정을 이용한 기법은 상대적으로 느리다. 점 이상 탐지 기법은 빠른 것을 사용하면 된다. 테스트 과정은 개체마다 맥락을 정하고 다시 점 이상 탐지 기법을 적용해야 하므로 상대적으로 시간을 더 소모할 수 있다.

자료 구조를 활용한 기법은 보통 앞의 기법들보다 학습 과정의 복잡도는 크지만 테스트 과정은 얻어진 모형 하나와의 비교가 전부여서 오히려 더 빠르다.

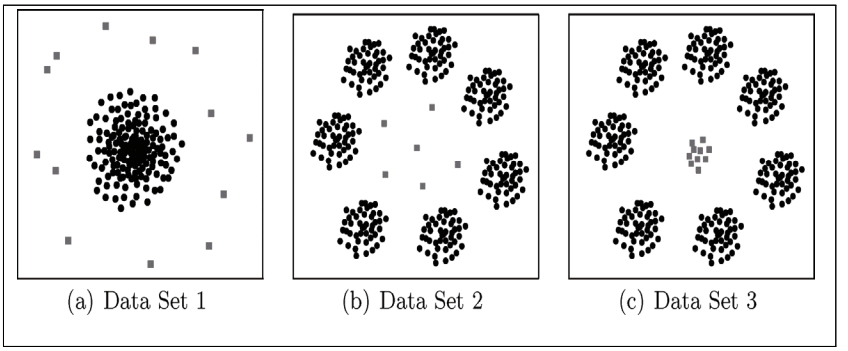
(참고 2) 맥락적 이상 탐지의 장단점

개체가 주변의 흐름을 따르는 경향이 있는 실제 자료에서 자연스럽게 정의되는 이상값의 개념을 유지할 수 있고, 점 이상 탐지 기법으로 찾기 힘든 이상값을 발견할 수 있다는 장점이 있다. 단점은 맥락이 반드시 정의되어야 한다는 것이다.

8. 이상 탐지 기법들의 상대적 장단점

현재까지 논의한 수많은 기법은 각각 특수한 장단점을 가지고 있다. 주어진 문제에 어떤 기법이 가장 적합한지 아는 것은 중요하지만, 문제 공간이 아주 복잡함을 고려하면 분석 기법들을 일일이 살펴보기는 불가능하다. 따라서 몇몇 상황으로 각 분야의 장단점을 살펴보았다.

[그림 3-8] 세 종류의 2차원 자료 예시



주: 원은 정상값을, 네모는 이상값을 나타낸다.
자료: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15. 52page

간단한 예제로 세 종류의 2차원 연속형 자료 〔그림 3-8〕을 살펴보자. 먼저 〔그림 3-8(a)〕의 자료에서 정상값은 정규분포로부터 생성되었고 조밀한 군집을 이루는 한편 매우 적은 이상값이 정상값 분포의 평균에서 아주 먼 곳에 있다. 따라서 정상값들을 포함하는 전형적인 학습 집합을 구할 수 있다. 대부분 기법의 가정들이 모두 성립해, 이 자료에는 어느 기법이든 이상값을 잘 탐지할 것이다.

〔그림 3-8(b)〕은 평균들이 원을 이루고 분산이 매우 작은 여러 정규분포에서 생성된 값들이 정상 자료를 구성한다. 일집단 분류 기반 기법은 전체 자료를 포함하는 원형 경계를 설정해 이상값을 찾아내지 못할 수도 있다. 하지만 각 군집이 하나의 클래스가 되면 다집단 분류 기반 기법으로 각 군집을 둘러싸는 경계를 학습해 가운데의 이상값들을 감지할 것이다. 이상값이 다른 개체와 충분히 떨어져 있어 군집화나 NN 기반 기법도 쓸만하다. 하지만 〔그림 3-8(c)〕처럼 이상값들이 군집을 이루면 군집화, NN 기반 기법은 이것들을 정상값으로 처리해 안 좋은 성능을 보일 것이다.

고차원 자료로 넘어가면 더욱 다양한 난제에 부딪힌다. 군집화 및 NN 기반 기법은 자료의 차원이 커지면 악영향을 받는데, 이는 고차원에서 거리 측도가 정상, 이상값을 구분하기 힘들기 때문이다. 이 문제는 스펙트럴 기법으로 자료를 저차원 공간으로 사영시키면 해결되지만, 성능이 사영된 공간에서 정상, 이상값이 구분되는지에 크게 의존한다는 또 다른 문제가 있다.

이때 분류 기반 기법이 좋은 대안이 될 수 있다. 하지만 최대의 효율을 위해 필요한 정상, 이상 모두에 대한 라벨을 구하기 어렵고, 자료의 불균형(imbalance)도 분류에 걸림돌이다. 오히려 이런 측면에서는 정상값 라벨만 사용하는 준지도 NN, 군집화 기법이 더 효율적일 수도 있다. 통계적 기법은 비지도임에도 자료가 저차원이고 통계적 가정이 성립해야

효과적이다. 정보 이론 기법은 한두 개의 이상값도 잡아낼 정도로 민감한 척도를 찾는 것이 관건이다.

NN과 군집화 기반 기법들은 개체 사이 거리를 정의할 수 있어야 하며, 그 거리로 정상값과 이상값을 쉽게 구분 가능하다고 가정한다. 이러한 거리를 정하기 어려우면 분류 기반이나 통계적 기법이 더 나은 선택이다.

이상 탐지에서 계산 복잡도는 특히 실제 자료에 적용할 때 아주 중요한 요소이다. 분류, 군집화, 통계적 기법은 학습 과정에 많은 시간을 쓰지만, 테스트는 빠르게 진행한다. 테스트는 실시간으로 해야 하지만 모형 학습은 오프라인이어도 무방하므로 이러한 기법들이 실전에 유용하다. 반대로 NN, 정보 이론, 스펙트럴 기법은 학습 과정이 없음에도 테스트가 오래 걸려 활용 범위에 제한이 있다.

마지막으로, 다수의 기법이 이상값이 매우 드물다고 가정하는데 그렇지 않은 경우도 존재한다. 예를 들어 컴퓨터 네트워크의 웜(worm)을 다루는 경우에는 이상값에 해당하는 웜 트래픽이 정상값보다 더 빈번하게 나타난다. 이처럼 이상의 비율이 높으면 비지도 기법은 부적절하고 대신 지도나 준지도 기법을 적용할 수 있다(Sun et al., 2007; Soule et al., 2005).

제2절 딥러닝(Deep learning)을 활용한 이상 탐지 기법

본 절에서는 최근 딥러닝을 이용한 anomaly/outlier 탐지에 대해 두 개의 논문을 소개한다. 앞에서 논의하였다시피 자료의 차원이 커지면 커질수록 차원의 저주가 생겨 거리를 정의하기 힘들다. 또한 고차원 자료를 저차원으로 축소시킨 뒤, 이상 탐지를 진행하는 2단계 기법에서 축소시

키는 사상(mapping)은 후에 진행할 이상 탐지에 대한 정보 없이 진행되기 때문에, 이상 탐지에 필요한 변수 또는 속성을 제거할 수 있다. 따라서 최근에는 2단계 기법보다는 딥러닝을 이용하여 차원 축소와 이상 탐지 방법을 동시에 학습하는 방법론이 연구되었다.

1. Deep Embedded Clustering(DEC)[Xie et al., 2016]

탐색적 자료 분석과 자료 시각화를 위한 비지도 기계학습 기법인 군집분석은 다음과 같이 여러 관점에서 연구되었다: 군집을 정의하는 것은 무엇인지, 적절한 거리 측도는 무엇인지, 각 개체들을 군집으로 그룹화하는 효율적인 방법은 무엇인지, 군집분석의 타당성을 평가하는 방법 등 다양한 거리 함수와 embedding 방법들이 연구되어 왔고, 상대적으로 특성 공간에서 군집분석을 시행하는 비지도 학습에 초점이 맞춰진 연구는 다양하지 않다.

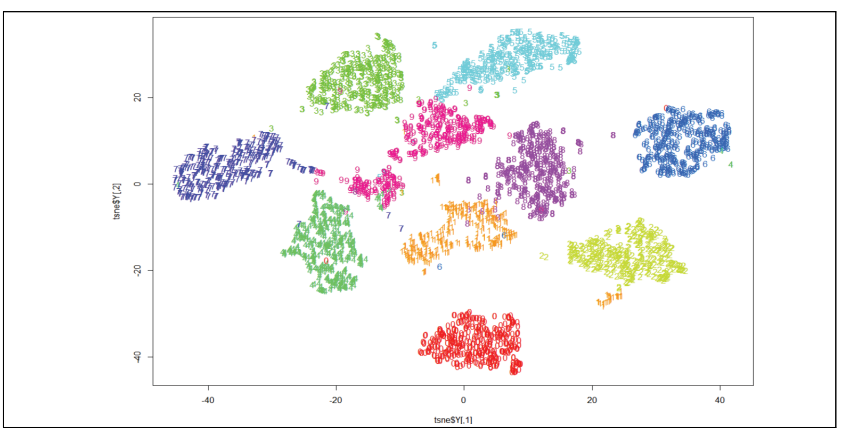
거리 또는 비유사도(dissimilarity)라는 개념은 이전 절에서도 언급하였다시피 자료를 군집화하기 위해 필수적인 항목이며, 거리는 특성 공간 내 나타나는 자료에 의존한다. 이러한 특성 공간의 선택은 일반적으로 분석자가 문제마다 결정해야 할 사항이다. 이 장에서는 자료를 기반으로 특성 공간 선택과 군집분석을 동시에 할 수 있는 비지도 학습 방법론을 제시하고자 한다.

비지도 군집분석을 위해, 원자료 공간에서 저차원의 특성 공간으로의 모수화된 비선형 사상을 정의하고, 저차원 특성 공간에 사영된 자료를 사용하여 군집분석을 시행하려 한다. 기존 연구들은 주로 원자료 공간에서 군집분석을 진행하거나 선형 차원 축소를 통해 저차원 공간에서 군집분석을 진행한 반면 해당 방법에서는 저차원으로의 비선형 사상을 deep neural network(DNN)로 정의하고 군집분석 목적함수에 대한 역전파

알고리즘을 통해 모형을 학습시키고자 한다. 이 군집분석 알고리즘을 Deep Embedded Clustering(DEC)이라고 한다.

DEC를 학습하기 위해 군집 할당과 저차원 특성 공간에 대해 동시에 최적화를 진행해야 한다. 각 개체마다 라벨이 없는 상태에서 학습을 진행하기 위해, 현 시점의 군집 할당 정보를 이용하여 보조의(auxiliary) 목표 분포를 사용하여 군집 할당 정보를 갱신한다. 이러한 과정은 군집분석에 대한 성능을 향상시키며 동시에 특성 공간에서 각 군집끼리 잘 분리되는 특성 표현(feature representation)을 찾아낼 수 있다.

[그림 3-9] t-SNE 예시



자료: Mark B. (2017). Multi-dimensional Reduction and Visualisation with t-SNE. https://data scienceplus.com/multi-dimensional-reduction-and-visualisation-with-t-sne/ 2018. 11. 30. 인출

데이터 시각화 및 차원 축소 방법인 t-SNE(Maaten & Hinton, 2008)는 원자료상의 자료 분포와 사영된 공간에서의 자료 분포 간 쿨백-라이블러 발산(Kullback-Leibler divergence 또는 KL divergence)을 최소화하여 저차원의 사영된 자료를 계산한다. t-SNE는 비모수적 알고리즘

이기 때문에 embedding 부분을 DNN으로 모형화한 parametric t-SNE(Maaten, 2009)가 제안되었다.

DEC는 parametric t-SNE의 목적함수를 변형시켜 군집분석에 사용한다. DEC에서는 원자료 공간의 거리를 저차원에서도 보존하는 embedding을 찾아내는 대신, 중심 기반(centroid-based) 확률 분포를 정의하고, 군집 할당과 특성 표현을 향상시키기 위해 중심 기반 확률 분포와 보조 목표 분포 간 KL 발산을 최소화한다.

먼저 주어진 N개의 자료 $\{x_i \in X\}_{i=1}^{N}$을 K개의 군집으로 나누는 문제를 고려하자. 각 군집마다 중심(centroid)을 $\mu_j, j = 1,\ldots,K$라고 표기하고, 원자료 차원에서 저차원으로 사영시키는 비선형 함수를 $f_\theta : X \rightarrow Z$라 하자. 여기서 θ는 학습 가능한 모수를 의미하고, Z는 저차원의 특성 공간을 의미한다. Deep neural network(DNN)는 함수 근사에 대한 이론적 성질(Hornik, 1991)과 잘 알려진 특성 학습 능력(Bengio et al., 2013)을 가지고 있기 때문에, f_θ을 모형화하기 위해 DNN을 사용한다.

DEC 알고리즘은 특성 공간 Z에서 k개의 군집의 중심 $\{\mu_j \in Z\}_{j=1}^{K}$와 자료를 Z로 사영시키는 DNN의 모수 θ를 동시에 학습하여 자료를 군집화한다. DEC는 다음과 같이 두 단계로 이루어진다. (1) deep autoencoder(Vincent et al., 2010)를 이용한 모수 초기화와 (2) 보조 목표 분포의 계산과 KL 발산을 최소화하는 과정을 반복하여 모수를 최적화한다. 먼저 θ와 $\{\mu_j \in Z\}_{j=1}^{K}$의 초기값이 주어졌다는 가정하에 (2)단계인 모수 최적화와 군집분석을 상세히 살펴보도록 하자.

가. KL 발산을 사용한 군집화

비선형 함수 f_θ에 대한 초깃값과 군집의 중심 $\{\mu_j \in Z\}_{j=1}^{K}$에 대한 초깃값이 주어졌을 때, 비지도 알고리즘을 사용하여 군집분석의 성능을 향상시키는 방법론을 살펴본다. 먼저 사영된 자료들과 군집 중심들에 대해 soft assignment를 계산하고, 보조 목표 분포를 사용하여 군집 할당이 높은 신뢰도를 가지도록 f_θ와 $\{\mu_j \in Z\}_{j=1}^{K}$를 학습시킨다. 이와 같은 과정을 분석자가 정한 특정 수렴 기준을 만족할 때까지 반복한다.

1) Soft assignment

Maaten & Hinton(2008)과 동일하게 사영된 자료 z_i와 중심 μ_j 사이의 유사도를 계산하기 위해 Student's t-분포를 사용한다.

$$q_{ij} = \frac{\left(1 + \| z_i - \mu_j \|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1 + \| z_i - \mu_{j'} \|^2/\alpha\right)^{-\frac{\alpha+1}{2}}},$$

여기서 $z_i = f_\theta(x_i) \in Z$는 주어진 자료 $x_i \in X$를 저차원 특성 공간으로 사영시킨 값이고, α는 t-분포의 자유도이다. q_{ij}는 i번째 개체가 j번째 군집에 할당될 확률로 해석할 수 있다. 비지도 학습에서는 α에 대한 교차검증을 진행할 수 없기 때문에, $\alpha = 1$로 설정한다.

2) KL 발산 최소화

보조 목표 분포를 사용하여 군집 할당의 신뢰도를 높이는 학습을 이용

하여 군집을 반복적으로 개선하는 방법을 살펴보자. 이를 위해 다음과 같이 soft assignment q_i와 보조 목표 분포 p_i의 KL 발산을 목표함수로 설정한다.

$$L = KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

목표 분포 P의 선택이 DEC의 성능에 굉장히 큰 영향을 끼친다. 단순히 고려할 수 있는 P는 각 p_i들을 가장 가까운 중심에 대한 델타 분포로 가정하는 것이다. 하지만 여기서 q_i가 soft assignment이기 때문에, softer probability 목표 분포를 설정해야 한다. 여기서 목표 분포는 다음과 같은 성질을 만족하도록 한다: (1) 예측을 강화하고 (즉, 군집의 순도(purity)를 향상), (2) 높은 신뢰도로 할당된 자료에 더 많은 가중치를 두고, (3) 대형 군집이 잠재 특성 공간을 왜곡하지 못하도록 각 중심에 대한 손실을 표준화한다.

DEC에서는 p_i를 다음과 같이 계산한다.

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}},$$

여기서 $f_j = \sum_i q_{ij}$로 soft 군집 빈도를 나타낸다. 이러한 학습 방법은 self-training(Nigam & Ghani, 2000)의 한 형태로 볼 수 있다. Self-training과 마찬가지로 DEC는 라벨이 없는 자료와 초기 분류기를 설정한 뒤, 신뢰도를 높이도록 학습하기 위해 분류기를 사용하여 자료에 라벨을 지정한다. 또한 각 반복 시행마다 높은 신뢰도 예측에서 학습하여 초기 추정치를 향상시키고 낮은 신뢰도를 가지는 개체에 대한 성능을 향상시키는 것을 실험적으로 확인하였다.

3) 최적화

DEC는 Stochastic gradient descent(SGD)를 이용하여 군집의 중심 μ_j와 DNN의 모수 θ를 동시에 최적화한다. 특성 공간으로 사영된 자료 z_i와 각 군집의 중심 μ_j에 대한 손실함수 L에 대한 미분은 다음과 같다.

$$\frac{\partial L}{\partial z_i} = \frac{\alpha+1}{\alpha}\sum_j \left(1+\frac{\| z_i-\mu_j \|^2}{\alpha}\right)^{-1}(p_{ij}-q_{ij})(z_i-\mu_j),$$
$$\frac{\partial L}{\partial \mu_j} = -\frac{\alpha+1}{\alpha}\sum_i \left(1+\frac{\| z_i-\mu_j \|^2}{\alpha}\right)^{-1}(p_{ij}-q_{ij})(z_i-\mu_j).$$

미분 $\partial L/\partial z_i$은 DNN으로 전달되고 역전파 방법을 통해 DNN의 모수에 대한 미분 $\partial L/\partial \theta$을 계산한다. 자료의 tol%(standard reference tolerance)보다 적은 자료에서만 군집 할당이 변할 때 최적화가 수렴했다고 판단하고 최적화 과정을 중단한다.

나. 모수 초기화

현재까지 DNN의 모수 θ와 군집 중심 $\{\mu_j\}$의 초깃값이 주어진 상태에서 DEC가 어떻게 진행되는지에 대해 살펴보았다. 이제는 모수와 중심을 어떻게 초기화하는지에 대해 알아보고자 한다. 먼저 DEC는 stacked autoencoder(SAE)를 사용하여 모수를 초기화하였다. 최근 연구에 따르면 SAE를 이용한 초기화 방법은 유의미하고 잘 분리된 표현을 일관성 있게 만들어 내는 것으로 나타났다(Vincent et al., 2010; Hinton & Salakhutdinov, 2006; Le, 2013). 따라서 SAE로 학습한 비지도 표현은 DEC를 사용한 군집분석 학습을 더 용이하게 한다.

SAE 네트워크를 초기화하기 위해, 각 층마다 denoising autoencoder

를 사용한다(Vincent et al., 2010). Denoising autoencoder는 다음과 같이 정의된 두 층의 인공신경망이다:

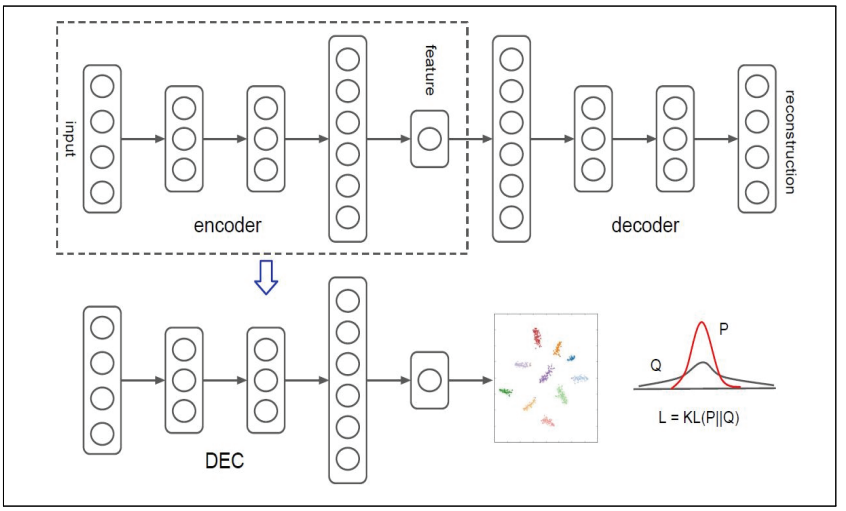
$$\begin{aligned}\tilde{x} &\sim Dropout(x)\\ h &= g_1\left(W_1\tilde{x}+b_1\right)\\ \tilde{h} &\sim Dropout(h)\\ y &= g_2\left(W_2\tilde{h}+b_2\right)\end{aligned}$$

여기서 $Dropout(\cdot)$(Srivastava et al., 2014)은 랜덤하게 선택된 입력값의 일부에 대해 0의 값을 부여하는 확률적 사상(mapping)이고, g_1, g_2는 각각 인코딩과 디코딩에 대한 활성함수(activation function)이고, $\theta = \{W_1, b_1, W_2, b_2\}$는 모형의 모수이다. 최소 제곱 손실함수 $\| x-y \|_2^2$을 이용하여 모형을 학습하고, 특정 하나의 층에 대해 학습이 끝난 뒤에는 그 층의 h를 입력값으로 설정하고 그 다음 층에 같은 과정을 반복한다. 여기서는 인코더/디코더에 대해 첫 번째 층에서의 g_2와 마지막 층에서의 g_1을 제외하고 활성함수를 rectified linear units (ReLUs) (Nair & Hinton, 2010)로 설정한다.

각 층마다 greedy 학습이 끝난 뒤, 모든 인코더 층을 연결시키고 그 다음 디코더 층을 학습 과정의 역순으로 연결시켜 deep autoencoder 네트워크를 형성하고 다시 복원 오차를 최소화하는 방향으로 학습시킨다(finetune). 최종적으로 〔그림 3-10〕과 같이 deep autoencoder에서 디코더 층을 버리고 인코더 층의 최종 출력값을 자료 공간에서 특성 공간으로의 초기 매핑으로 사용한다.

군집 중심의 초기화는 원자료를 초기화가 완료된 DNN에 입력하여 사영된 자료를 계산한 뒤, 특성 공간 Z에서 K-평균 군집분석을 시행하여 K개의 초기 중심값 $\{\mu_j\}_{j=1}^{K}$를 구한다.

〔그림 3-10〕 DEC의 네트워크 구조



자료: Xie, J., Girshick, R., & Farhadi, A. (2016, June). Unsupervised deep embedding for clustering analysis. In International conference on machine learning (pp. 478-487). 481page

다. DEC 기반 이상 탐지 방법

DEC를 이상치 탐색 문제에 적용하기 위해, 학습된 DEC를 바탕으로 개체 x_i에 대해 사영 공간의 사영된 값 z_i와 x_i가 포함되는 군집의 중심 μ_j를 구한 뒤, z_i와 μ_j의 거리를 이상 탐지의 기준으로 사용한다. 즉, i번째 자료가 포함된 군집의 중심에서 많이 떨어져 있을 때, 이상치일 확률이 높다고 판단한다.

2. Deep Autoencoding Gaussian Mixture Model(DAGMM) [Zong et al., 2018]

해당 장에서는 이상값 탐색의 핵심 아이디어*(주어진 입력 자료들에 대*

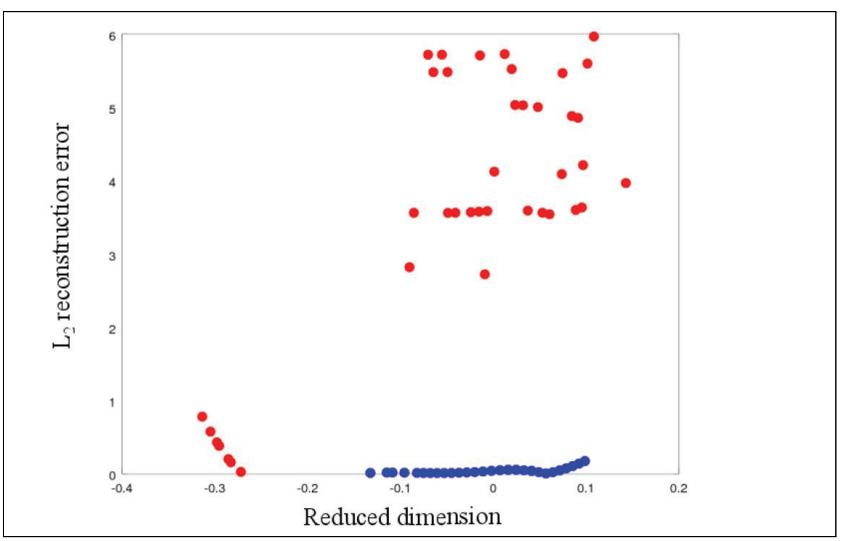
해서, 이상값들은 낮은 확률 밀도 영역에 위치한다)를 바탕으로 개발한 방법을 소개한다. 앞서 소개하였듯이 수많은 이상 탐지 연구가 있었지만, 사람의 보조 없이 다차원 또는 고차원 자료에 대해 로버스트한 이상값 탐지를 시행하는 일은 여전히 어려운 과제로 남아 있다. 특히, 입력 자료의 차원이 커지면 커질수록 원자료상에서 밀도를 추정하는 것이 점점 어려워지고, 모든 입력 자료에 대해 나타날 확률이 굉장히 낮아져 이상값으로 잘못 판단할 수도 있다(Chandola et al., 2009). 차원의 저주에서 생기는 문제를 해결하기 위해, 일반적으로 2단계 접근법을 많이 사용한다(Candes et al., 2011). 먼저 차원 축소를 시행한 뒤, 저차원의 잠재공간에서 밀도를 추정한다. 그러나 첫 번째 단계에서 진행하는 차원 축소는 후에 진행할 밀도 추정에 대한 어떤 정보도 인식하지 못하기 때문에, 이상 탐지를 위한 중요 정보가 차원 축소 단계에서 사라질 수도 있다. 그렇기 때문에 이러한 두 단계 접근 방법은 최적의 성능을 얻지 못한다. 따라서 계산적으로는 힘들지라도 차원 축소와 밀도 추정을 동시에 수행할 수 있도록 두 과정을 결합하는 것이 필요하다. 최근 연구들(Zhai et al., 2016; Yang et al., 2017; Xie et al., 2016)에서 DNNs의 강력한 모델링 기능을 활용하여 차원 축소와 밀도 추정을 동시에 진행하는 방법론들을 연구하였지만, 원자료에서의 중요 정보를 저차원의 축소된 공간에서 유지할 수 없었고, 너무 단순한 밀도 추정 모형을 사용하여 밀도함수를 정확히 추정하기가 불가능하다.

여기에서는 비지도 이상 탐지에서 앞서 언급한 문제를 해결하는 딥러닝 기반 모형 Deep Autoencoding Gaussian Mixture Model (DAGMM)을 소개한다.

먼저 DAGMM은 축소된 차원과 복원 오차에 대한 특성을 유지하여 입력값의 중요 정보를 저차원상에서도 보존한다. 〔그림 3-11〕과 같이 이상

치는 정상 자료와 두 가지 측면에서 다르게 행동하는 것을 볼 수 있다: (1) 이상값과 정상자료에서 변수들의 상관관계가 다른 방향으로 나타나 이상치의 경우 축소된 차원에서 크게 벗어나게 되고, (2) 이상값들이 정상자료와 비교하였을 때 복원하기 힘들다. DAGMM에서는 차원 축소를 위한 압축 네트워크(compression network)에 autoencoder를 사용하여 저차원상의 사영된 자료와 축소된 저차원상에서 원자료 공간으로의 복원 에러에 대한 특성 정보를 계산할 수 있다.

[그림 3-11] 개인 사이버 보안 자료에 대한 저차원상의 차원 축소 결과



주: (1) 각 개체의 원자료 차원은 20차원이고, 가로축은 deep autoencoder를 사용하여 1차원으로 차원을 축소한 결과이다; (2) 세로축은 각 개체마다 1차원에서 복원했을 때 원자료와의 차이이다; (3) 각 빨간색/파란색은 각각 비정상/정상 개체임을 나타낸다.

자료: Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. 2 page

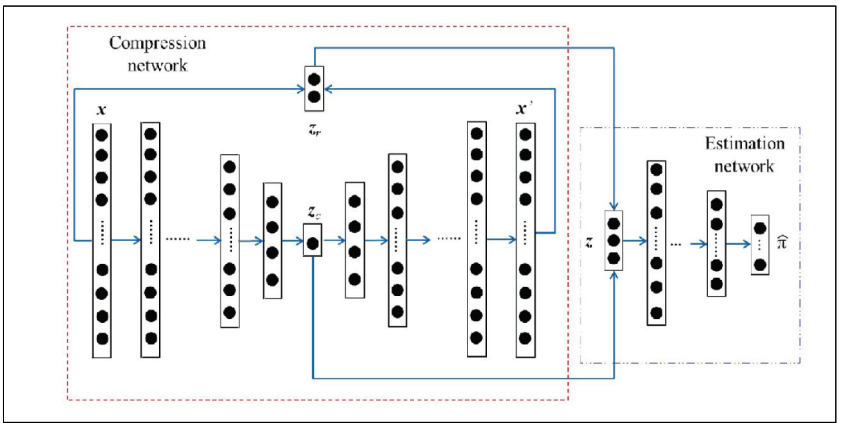
두 번째로, DAGMM은 학습된 저차원 공간에서 가우시안 혼합 모형(Gaussian Mixture Model, GMM)을 활용하여 복잡한 구조를 가진 입

력 자료에 대한 밀도 함수 추정을 수행한다. GMM은 밀도 추정에 대해 강력한 성능을 가지고 있고, 일반적으로 GMM은 Expectation-Maximization (EM)(Huber, 2011)을 사용하여 학습시킨다. 그러나 차원 축소와 GMM을 이용한 밀도 함수 추정을 동시에 최적화하기 힘들기 때문에, 종종 앞서 언급한 2단계 접근법을 사용하여 적합시킨다. 이 학습 문제를 해결하기 위해, DAGMM은 각 개체에 대해 압축 네트워크를 적용하여 저차원 입력을 계산한 뒤, 혼합 밀도 함수를 추정하는 추정 네트워크(estimation network)를 사용하고, 입력 자료를 저차원으로 축소시킨 뒤 에너지/가능도 평가를 가능하게 하여 GMM의 모수를 직접 추정할 수 있다.

마지막으로 DAGMM은 end-to-end 학습에 적합하다. 일반적으로 end-to-end 학습으로 deep autoencoder를 학습시키면 성능이 좋지 않은 local optima에 쉽게 빠질 수 있기 때문에, pre-training을 많이 사용한다(Vincent et al., 2010; Yang et al., 2017; Xie et al., 2016). 그러나 pre-training을 사용하면 후에 차원 감소 및 밀도 추정을 위한 fine-tuning 최적화를 진행할 때 네트워크에 대한 모수가 크게 변하지 않는다는 단점이 있다. 실험에 의하면, 추정 네트워크에 의해 생기는 정규화(regularization)로 압축 네트워크의 autoencoder가 local optima로 빠지지 않는 것을 확인하였다.

Deep Autoencoding Gaussian Mixture Model(DAGMM)은 두 개의 주요 요소인 압축 네트워크와 추정 네트워크로 구성되어 있다. 〔그림 3-12〕와 같이, DAGMM은 다음과 같이 진행된다: (1) 압축 네트워크는 deep autoencoder를 사용하여 입력 자료의 차원을 축소하고, (2) 추정 네트워크는 차원이 축소된 자료를 입력값으로 하여, GMM의 가능도/에너지를 예측한다.

〔그림 3-12〕 DAGMM의 구조



자료: Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. 4 page

가. 압축 네트워크

압축 네트워크는 deep autoencoder를 통한 축소된 저차원의 특성값 z_c와 복원된 자료에 대한 특성값 z_r을 생성한다. 즉, 주어진 입력 자료 x에 대해 압축 네트워크는 저차원 특성 벡터 z를 아래와 같이 계산한다.

$$z_c = h(x;\theta_e);\ x' = g(z_c;\theta_d)$$
$$z_r = f(x, x');$$
$$z = [z_c, z_r]$$

여기서 z_c는 deep autoencoder에서 계산한 저차원 특성값이고, z_r은 복원 오차에서 파생된 특성 변수들을 나타낸다. 또한 θ_e와 θ_d는 각각 deep autoencoder의 인코더, 디코더의 모수들이고 x'는 x를 압축시킨 후 복원된 자료, $h(\cdot)$은 인코딩 함수, $g(\cdot)$은 디코딩 함수, $f(\cdot)$은 복원 오차 특성을 계산하는 함수이다. 특히 z_r은 다차원으로 설정할 수 있

다. 예를 들어 유클리디안 거리, 상대 유클리디안 거리, 코사인 유사도 등으로 복원 에러에 대한 특성을 여러 차원으로 정의할 수 있다. 마지막으로 압축 네트워크는 그 다음의 추정 네트워크의 입력값 z를 계산한다.

나. 추정 네트워크

입력 자료의 저차원 특성 벡터를 바탕으로 추정 네트워크는 총군집 수가 K인 가우시안 혼합 모형(GMM)을 사용하여 밀도함수를 추정한다. 학습하는 과정에서는 가우시안 혼합 분포의 각 군집의 비율 ϕ_k, 평균 벡터 μ_k, 공분산 행렬 Σ_k, $k = 1, ..., K$를 알지 못하기 때문에, 추정 네트워크는 GMM의 모수를 추정하고 가능도/에너지 함수를 평가한다. 이를 위해, 추정 네트워크는 다층 인공신경망 네트워크(multi-layer neural network, MLN)를 이용하여 각 표본의 혼합 분포를 다음과 같이 예측한다. 주어진 저차원 특성 변수 z와 군집의 수 K에 대해,

$$p = MLN(z;\theta_m), \hat{\gamma} = softmax(p),$$

여기서 $\hat{\gamma}$은 K차원 벡터이며, 각 성분은 개체가 각 군집에 들어갈 확률에 대한 추정값을 나타내고, p는 모수가 θ_m인 다층 인공신경망 네트워크의 출력값을 의미한다. 총 N개의 개체가 주어져 있을 때, GMM의 모수는 아래와 같이 추정할 수 있다

$$\widehat{\phi_k} = \sum_{i=1}^{N} \frac{\widehat{\gamma_{ik}}}{N}, \widehat{\mu_k} = \frac{\sum_{i=1}^{N} \widehat{\gamma_{ik}} z_i}{\sum_{i=1}^{N} \widehat{\gamma_{ik}}}, \widehat{\Sigma_k} = \frac{\sum_{i=1}^{N} \widehat{\gamma_{ik}} (z_i - \widehat{\mu_k})(z_i - \widehat{\mu_k})^\top}{\sum_{i=1}^{N} \widehat{\gamma_{ik}}}$$

여기서 $\widehat{\gamma_j}$은 저차원 특성 벡터 z_i의 각 군집에 대한 비율이고, $\widehat{\phi_k}, \widehat{\mu_k}$, $\widehat{\Sigma_k}$은 각각 GMM의 k번째 군집의 비율, 평균 벡터, 공분산 행렬이다.

추정된 모수를 사용하여 샘플 에너지는 다음과 같이 계산한다.

$$E(z) = -\log\left(\sum_{k=1}^{K}\widehat{\phi}_k\frac{\exp\left(-\frac{1}{2}(z-\widehat{\mu_k})^\top\widehat{\Sigma_k}(z-\widehat{\mu_k})\right)}{\sqrt{\left|2\pi\widehat{\Sigma_k}\right|}}\right)$$

여기서 $|\cdot|$는 행렬식(determinant)을 의미한다.

예측 단계에서는 학습된 GMM의 모수를 사용하여 예측 자료의 샘플 에너지를 계산하고, 사전에 분석자가 선택한 기준값보다 큰 에너지를 갖는 샘플을 비정상 자료로 판단한다.

다. 목적 함수

주어진 N개의 샘플을 가지는 자료를 사용하여, DAGMM 학습을 위한 목적 함수는 다음과 같이 주어진다.

$$J(\theta_e, \theta_d, \theta_m) = \frac{1}{N}\sum_{i=1}^{N}L(x_i, x_i') + \frac{\lambda_1}{N}\sum_{i=1}^{N}E(z_i) + \lambda_2 P(\widehat{\Sigma})$$

위 목적 함수는 세 개의 구성 요소로 이루어진다.

- $L(x_i, x_i')$은 압축 네트워크에서 deep autoencoder의 모수를 추정하기 위한 복원 오차를 나타낸다. 직관적으로 만약 복원 오차가 적은 압축 네트워크가 입력값을 저차원으로 압축시킬 때 입력 개체의 정보를 더 많이 보존한다고 할 수 있다. 그러므로 복원 오차를 최소화하는 작업이 필요하고, 본 연구에서는 이를 위해 L_2-norm ($L(x_i, x_i') = \| x_i - x_i' \|_2^2$)을 사용한다.
- $E(z_i)$는 입력 개체들의 가능도를 모형화한다. 샘플 에너지를 최소화함으로써, 입력 자료의 가능도를 최대화하는 압축 네트워크와 추정

네트워크의 최상의 조합을 찾고자 한다.

- DAGMM은 또한 GMM과 마찬가지로 특이점(singularity) 문제가 있다. 공분산 행렬의 대각 성분이 0이 되면 무의미한(trivial) 해를 가진다. 이런 문제점을 해결하기 위해, 목적 함수에 공분산 행렬의 대각 성분에 대한 벌점화 함수를 추가하였다:

  $$P(\hat{\Sigma}) = \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{1}{\hat{\Sigma}_{kjj}}$$

  여기서 d는 압축 네트워크로 구한 저차원 특성 벡터의 길이를 의미한다.
- λ_1과 λ_2는 DAGMM의 조율모수이다.

라. DAGMM의 학습

일반적으로 pre-training에 영향받는 deep autoencoder 기반 모형들(Yang et al., 2017; Xie et al., 2016)과는 다르게, DAGMM은 end-to-end 학습을 사용한다. pre-training으로 사전에 학습된 압축 네트워크의 모수들은 이미 차원 축소만 진행하는 deep autoencoder의 local optima(또는 global optima)로 추정되어 있기 때문에, 그 다음 진행하는 밀도 추정 함수의 성능을 향상시키는 deep autoencoder의 네트워크 모수로 추정하기 힘들고, 따라서 이상치 탐색 성능을 저하시킨다. 또한 실험을 통해 압축 네트워크와 추정 네트워크를 동시에 추정함으로써 서로 성능을 향상시키는 것을 확인하였다.

제 4 장

보건사회 분야 자료의 이상 탐지 기법에 대한 탐색적 분석

4 보건사회 분야 자료의 이상 탐지 기법에 대한 탐색적 분석

이 장에서는 보건사회 분야 자료를 활용하여 앞에서 서술한 이상 탐지 기법을 적용해 보고, 활용 가능성을 검토해 보는 탐색적 분석을 실시하였다. 우리나라는 현재 고령사회로 진입하였기 때문에 향후 노인들에 대한 지원 및 정책이 확대될 수밖에 없다. 고령화 및 노령인구의 증가는 치매 노인 증가, 노인 학대 문제 등 돌봄 수요와 밀접한 관련이 있다. 이러한 차원에서 정책대상인 노인에 초점을 두어 보건 분야에서는 치매 조기 진단을 위한 자료(ADNI의 FDG-PET), 복지 분야에서는 노인 학대 노출 특성 분석을 위한 자료(2017 노인실태조사)를 활용하여 분석하였다. 자세한 설명은 각 절에서 기술하였다.

제1절 치매 조기 진단을 위한 이미지 자료(FDG-PET) 활용성 분석

보건 분야의 치매 조기 진단을 위한 이미지 자료 활용성 분석을 위해 사용한 데이터 및 anomaly 개념, 분석 프로세스, 분석 결과의 의미를 하나의 표로 정리하였다. 보건 분야 분석 개념도는 다음과 같다.

〈표 4-1〉 보건 분야 분석 개념도

내용	세부 내용	설명
Data	사용 데이터	ADNI의 FDG-PET 영상 자료(이미지 자료로 변환된 자료) + scalar 자료
개념	anomaly 개념	NCI군이 알츠하이머병으로 전환된 환자
	normal 개념	NCI군이 알츠하이머병으로 전환되지 않은 환자
프로세스	이상 탐지 기법 적용을 위한 자료 속성 파악	입력자료 성질: 수치형 자료 + 이미지 자료 이상의 종류: point anomaly 자료 라벨: 지도 이상 탐지 모형의 출력값 : 정상/이상 라벨 부여
	분석 방법 (사용한 이상 탐지 기법)	1. 스펙트럴 이상 탐지 기법: PCA 2. 분류 기반 이상 탐지 기법: Lasso
	분석 결과 제시	Accuracy Sensitivity Specificity AUC
의미	활용성	수치형 자료만 사용하는 것보다 이미지 자료와 함께 분석 시, 치매 조기 진단 예측력을 높일 수 있음. 이는 이상 탐지 기법에 정형 데이터뿐만 아니라, 비정형 데이터 분석도 중요함을 시사

1. 데이터 및 분석 개요

알츠하이머병(Alzheimer's Disease)은 치매를 일으키는 퇴행성 뇌질환의 일종으로, 1907년 독일 출신의 정신과 의사인 알로이스 알츠하이머(Alois Alzheimer)가 이 병에 대해 최초로 보고하였다. 알츠하이머는 서서히 발병하여 점진적으로 그 증세가 악화되는 것이 특징이며, 가장 흔히 나타나는 초기 증세로는 최근 일에 대한 기억력 상실이 있다. 병이 점차 진행되면서 언어 능력 및 인지 기능 상실이 나타나고, 말기에 이르면 신경학적 장애뿐 아니라 다양한 신체적인 합병증이 동반될 수 있다(Wikipedia, 2018).

인간의 평균수명이 길어지면서 치매 환자의 수도 증가하는 추세인데, 우리나라의 경우 노인성 치매 환자가 매년 약 4만 명 이상 증가하고 있다(한송원, 2018). 한편, 미국의 경우 2060년 알츠하이머성 치매를 앓는 환자가 약 1,400만 명에 이를 것이라 추정되며, 이를 인구당 비로 계산하면 3.3%이다. 참고로 2014년 전체 미국 인구 중 알츠하이머 치매를 앓는 비율은 1.6%이다(U.S. CDC, 2018).

알츠하이머성 치매 환자 증가에 따른 의료비 증가 문제를 해결하기 위하여 미국은 2000년대 초반부터 다양한 노력을 시도해 왔는데, 그중의 하나가 알츠하이머병뇌영상자료계획(Alzheimer's Disease Neuroimaging Initiative, ADNI)이다.

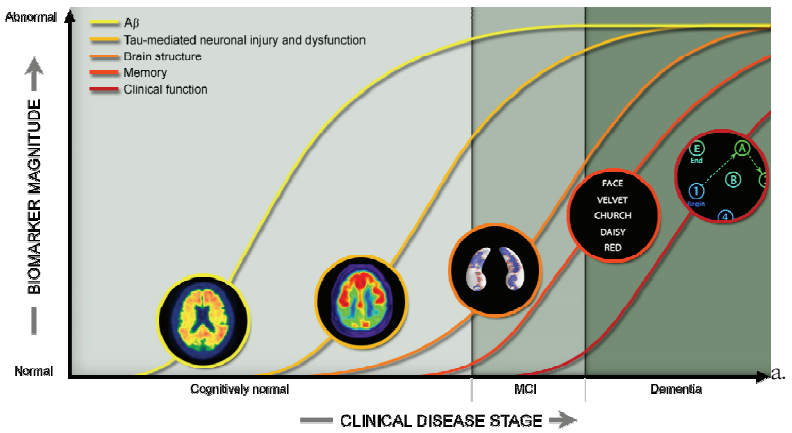
ADNI는 복합지역연구(multisite study)로서 알츠하이머병의 예방과 치료를 위한 임상시험(clinical trial)을 개선하기 위한 목적에서 만들어졌다. ADNI는 알츠하이머병을 초기에 진단하는 것뿐 아니라, 병의 생체표지자(biomarkers)를 추적(track), 입증(validation), 표준화(standardization)하는 데 주된 목적이 있다.

ADNI 연구의 참가자는 미국 또는 캐나다에 거주하는 55~90세 남성 또는 여성으로, 알츠하이머병으로 인한 치매 환자, 경미한 기억력 장애를 가지고 있는 환자(Mild Cognition Impairment, MCI),[4] 정상 표준(Normal Controls)으로 구성되어 있다. 자발적 참여를 통해 각 참가자에 대해 2004년부터 다양한 종류의 의료 데이터, 예를 들어 각종 영상 자료, 유전 정보, 신경심리검사 등을 장기적으로 수집해 왔는데, 이러한 데이터는 누구나 접근이 가능하도록 하는 'open data access'를 원칙으로 하고 있다. ADNI의 역사, 연구 참여 등록, 연구 디자인(study design), 알츠하이머 생체표지자(biomarker) 등에 대해 더 자세한 정보를 알고자

4) 알츠하이머로 발전 가능성이 높은 환자군.

한다면 http://adni.loni.usc.edu/를 참고하기 바란다. Petersen 등 (2010) 또한 ADNI의 목적 및 ADNI 자료의 탐색적 분석 결과를 자세히 기술하였다.

[그림 4-1] 알츠하이머 발병의 생체표지자 변화 그래프



a. β-amyloid (Aβ) measured in cerebrospinal fluid or by amyloid PET imaging

b. Neurodegeneration indicated by tau protein measured in cerebrospinal fluid, or by synaptic dysfunction, measured by FDG-PET

c. Brain atrophy, mostly in the medial temporal lobe, measured by structural MRI

d. Memory loss, measured by cognitive tests. e.g.)minimental score

e. Clinical function, indicated by general cognitive decline measured by cognitive tests.

자료: Alzheimers Disease Neroimaging Initiative. (2017). About Biomarkers. Retrive from http://adni.loni.usc.edu/study-design/#background-container 2018. 9. 2.

앞서 언급하였듯이 ADNI의 주된 목적은 알츠하이머병이 진행됨에 따라 다양한 영상 자료 및 임상적 실험 결과가 시간적으로 변화해 가는 과정을 추적함으로써 알츠하이머의 발병 및 진행 과정을 병리학적으로 이해하는 데 있다. 〔그림 4-1〕은 ADNI 자료를 바탕으로 시행한 단면 또는 종단 연구 결과들을 종합하여 만들어진 알츠하이머 생체표지자들의 변화 그래프를 보여 준다.

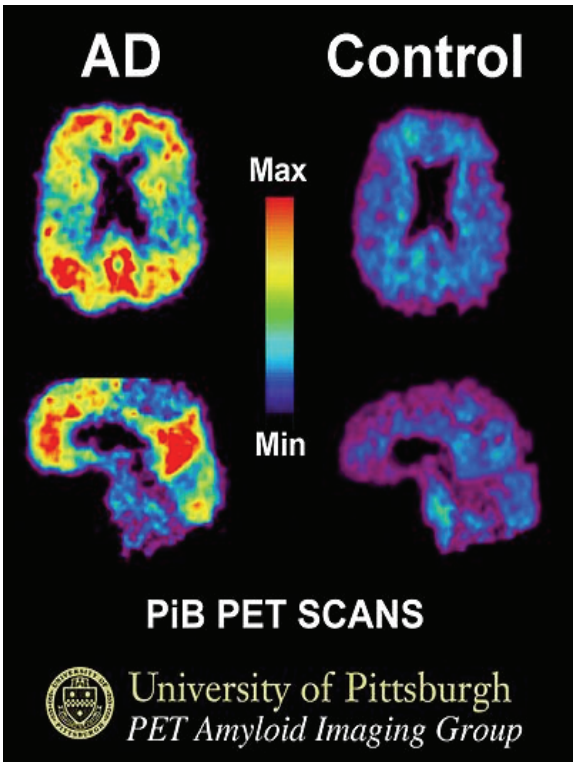
양전자단층촬영(Positron Emission Tomography, PET)은 양전자를 방출하는 방사성 물질을 이용하여 인체의 생리적, 화학적, 기능적 현상을 3차원으로 나타낼 수 있는 의료영상방법이다. PET에 사용되는 방사성 의약품의 종류는 다양한데, 본 연구에서 이용한 영상 자료의 방사성 물질은 FDG(Fludeoxyglucose)이다. FDG는 글루코스(glucose)의 유사물질로서 PET 영상 이미지에 가장 널리 이용되는 방사성 물질이다. FDG-PET의 농축량은 세포조직들의 지역적 글루코스 소모량을 나타내고, 이는 달리 말해 세포들의 신진대사 활동량을 뜻한다. 따라서 뇌의 FDG-PET 영상 자료는 뇌세포들의 신진대사량의 공간적 패턴을 시각화하여 보여 주는데, 알츠하이머 환자의 경우 뇌의 특정 부위에서 포도당 소모가 정상 표준에 비하여 눈에 띄게 적은 것으로 알려져 있다(〔그림 4-2〕 참조). 실제로 FDG-PET 영상 자료는 알츠하이머성 치매를 조기 진단하는 데 널리 이용되고 있다.

〔그림 4-1〕에서 보여지듯 병이 진행됨에 따라 FDG-PET 이외의 다양한 종류의 생체표지자에서 변화가 일어나는데, 〔그림 4-2〕는 단백질의 일종인 β-amyloid가 병이 악화됨에 따라 축적되어 가는 것을 시각적으로 보여 주고,[5] 〔그림 4-3〕은 병의 발병으로 인해 뇌의 구조가 변해 가는 것을 보여 준다.[6]

---

5) 〔그림 4-1〕a. 생체표지자에 해당.

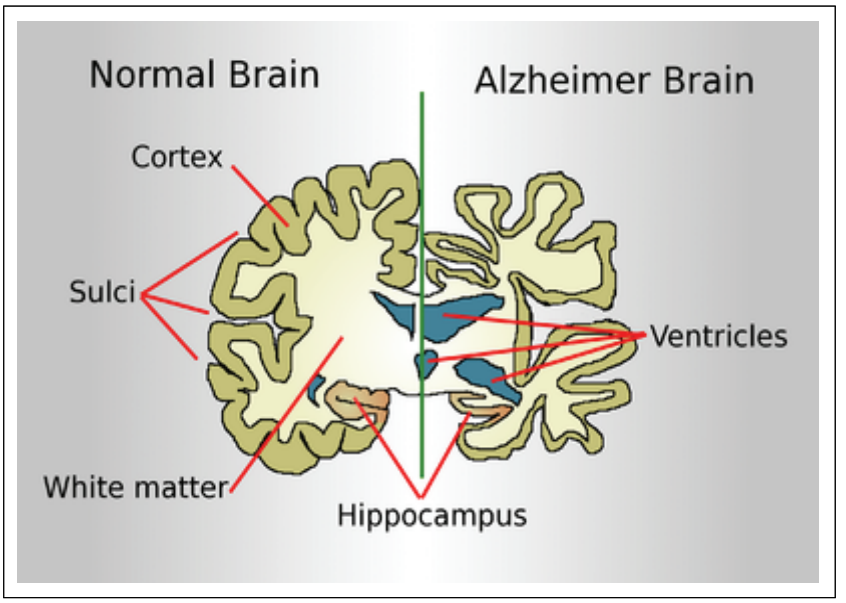
〔그림 4-2〕 β-amyloid 축적



자료: Ravi, S., &Rif, R. (2013). PET scans accurately identify amyloid deposition after traumatic brain injury. Retrieved from https://www.2minutemedicine.com/pet-scans-accurately-identify-amyloid-deposition-after-traumatic-brain-injury/ 2018. 9. 20. 인출.

6) 알츠하이머병이 심화됨에 따라 뇌가 쪼그라듦.

[그림 4-3] 알츠하이머 발병으로 인한 뇌의 구조적 변화



자료: Wikipedia. (2018). Alzheimerss Disease neuroimaging unitiative. Retrive from https://en.wikipedia.org/wiki/Alzheimer%27s_Disease_Neuroimaging_Initiative 2018. 9. 20.

알츠하이머가 발병하기 전 사람들은 일상생활에는 문제가 없지만 경미한 정도의 기억력 상실을 경험한다. 이러한 과도기적인 인지 기능 상실의 상태를 설명하기 위하여 경미한 기억력 장애를 가지고 있는 환자(Mild Cognition Impairment, MCI)군이 소개되었다. 많은 연구 결과에 따르면 MCI를 진단받은 환자군 중 연간 10~15%의 비율로 치매 환자로 전환한다고 한다. 이러한 현상으로 인해 MCI 환자들은 특히 연구 대상으로서 집중적인 관심을 받아 왔다.

MCI 환자들 중 알츠하이머병으로 전환 가능성이 높은 환자들을 미리 예측할 수 있다면 이들을 대상으로 빠른 치료를 실시할 수 있고, 결국 알츠하이머의 발병을 예방할 수 있을 것이다. 따라서 본 절에서는 ADNI의

FDG-PET 영상 자료를 바탕으로 MCI를 경험하는 환자들 중 3년 이내에 치매 환자로 전환하는 환자와 그렇지 않은 환자를 예측하여 분류하는 분석을 시범적으로 시행하였다.

본 분석에서 활용된 자료는 다음과 같이 정리할 수 있다. 먼저 3년 이내에 알츠하이머병으로 전환 여부를 예측하는 진단의 도구로 이용한 각 개체의 FDG-PET은 3차원의 이미지이다. 이들은 이미지 전처리 과정을 거쳐 126*126*96 사이즈의 수리 데이터로 전환되었다. 분석이 바로 가능한 이러한 데이터[7]는 ADNI의 데이터베이스인 http://adni.loni.usc.edu/data-samples/에서 다운로드하였다. 3차원 이미지의 최소 단위를 복셀(voxel)[8]이라 한다. 따라서 각 개체의 FDG-PET 이미지를 구성하는 복셀의 개수는 총 152만 4,096(=126*126*96)개이다.

인간의 뇌는 복잡한 네트워크의 구조를 가지고 있다. 즉 인간의 뇌는 공간적 분포를 가지지만 기능적으로 유기적 연대성을 갖는 다양한 크기의 해부학적 영역으로 구성되어 있다. 뇌 영상 자료를 보면 이웃하는 영역들의 색깔이 비슷한 것을 알 수 있는데, 이는 이웃하는 영역 또는 복셀들은 비슷한 값을 가짐을 뜻하고, 이러한 현상을 통계학적 용어로 '높은 공간적 상관관계(spatial correlation)'라 한다.

공간 데이터 분석에서 높은 공간적 상관관계를 해결하는 문제는 중요하게 다루어져 왔고 많은 발전을 거듭해 왔다. 그러나 통계학적 배경이 약한 대부분의 의학 분야 전문가들은 기술적 어려움으로 인해 뇌 영상 자료를 해부학적 구조를 이용하여 사전에 분할(segmentation)하여, 각 세그먼트의 대푯값[9]을 이용하여 변수 간의 높은 상관관계의 문제를 해결하고자 시도하였다. 이 방법은 결과의 해석을 용이하게 한다는 장점이 있는

7) preprocessed data
8) 3D analogue of pixel
9) 평균

반면, 영상 자료의 공간적 특성 및 정보를 제대로 활용할 수 없다는 단점이 있다. 나아가 측정의 오류로 인해 발생하는 추정의 편향(estimation bias) 문제들을 제대로 다루지 못한다는 한계도 있다. 따라서 본 분석에서는 분석의 단위를 세그먼트가 아닌 복셀로 설정하여 공간정보를 충분히 활용하고자 한다.

복셀을 분석의 단위로 설정하면 추정 계수의 인식 가능성(identifiability) 문제에 직면한다. 왜냐하면 개체의 개수 n보다 변수의 개수 p가 월등히 크기 때문이다. 여기서 변수의 개수는 이미지당 복셀의 수와 같고 통계학에서는 이러한 현상을 $n \ll p$ 문제라 한다. 통계적 인식 가능성의 문제를 해결하기 위해서 본 분석에서는 기저함수(basis function) 및 경험직교함수(empirical orthogonal function)를 이용하여 차원 축소(dimension reduction)를 시도하였는데 자세한 언급은 아래 분석 결과에서 하겠다.

앞서 알츠하이머 환자의 경우 뉴런의 파괴로 인해 뇌의 특정 부위에서 포도당 소모가 정상 표준에 비하여 눈에 띄게 줄어든다고 언급하였다. 여기서 말하는 뇌의 특정 부위는 히포캠퍼스(Hippocampus), 메디알 템보랄 롭(Medial Temporal Lobe) 등이 있는데, 특히 뒤의 대상 피질(Posterial Cingulate Cortex, PCC) 부위에서 포도당의 소모 감소가 눈에 띄게 나타나는 것으로 알려져 있다. 〔그림 4-4〕는 알츠하이머병이 진행됨에 따라 PCC 영역의 뇌세포들의 포도당 소모가 눈에 띄게 줄어드는 것을 시각적으로 보여 준다.

PCC는 뇌의 아래 뒤쪽에 위치하는 피질이다. 따라서 본 분석에서는 3차원의 뇌 영상 자료 중 PCC가 위치한 뇌의 아래 부위에 집중하고자 한다. 따라서 3D 자료 중 뇌의 아래쪽에 해당하는 뇌 축의 슬라이스(axial slice) 번호 60번만 이용하였고,[10] 따라서 분석에 포함된 총복셀의 개수

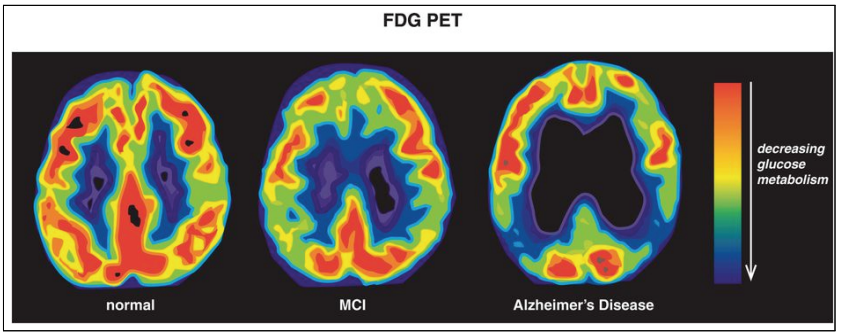
10) 뇌의 위쪽으로 갈수록 슬라이스 번호 줄어듦.

는 (15,876=126*126)이다.

반응변수는 알츠하이머로의 전환 유무를 나타내는 지표로서 전환(1) 또는 비전환(0)으로 표기되었다. 이때 전환 여부는 FDG-PET 영상을 찍은 시점에서 3년이 지나 평가되었다. FDG-PET 영상 자료 이외에 알츠하이머 발병에 영향을 줄 수 있는 성별, 나이, 유전자형(genotype), 간이정신상태평가(minimental state examination, MMSE) 점수를 모형에 포함하여 조절변수로 이용하였다.

간이정신상태평가 점수는 가장 널리 사용되고 있는 치매를 선별하는 검사 도구로서 인지 기능의 손상 정도를 측정하는 것을 목적으로 한다. 이 점수는 1975년에 Folestein 부부와 McHugh에 의해 처음으로 소개되었다(Folestein, Folestein, & McHugh, 1975). 이 평가 점수는 30점이 만점이고 일반적으로 24점 이상이면 정상, 20~23점이면 치매 의심, 19점 이하인 경우 치매 판정으로 여겨진다. 즉 간이정신상태평가 점수가 낮을수록 치매 진단의 확률이 높아진다고 이해할 수 있다.

[그림 4-4] 포도당 소비 패턴 영상 자료



자료: Basic medical key. (2018). Retrive from https://basicmedicalkey.com/dementia-and-its-treatment/ 2018. 9. 2.

2. 이상(anomaly) 자료 정의 및 기초 통계

본 분석에서는 이상(anomaly)의 개념을 일정 기간 동안 알츠하이머 환자 전환의 유무로 정의하였다. 즉 MCI에서 3년 이내에 알츠하이머성 치매로 전환한 환자를 비정상, 반대로 전환하지 않은 환자는 정상이라 정의하였다. 왜냐하면 알츠하이머가 발병하면 뇌 신경세포들이 손상되기 때문에 이러한 환자들의 뇌는 정상에 대해 비정상이기 때문이다(〔그림 4-2〕, 〔그림4-3〕, 〔그림4-4〕 참조).

ADNI 연구에 포함된 총참가자는 정상표준이 229명, MCI 그룹이 398명, 알츠하이머 환자가 192명이다. 각 그룹에 대한 기초 통계량 분석 결과는 [표 4-2]에 제시한다. 분류 분석 시 MCI 그룹만 이용하였지만 비교를 위하여 정상표준과 알츠하이머 환자의 기초 통계량도 표에 제시하였다. 표의 결과를 보면 각 그룹별로 나이, 교육 연수, 성별 등은 큰 차이가 있어 보이지 않는다. 실제로 이들의 차이는 통계적으로 유의하지 않았다. 단, ApoE4 유전자형의 보유 여부는 알츠하이머의 발병과 밀접한 연관이 있어 보인다. 즉, 알츠하이머 환자 그룹에서 ApoE4를 보유한 환자의 비율은 정상표준 그룹에 대해 두 배 이상으로 월등히 높다. 실제로 그룹 간 차이를 통계적으로 가설검증한 결과 이들의 차이는 매우 유의한 것으로 드러났다.

〈표 4-2〉 ADNI 연구 참가자의 인구 통계학적 특성

characteristic	정상표준	MCI	알츠하이머
총명수	229	398	192
나이 평균	75.8	74.7	75.3
교육 연수	16	15.7	14.7
여성 비율(%)	48	35.4	47.4
ApoE4 보유 비율(%)[11]	26.6	53.3	66.1

3. 분석 결과

분석에 포함된 개체 수는 398명으로 이 중 120명의 MCI 환자가 3년 이내에 알츠하이머 환자로 전환하였고, 나머지 278명은 전환하지 않았다. 이를 연간 전환율로 환산하면 약 10% 정도 된다.

본 분석에서는 알츠하이머로 전환 유무를 분류할 때 이미지 자료가 포함된 모형과 포함되지 않은 모형의 분류 정확도(classification)를 비교하였다. 이러한 분석의 목적은 이미지 자료가 치매의 조기 진단에 대해 가지는 중요성 및 의의를 강조하기 위함이다.

뇌 영상 자료는 전형적으로 평평한(smooth) 영역들이 공간적으로 결합한 형태를 띠고 있다. 달리 말해 같은 해부학적 영역에 속하는 뇌 조직들은 영상 자료에서 비슷한 색으로 나타나고, 다른 해부학적 영역에 속하는 조직들은 다른 색을 띠기 때문에 영역 간 경계선은 다소 분명한 편이다.

뇌는 영역별 기능이 다르기 때문에 특정한 뇌 질환에 관한 진단 및 연구를 할 때 그 질병과 관련된 영역에 집중하는 것이 매우 중요하다.[12] 달리 말해 알츠하이머 전환 유무를 이미지를 바탕으로 판단하고자 하는 본 연구에서는 뇌의 영역 중 기억력과 인지 능력과 관련된 영역을 찾아 병이 진행됨에 따라 그 영역에서의 변화에 집중하는 것이 필요한데, 이러한 분석 목적을 달성하기 위해서는 희박주성분분석(sparse principal component analysis) 추정 방법론을 이용하여 분류를 시도하는 것이 적절해 보인다.

희박주성분분석을 활용하면 알츠하이머 전환 여부와 관련이 없는 뇌의 영역에서 회귀계수(coefficient)[13] 값은 0으로 줄어드는(shrink) 반면,

11) APOE 유전자형의 보유 여부는 알츠하이머병 발병과 밀접한 관련이 있다고 알려져 있음.
12) 통계학에서는 이를 희박성(sparsity)이라 부름.

전환 여부에 큰 영향을 미치는 영역에서의 절대계수 값 $|\beta_{jk}|$은 $|\beta_{jk}| \gg 0$을 만족하므로, 앞서 설명한 회귀계수 이미지(coefficient image)의 희박성을 만족한다. 일반적으로 희박 회귀계수는 해석이 용이하고 예측력을 높이는 장점이 있는 것으로 알려져 있다.

회귀계수의 성김성(sparsity)을 위해서는 다양한 방법들을 고려해 볼 수 있는데, 본 분석에서는 웨이블릿 변환(wavelet transformation) 및 l_1 벌칙(penalty)을 이용하였다. 즉 각 개체의 뇌 이미지 및 주성분들을 웨이블릿 기저함수로 표현하여 차원 축소를 시도한 후 이를 통해 얻어진 웨이블릿 계수들을 바탕으로 회귀모형을 재설정하였다. 이때 회귀계수들의 l_1 놈(norm)에 제약을 둠으로써, 즉 $|\beta_{jk}| < \lambda$를 제약함수로 설정함으로써, 대부분의 영역에서의 회귀계수 값이 0으로 수렴하도록 강요하였다. 여기서 λ는 희박성의 정도를 조절하는 매개변수로서, $\lambda \to \infty$이면 희박성이 점점 강해져 $\beta_{jk} = 0$, for all j,k를 만족하게 된다. λ는 AIC, BIC, CV 등의 방법으로 선택할 수 있다.

아래 〔표 4-3〕은 알츠하이머 전환 여부 분류 분석의 결과를 보여 준다. FDG-PET 영상자료가 알츠하이머성 치매의 조기 진단에 가지는 예측력을 평가하기 위하여 이미지가 포함된 모형과 포함되지 않은 모형을 각각 적합하였고, 각 모형의 분류 결과 다양한 각도에서 평가하였다. 분류의 정확도(accuracy), 민감도(sensitivity), 특이성(specificity)을 계산하는 데 사용한 식별한계점(discrimination threshold)은 0.5이다.

민감도[14]를 먼저 살펴보면 λ가 AIC에 의해 선택된 경우 수치적 변수들(scalar variable)[15]만 포함된 모형의 경우는 민감도가 0.667이고

13) β_{jk}: jk는 위치를 나타내는 index임.
14) true positive, 즉 실제 전환한 개체 중 전환으로 분류된 개체 중 비율.
15) 나이, 교육 연수, 성별, ApoE4 보유 여부, 간이정신상태평가점수.

FDG-PET 이미지도 포함된 모형의 민감도는 0.692이다. 이는 3년 뒤 알츠하이머로 전환한 사람이 1,000명이라 가정하였을 때, 수치적 변수들만 활용하면 667명을 미리 판별할 수 있고, 이미지도 함께 활용하면 692명을 조기 발견할 수 있다는 것을 의미한다.

정확도를 살펴보면 AIC, 수치적 변수 모형의 경우 0.852이고 수치적+이미지 모형의 경우 0.859로서 1,000명 중 맞게 분류[16]된 개체 수가 두 번째 모형에서 7만큼 높다.

〈표 4-3〉 알츠하이머 전환 여부 분류 분석 결과

모형	방법	classification result				error deviance
		Accuracy	Sensitivity	Specificity	AUC	
scalar + Image	AIC	**0.859**	**0.692**	0.928	**0.929**	**128.99**
	BIC	0.854	0.683	0.925	0.917	137.58
	CV	0.857	0.658	**0.938**	0.906	145.08
Image only	AIC	0.801	0.483	0.932	0.814	188.52
	BIC	0.813	0.508	0.938	0.808	190.94
	CV	0.782	0.433	0.925	0.783	201.88
scalar only	AIC	0.852	0.667	0.928	0.904	144.01
	BIC	0.842	0.658	0.918	0.901	146.71
	CV	0.837	0.658	0.911	0.888	156.14

주: ROC curve로부터 계산한 AUC(area under curve)임.

16) 전환-전환 분류, 비전환-비전환 분류 비율.

〈표 4-4〉 알츠하이머 전환 여부 예측 분류 분석 결과

모형	방법	classification result				error deviance
		Accuracy	Sensitivity	Specificity	AUC	
scalar + Image	AIC	0.837	**0.650**	0.914	**0.889**	**158.45**
	BIC	0.833	0.642	0.911	0.864	172.20
	CV	**0.842**	**0.650**	**0.921**	0.880	161.38
Image only	AIC	0.774	0.458	0.904	0.758	211.13
	BIC	0.767	0.442	0.901	0.771	207.66
	CV	0.757	0.333	0.932	0.752	212.50
scalar only	AIC	0.816	0.600	0.904	0.876	163.00
	BIC	0.818	0.608	0.904	0.865	170.78
	CV	0.820	0.625	0.901	0.867	168.95

위 [표 4-4]는 알츠하이머 전환 여부를 10-fold CV를 바탕으로 예측하여 분류한 분석 결과를 보여 준다. [표 4-2]와 달리 모델 적합과 예측에 다른 개체를 이용하여 예측(prediction)을 시도하였다. [표 4-3]은 단순 적합(fitting)의 결과로 예측을 시도하지는 않았다. 예상하는 바에 부합하게 모든 방법과 모형에 대해 예측의 경우 분류의 결과가 적합에 비해 나쁘다. 아래 [표 4-4]의 분류의 정확도(accuracy), 민감도(sensitivity), 특이도(specificity)를 계산하는 데 사용한 식별한계점(discrimination threshold)은 0.5이다.

민감도를 먼저 살펴보면 λ가 AIC에 의해 선택된 경우 수치적 변수들(scalar variable)[17]만 포함된 모형의 경우는 민감도가 0.6이고 FDG-PET 이미지도 포함된 모형의 민감도는 0.65이다. 이는 3년 뒤 알츠하이머로 전환한 사람이 1,000명이라 가정하였을 때, 수치적 변수들만

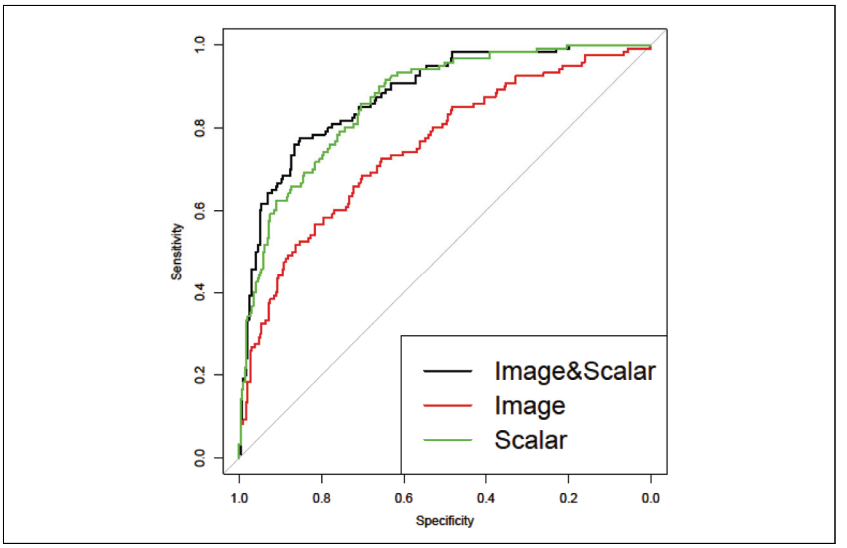
17) 나이, 교육 연수, 성별, ApoE4 보유 여부, 간이정신상태평가 점수

활용하면 600명을 미리 예측할 수 있고, 이미지도 함께 활용하면 650명을 조기 발견하여 알맞은 치료를 미리 시행할 수 있음을 의미한다.

정확도를 살펴보면 AIC, 수치적 변수 모형의 경우 0.816이고 수치적+이미지 모형의 경우 0.837로서 1,000명 중 맞게 분류[18]된 개체 수가 두 번째 모형에서 21만큼 높다.

〔그림4-5〕는 식별한계점을 0에서 1로 증가시킴에 따라 변화하는 예측 분류의 민감도와 특이성을 계산한 것을 시각화하여 보여 주는 ROC curve이다. 실선이 왼쪽 위 모서리에 가까울수록 예측이 잘되었음을 뜻하는데, 대부분의 식별한계점에서 Image+Scalar 모형이 더 나은 예측력을 보여 주는 것을 알 수 있다.

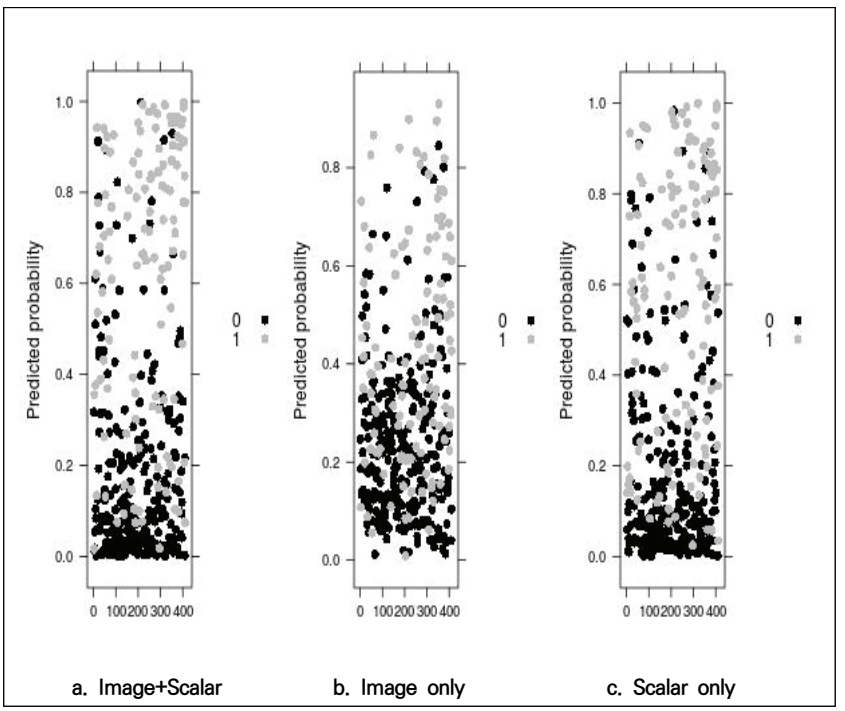
〔그림 4-5〕 알츠하이머 전환 여부 예측 분류 분석 ROC curve



18) 전환-전환 분류, 비전환-비전환 분류 비율

〔그림 4-6〕은 각 개체의 예측된 알츠하이머 전환 확률값을 보여 준다. 모든 모형에서 실제 전환한 개체(1로 코딩, 회색 점으로 표시)의 전환 확률이 전환하지 않은 개체(0으로 코딩, 검정색 점으로 표시)의 확률보다 큰 경향이 있고 따라서 모형이 전환 여부를 잘 예측하고 있다는 것을 알 수 있다. 단, Image+Scalar의 경우 두 그룹을 더 잘 분류하는 것으로 보인다. 왜냐하면 전환 개체의 전환 확률이 1에 더 가까운 경향이 있고, 반대로 전환하지 않은 개체의 경우 그 확률이 0에 더 가까운 경향이 있기 때문이다.

〔그림 4-6〕 알츠하이머 전환 확률 예측 결과



4. 시사점

치매 조기 진단을 위한 자료 분석에서는 NCI군이 알츠하이머병으로 전환된 환자를 anomaly로 정의하였고, normal의 개념은 NCI군이 알츠하이머병으로 전환되지 않은 환자로 정의하였다.

분석을 위해 3차원의 영상 자료를 2차원의 이미지 자료로 변환하였으며, 이미지 자료뿐만 아니라 수치형 자료도 함께 입력변수로 활용하였다. 이상 탐지 기법 적용을 위한 자료 속성은 point anomaly이고, 지도 이상 탐지에 속한다고 볼 수 있다. 분석 방법은 웨이블릿 변환(wavelet transformation) 및 l_1 penalty를 이용하였고, 정확도, 민감도, 특이도, AUC로 분석 결과를 제시하였다.

그 결과, 이미지 자료만 가지고 치매 가능성을 예측하는 것보다는 인구학적인 정보만 가지고 치매 가능성을 예측하는 것이 정확도 측면에서 좋고, 이미지 자료와 인구학적 정보를 둘 다 이용했을 때의 예측력은 더 높아진다는 것을 알 수 있다.

즉, 수치적 자료로 고려한 인구학적 요인들과 이미지 자료까지 고려하면 치매를 조기 진단하는 데 예측 가능성을 높인다는 점에서 중요한 시사점을 줄 수 있다. 이는 정형 자료와 비정형 자료가 연계되면 활용성 측면에서 더 긍정적인 효과가 발생할 수 있다는 것을 보여 준다.

하지만, 이에 대한 제한점도 존재한다. 어떠한 이상 탐지 기법을 적용하느냐에 따라서 모형 성능이 바뀔 수 있고, 영상 자료를 이미지 자료로 변환할 경우 어떤 feature를 선택하느냐에 따라 분석 결과가 달라질 수 있다. 이런 부분을 염두에 두고 관련 연구를 진행한다면 보다 발전된 연구 성과를 보일 수 있을 것이라 생각한다.

제2절 노인 학대 노출에 대한 이상(anomaly) 재정의와 특성 분석

복지 분야의 노인 학대 노출에 대한 탐색적 자료 분석을 위해 사용한 데이터 및 'anomaly' 개념, 분석 프로세스, 분석 결과의 의미를 하나의 표로 정리하였다. 복지 분야 분석 개념도는 다음과 같다.

〈표 4-5〉 복지 분야 분석 개념도

내용	세부 내용	설명
Data	사용 데이터	2017 노인실태조사
개념	'anomaly' 개념	- 정상(학대 경험 없음)인데 비정상(학대 경험 있음)으로 예측된 대상자(오분류) - 비정상(학대 경험 있음)인데 정상(학대 경험 없음)으로 예측된 대상자(오분류)
	'normal' 개념	- 정상(학대 경험 없음)인데 정상(학대 경험 없음)으로 예측된 대상자(정분류) - 비정상(학대 경험 있음)인데 비정상(학대 경험 있음)으로 예측된 대상자(정분류)
프로세스	이상 탐지 기법 적용을 위한 자료 속성 파악	- 입력 자료 성질: 연속형 자료 + 범주형 자료 - 이상의 종류: 맥락적 이상(contextural anomaly) - 자료 라벨: (맥락적 이상 개념을 적용한) 지도 이상 탐지 - 모형의 출력값: 정상/이상 라벨 부여
	분석 방법 (사용한 이상 탐지 기법)	1. 이상 탐지의 통계적 기법: 혼합 모수적 방법 2. NN(nearest neighbor) 기반 이상 탐지 기법: LOF 3. 스펙트럴 이상 탐지 기법: t-SNE 4. 군집화 방법(clustering): DBSCAN 알고리즘
	분석 결과 제시	1. 4개의 class로 나누어 이상치가 어디에 속하는지 특성 분석 2. 학대 경험 없는 집단에서 이상값 상위 100명 plot 및 특성 분석/학대 경험 있는 집단에서 이상값 상위 100명 plot 및 특성 분석 3. plot에 anomaly 표시 및 분석 4. plot에 clustering 결과 표시 및 분석
의미	활용성	- 분석 방법에 따라 대분류->중분류->소분류로 정의할 수 있을 것임 - 대분류를 한다면 1번 -> 4번 방법으로 톱 다운(top down) 설명 가능

내용	세부 내용	설명
		- 소분류를 한다면 4번 -〉 1번 방법으로 보텀 업(bottom up) 설명 가능 - 이는 특성에 대한 대분류 또는 상세 분류(세부 속성)를 하기 위해서는 단계별로 이상 탐지 기법을 적용하는 것이 바람직하다는 것을 시사

1. 데이터 소개[19)]

노인복지법 제5조 노인실태조사 실시의 법제화(2007년 1월)로 3년 마다 조사를 실시하도록 되어 있으며 그 일환으로 2017년도 노인실태조사가 실시되었다. 본 조사가 법제화된 2007년 이후 2008년, 2011년, 2014년에 이어 4번째로 실시되는 조사로, 노인에 대한 심층적 이해를 위한 경험적 기반을 마련함으로써 노인 정책 수립에 필요한 기초자료를 제공하는 것을 목표로 한다. 즉 본 조사를 통하여 노인의 생활 현황과 욕구를 다각적으로 파악하고 노인 특성의 변화 추이를 예측하여, 현재의 노인 정책 및 향후 다가올 고령사회에 대응하기 위해 정책 개발에 필요한 기초자료를 제공하는 것을 목적으로 하고 있다. 본 조사는 사전조사와 전문가 검토 등을 거쳐 통계변경승인(승인번호 제11771호)을 받아 확정된 조사표를 활용하여 2017년 6월 12일~8월 28일 기간 동안 934개 조사구의 65세 이상 1만 299명(대리응답 226명 포함)에 대한 직접면접조사를 완료하였다.

2017년 노인실태조사는 노인의 삶의 질 향상을 위한 정책 개발과 향후 고령사회에 대응하기 위한 정책 대응을 위해 매우 중요한 자료로 활용되는 조사이다. 한국 사회의 급격한 인구고령화와 더불어 노인의 양적 증대 및 노인 내부의 다양성이 증대하고 있어, 이를 반영한 맞춤형 정책 개발

19) [정경희, 오영희, 강은나, 김경래, 이윤경, 오미애, 황남희, 김세진, 이선희, 이석구, 홍송이 (2017). 2017년도 노인실태조사. 보건복지부·한국보건사회연구원.] 보고서 내용 발췌.

이 요구되고 있다. 따라서 2017년도 조사는 사회경제문화적 변화 및 정책 변화라는 맥락 속에서 노인의 특성과 생활 현황을 다각적으로 파악하여 시계열적인 변화 추이를 파악하고 노인 내부의 다양성을 정확히 파악하는 데 초점을 두었다.

2017 노인실태조사는 노인 학대 경험에 대한 질문도 포함하고 있는데, 이 장에서는 2017 노인실태조사의 데이터로 노인 학대 경험 유무(y값)를 분석하여 위기 노인의 특성을 다각도로 살펴보고자 한다. 노인 학대의 원인은 크게 노인의 인구사회학적 특성, 노인의 경제 및 건강, 심리사회적 기능 요인 등 노인의 의존성과 관련된 요인, 가족 상황적 원인, 사회적 관계망 요인, 사회문화적 원인으로 구분 지어 설명할 수 있다(정경희, 2017).

여기에서는 노인 학대의 원인을 찾고자 하는 것이 아니라, 2017년 노인실태조사를 이용하여 이상(anomaly)에 대한 개념을 다시 정의하고, 기계학습 기반 이상 탐지 기법을 적용하여 다양한 분석을 시도하고 그 특성을 파악하는 데 의미가 있다.

2. 이상(anomaly) 자료 정의 및 기초 통계

여기에서는 이상점, 이상값의 정의 및 개념을 1절 보건 분야 데이터와 다르게 접근하고자 한다. 1절에서는 이상(anomaly)의 개념을 NCI군이 알츠하이머병으로 전환된 환자로 정의하였다. 2절에서도 학대 경험 유무 정보를 활용하여 이상(anomaly)의 개념을 학대 경험이 있는 노인으로 할 수 있다. 하지만, 여기에서는 맥락적 이상(contextual anomaly)으로 접근하여 자료 내에서 개체(조사 대상자)가 특정 맥락에서 이상하다고 판단되는 경우를 이상(anomaly) 개념으로 정의하려고 한다. 3장에서 언급

한 것처럼, 학습 자료에서 모형을 학습해 주어진 맥락에서의 행동을 예측한다. 그리고 구한 예측값과 관측값의 차이가 유의미하면 이상값으로 볼 수 있다. 맥락적 속성이 행동적 속성 예측에 이용되는 것인데, 2017년 노인실태조사 자료에서는 학대 경험 유무를 반응변수(response variable)로 놓고 예측 모형을 적용하여 예측값을 구하고 예측값과 관측값(실제값)의 차이가 나는 경우를 이상(anomaly)이라고 정의하였다.

이를 이항 분류 문제의 분류 행렬로 살펴보면 반응변수의 실제값과 예측값으로 다음과 같이 나타낼 수 있다. 여기에서 anomaly 개념은 정상(학대 경험 없음)인데 비정상(학대 피해 경험 있음)으로 분류(n_{12})되거나, 비정상(학대 경험 있음)인데 정상(학대 피해 경험 없음)으로 분류(n_{21})된 케이스들이 이상값이다.

〈표 4-6〉 복지 데이터에서의 이상(anomaly) 정의

		분류 예측(Predicted)	
		정상	비정상(학대)
분류 결과 (True)	정상	n_{11}	n_{12}
	비정상(학대)	n_{21}	n_{22}

실제 분석을 위해 학대받은 경험 문항에서 결측치를 제외한 1만 83명의 자료를 활용하였다. 학대받은 경험이 있다고 응답한 대상자는 989명으로 전체의 9.8%였다. 예측모형을 위해 사용한 설명변수는 다음과 같다.

〈표 4-7〉 2017년 노인실태조사 예측 모형에 사용한 설명변수

변수명	설명	리코딩(recoding)	비고
ID_add1	리코딩 지역 (동부 1, 읍면부 2)	읍면부 0	ID_add1_adj
ID_SEX	리코딩_성	여성 0	ID_SEX_adj
ID_AGE	리코딩_연령		
ID_MARRIAGE	리코딩_배우자 유무	배우자 없음 0	ID_MARRIAGE_adj
id_GA_H2	리코딩_가구 형태 5분류	1노인독거, 2노인부부, 3기혼자녀 동거 4미혼자녀 동거 5기타	독거가구여부로 변수생성해서 사용(id_GA_H2_adj)
ID_EDU	리코딩_교육 수준		
ID_WORK	리코딩_취업 여부	미취업 0	ID_WORK_adj
ID_ECO	리코딩_가구소득분위		
ID_HEALTH	리코딩_기능상태제한 여부	기능상태제한 있음 1, 기능상태제한 없음 0	ID_HEALTH_adj
KID	생존자녀_유무		
GKID	생존손자녀_유무		
KIN	가까운 친인척_유무		
FRND	친한 친구이웃지인_유무		
B1	평소의 건강상태		
GDS_R	우울 유무		
C1	현재 흡연 여부	아니요 0	C1_adj
N3_PT	공적이전 유무		
N3_BS	기초보장 유무		
N3_PP	사적연금 유무		
N3_PT_S	가구_공적이전소득(총액)		
N3_BS_S	가구_기초보장액(총액)		
N3_PP_S	가구_사적연금소득액(총액)		
NC_P_PT	개인_공적이전 유무		
NC_P_PP	개인_사적연금 유무		
NC_P_PT_W	개인_공적이전소득액(총액)		
NC_P_PP_W	개인_사적연금소득액(총액)		
G4	지난 1년간 자녀와의 갈등 경험	아니요 0	G4_adj
L1	거주 형태(가구)	1자가, 2전세, 3보증금 있는 월세, 4보증금 없는 월세, 5무상	
N6_1_4_a	본인_부채_유무	없음 0	N6_1_4_a_adj

위 정의에 따라 분류 테이블을 작성하기 위해 예측 모형은 예측 성능이 좋은 부스팅 방법을 적용하였고, 10 fold CV로 학습 자료와 테스트 자료를 구분하여 과적합(overfitting) 문제가 발생되지 않도록 하였다. 그리고, 각각의 fold에서 예측확률값 상위 10%만을 비정상(학대 경험 있음)으로 분류하였다. 물론, 예측 모형을 어떤 방법을 적용하느냐, 분류 기준(cut off point)을 어떻게 하느냐에 따라 위 표에서 정의한 테이블 케이스들이 달라진다. 실제로, lift 개념으로 상위 10%를 분류 기준으로 정하였을 때와 G-mean($\sqrt{sensitivity \times specificity}$) measure로 정하였을 때의 분류 결과는 다음과 같다.

〈표 4-8〉 분류 기준(lift 상위 10%)에 따른 분류 행렬

(단위: 명)

구 분		분류 예측(Predicted)	
		정상	비정상(학대)
분류 결과 (True)	정상	8,385	709
	비정상(학대)	688	301

〈표 4-9〉 분류 기준(G-mean)에 따른 분류 행렬

(단위: 명)

구 분		분류 예측(Predicted)	
		정상	비정상(학대)
분류 결과 (True)	정상	5,863	3,231
	비정상(학대)	360	629

이와 같이 분류 기준에 따른 상이한 결과는 이상(anomaly)에 속한 케이스들도 달라지게 된다. 하지만, 본 연구에서는 예측을 잘하고자 하는 목적이 아니라, 맥락적 의미에서 예측 모형으로 분류된 이상(anomaly) 값들에 대한 특성을 파악하기 위한 것으로 합리적이라고 생각하는 분류

기준(lift 상위 10%) 및 예측 모형(부스팅) 방법을 이용하여 분석하였다.

우선, 학대 경험 유무에 따른 설명변수에서의 평균값을 비교해 보면 다음과 같다.

〈표 4-10〉 학대 경험 유무에 따른 T-Test 결과

변수명	y (학대 경험 유무)	평균	평균의 표준오차
리코딩 지역	0	.66	.005
	1	.65	.015
리코딩_성 *	0	.40	.005
	1	.38	.015
리코딩_연령	0	2.49	.013
	1	2.55	.038
리코딩_배우자 유무 *	0	.64	.005
	1	.50	.016
독거가구 여부 *	0	.24	.004
	1	.36	.015
리코딩_교육 수준 *	0	3.35	.014
	1	3.12	.042
리코딩_취업 여부	0	.32	.005
	1	.32	.015
리코딩_가구소득분위 *	0	2.91	.015
	1	2.62	.045
리코딩_기능상태제한 여부 *	0	.26	.005
	1	.34	.015
생존자녀_유무	0	.98	.002
	1	.97	.005
생존손자녀_유무 *	0	.93	.003
	1	.91	.009
가까운 친인척_유무 *	0	.47	.005
	1	.39	.016
친한 친구이웃지인_유무 *	0	.57	.005
	1	.49	.016
평소의 건강상태 *	0	3.06	.010
	1	3.27	.031
우울 유무 *	0	.20	.004
	1	.38	.015
현재 흡연 여부	0	.09	.003

변수명	y (학대 경험 유무)	평균	평균의 표준오차
	1	.10	.010
공적이전 유무 *	0	.95	.002
	1	.96	.006
기초보장 유무 *	0	.06	.003
	1	.15	.011
사적연금 유무	0	.02	.001
	1	.02	.004
가구_공적이전소득(총액) *	0	710.66	8.656
	1	662.30	21.546
가구_기초보장액(총액) *	0	28.71	1.307
	1	67.83	5.840
가구_사적연금소득액(총액) *	0	13.70	1.605
	1	6.28	2.149
개인_공적이전 유무 *	0	.87	.003
	1	.92	.009
개인_사적연금 유무 *	0	.01	.001
	1	.01	.003
개인_공적이전소득액(총액)	0	430.87	6.703
	1	409.93	17.215
개인_사적연금소득액(총액)	0	8.38	1.209
	1	4.27	1.833
지난 1년간 자녀와의 갈등 경험 *	0	.05	.002
	1	.25	.014
거주 형태(가구) *	0	1.71	.014
	1	1.92	.044
본인_부채_유무 *	0	.25	.005
	1	.27	.014

주: 유의수준 10%를 기준으로 *변수명에 표시.

이상 탐지 기법을 적용하기 위해 맥락적 이상(anomaly)의 개념으로 분류하면 4개의 집단으로 나누어지는데, 학대 경험이 없는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{11})과 학대 피해 경험이 있는 것으로 예측된 집단(n_{12})으로 분류되고, 학대 피해 경험이 있는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{21})과 학대 피해 경험이 있는 것으로 예측된 집단(n_{22})으로 분류된다.

4개 집단에 따른 설명변수별 평균값은 다음 표들과 같다.

〈표 4-11〉 학대 경험이 없는 대상자 분류 결과

변수명	N11(8,385명)		N12(709명)	
	평균	표준편차	평균	표준편차
리코딩 지역	.65	.477	.72	.448
리코딩_성	.41	.492	.34	.473
리코딩_연령	2.48	1.212	2.55	1.208
리코딩_배우자 유무	.66	.475	.42	.493
독거가구 여부	.23	.420	.39	.488
리코딩_교육 수준	3.37	1.322	3.06	1.313
리코딩_취업 여부	.32	.467	.27	.443
리코딩_가구소득분위	2.95	1.384	2.43	1.428
리코딩_기능상태제한 여부	.25	.433	.37	.483
생존자녀_유무	.98	.134	.94	.241
생존손자녀_유무	.94	.243	.82	.386
가까운 친인척_유무	.48	.500	.31	.465
친한 친구이웃지인_유무	.58	.493	.44	.496
평소의 건강상태	3.03	.973	3.47	.941
우울 유무	.17	.376	.53	.499
현재 흡연 여부	.09	.289	.12	.320
공적이전 유무	.95	.226	.96	.192
기초보장 유무	.04	.190	.33	.472
사적연금 유무	.02	.133	.01	.118
가구_공적이전소득(총액)	713.21	834.506	680.50	709.657
가구_기초보장액(총액)	17.29	95.748	163.72	266.798
가구_사적연금소득액(총액)	14.37	158.323	5.75	63.257
개인_공적이전 유무	.87	.337	.94	.246
개인_사적연금 유무	.01	.108	.01	.106
개인_공적이전소득액(총액)	435.72	648.296	373.43	516.292
개인_사적연금소득액(총액)	8.70	118.824	4.62	59.694
지난 1년간 자녀와의 갈등 경험	.00	.051	.64	.480
거주 형태(가구)	1.66	1.299	2.22	1.438
본인_부채_유무	.24	.429	.28	.450

〈표 4-12〉 학대 경험이 있는 대상자 분류 결과

변수명	N21(688명)		N22(301명)	
	평균	표준편차	평균	표준편차
리코딩 지역	.64	.480	.66	.473
리코딩_성	.38	.485	.37	.483
리코딩_연령	2.55	1.203	2.56	1.181
리코딩_배우자 유무	.56	.496	.35	.476
독거가구 여부	.31	.464	.48	.500
리코딩_교육 수준	3.22	1.346	2.89	1.284
리코딩_취업 여부	.36	.479	.25	.433
리코딩_가구소득분위	2.76	1.426	2.31	1.390
리코딩_기능상태제한 여부	.31	.463	.40	.490
생존자녀_유무	.97	.164	.97	.161
생존손자녀_유무	.92	.274	.89	.317
가까운 친인척_유무	.42	.494	.32	.467
친한 친구이웃지인_유무	.52	.500	.42	.494
평소의 건강상태	3.17	.977	3.50	.972
우울 유무	.29	.452	.59	.493
현재 흡연 여부	.10	.295	.11	.309
공적이전 유무	.96	.201	.98	.140
기초보장 유무	.07	.260	.31	.464
사적연금 유무	.02	.136	.01	.081
가구_공적이전소득(총액)	667.03	729.455	651.51	541.619
가구_기초보장액(총액)	33.51	132.846	146.28	248.592
가구_사적연금소득액(총액)	8.06	78.604	2.19	29.516
개인_공적이전 유무	.90	.295	.95	.225
개인_사적연금 유무	.01	.114	.00	.058
개인_공적이전소득액(총액)	418.56	578.738	390.21	444.676
개인_사적연금소득액(총액)	5.44	66.641	1.59	27.667
지난 1년간 자녀와의 갈등 경험	.01	.100	.80	.403
거주 형태(가구)	1.82	1.401	2.13	1.367
본인_부채_유무	.25	.432	.31	.461

학대 경험이 없는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{11})과 학대 피해 경험이 있는 것으로 예측된 집단(n_{12})으로 분류된 각각의 특성을 살펴보면, n_{11}집단은 남성의 비중이 상대적으로 높고, 배우자가 있는 경우가 많으며 교육 수준, 가구소득분위도 상대적으로 높다. 반면에, n_{12}집단은 독거가구 비중이 높으며 기능상태제한이 있는 경우가 많으며 우울한 정도가 상대적으로 높았다. 그리고 기초보장을 받고 있는 경우가 많으며 지난 1년간 자녀와의 갈등 경험이 많았다. 이러한 결과는 학대 피해 경험이 있는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{21})과 학대 피해 경험이 있는 것으로 예측된 집단(n_{22})으로 분류된 특성과도 관련이 있다. n_{11}집단의 특성은 n_{21}집단의 특성과 비슷하며, n_{12}집단의 특성은 n_{22}집단의 특성과 유사하다고 할 수 있다. 이러한 결과는 예측 모형에서 n_{12}와 n_{21}에 해당되는 대상자들이 오분류된 것과 무관하지 않다.

3. 탐색적 자료 분석

여기에서는 3장의 분석 방법들을 적용해서 앞에서 정의한 맥락적 이상(anomaly) 자료에 대한 탐색적 자료 분석을 해 보고자 한다. 사용한 방법은 4가지로, 이상 탐지의 통계적 기법의 혼합 모수적 방법, NN(nearest neighbor) 기반 이상 탐지 기법의 LOF, 스펙트럴 이상 탐지 기법의 t-SNE, 군집화 방법(clustering)의 DBSCAN 분석 기법을 적용하였다. 자료 분석에서는 학대 경험 유무(y) 정보를 제외하고 설명변수 정보만 활용하여 분석하였으며, 결과를 제시할 때는 앞서 정의한 4개의 집단 정보를 토대로 특성을 살펴보았다.

가. 혼합 모수적 방법을 이용한 탐색적 자료 분석

우선, 혼합 모수적 방법을 적용하여 위 4개 군집 class를 표시해 보면 다음과 같다.

〈표 4-13〉 Mixture model 적용 군집 분류

변수명	Class 1 평균	Class 2 평균	Class 3 평균	Class 4 평균
리코딩 지역	.62	.65	.70	.80
리코딩_성	0.00	1.00	.42	.38
리코딩_연령	2.63	2.53	2.31	2.31
리코딩_배우자 유무	.42	.90	.68	.47
독거가구 여부	.38	.08	.21	.45
리코딩_교육 수준	2.78	3.94	3.47	3.65
리코딩_취업 여부	.30	.39	.29	.21
리코딩_가구소득분위	2.64	3.18	3.00	2.48
리코딩_기능상태제한 여부	.35	.15	.25	.26
생존자녀_유무	1.00	1.00	1.00	.44
생존손자녀_유무	.97	.94	.91	.43
가까운 친인척_유무	.48	.45	.45	.44
친한 친구이웃지인_유무	.57	.56	.55	.57
평소의 건강상태	3.21	2.87	3.08	3.20
우울 유무	.23	.16	.25	.24
현재 흡연 여부	0.00	0.00	.30	.11
공적이전 유무	1.00	1.00	.84	.92
기초보장 유무	0.00	0.00	.19	.32
사적연금 유무	0.00	0.00	0.00	.44
가구_공적이전소득(총액)	547.52	886.33	758.44	719.99
가구_기초보장액(총액)	0.00	0.00	90.65	145.90
가구_사적연금소득액(총액)	0.00	0.00	0.00	329.50
개인_공적이전 유무	1.00	1.00	.61	.83
개인_사적연금 유무	0.00	0.00	0.00	.29
개인_공적이전소득액(총액)	344.01	737.64	276.20	373.55
개인_사적연금소득액(총액)	0.00	0.00	0.00	202.70
지난 1년간 자녀와의 갈등 경험	0.00	0.00	.24	.02
거주 형태(가구)	1.84	1.45	1.80	1.93
본인_부채_유무	.20	.26	.30	.30

군집 결과, class 1은 모두 여성이고 독거가구 비중이 높으며 학력이 낮은 편이며 건강에 기능 제한이 있고 공적이전소득이 낮은 집단이라고 할 수 있다. class 2는 모두 남성이고 건강한 집단이며, 공적이전소득금액이 높은 집단이다. class 3은 자녀와의 갈등이 다른 집단에 비해 높으며, 학대 피해 경험 비율이 다른 집단에 비해 높은 집단이다. class 4는 독거가구 비중이 높으며 가구의 사적이전소득이 높고, 가구 기초보장액 역시 높은 집단이다.

〈표 4-14〉 Mixture model 적용 군집 분류에 따른 학대 경험 유무 비율

변수명	Class 1 평균	Class 2 평균	Class 3 평균	Class 4 평균
학대 경험 유무	.08	.07	.15	.10

〈표 4-15〉 4개 집단과 군집분석 결과 교차분석

(단위: 명,%)

집단 분류 \ 군집 분류		군집 class 1	2	3	4	합
4개 집단	n_{11}	**3676 (43.84%)**	2469 (29.45%)	1937 (23.10%)	303 (3.61%)	8385 (100%)
	n_{12}	43 (6.06%)	9 (1.27%)	604 (85.19%))	53 (7.48%)	709 (100%)
	n_{21}	319 (46.37%)	177 (25.73%)	161 (23.40%)	31 (4.51%)	688 (100%)
	n_{22}	11 (3.65%)	1 (0.33%)	**279 (92.69%)**	10 (3.32%)	301 (100%)

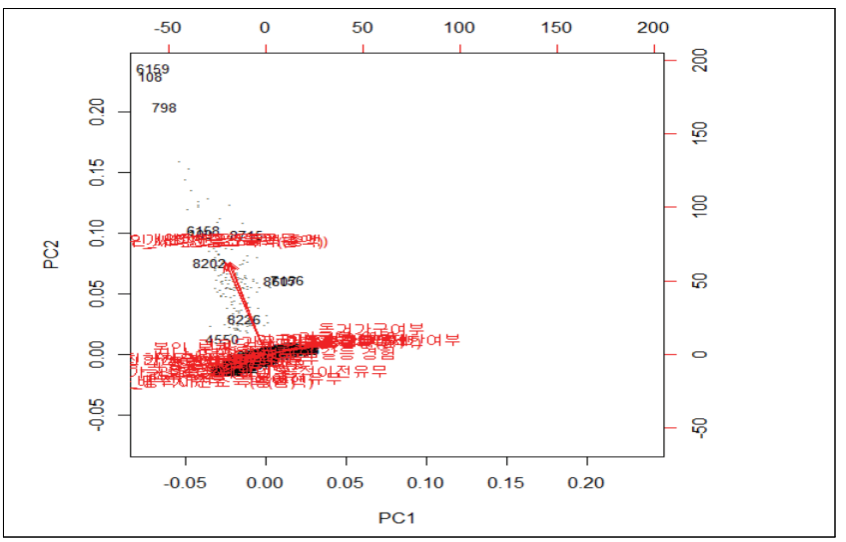
맥락적 이상(anomaly) 개념으로 분류된 4개 집단과 군집분석의 결과를 교차분석해 보면, n_{11} 집단과 n_{21} 집단은 군집 1 class에 많이 분포해 있으며, n_{12} 집단과 n_{22} 집단은 군집 3 class에 주로 분포해 있다고 할 수

있다. 이는 학대 경험이 없는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{11})과 학대 피해 경험이 있는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{21})의 특성이 비슷하다는 것을 알 수 있다. 그리고, 학대 경험이 없는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{12})과 학대 피해 경험이 있는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{22})의 특성이 유사하다는 것을 알 수 있다.

나. 상대 밀도(LOF)를 이용한 탐색적 자료 분석

상대 밀도를 이용한 탐색적 자료 분석에서는 LOF를 학대 경험이 없는 대상자와 학대 경험이 있는 대상자 데이터셋으로 나누어 분석을 하였다. LOF는 값의 크기가 상위 100개의 이상값을 각각 추출하였으며 인덱스를 표시하면 다음과 같다.

[그림 4-7] 학대 경험이 없는 대상자 데이터에서의 LOF



그림에서 보면 가구 공적이전소득(총액) 변수와 개인 공적이전소득(총액) 변수가 이상값을 분류하는 데 중요한 변수라고 할 수 있다. LOF에 의한 상위 100명은 대부분 공적이전소득이 있고, 기초보장 수급이 없는 대상자들이다.

〈표 4-16〉 학대 경험이 없는 대상자 데이터에서 LOF에 의한 상위 100 이상값의 2개 집단과 군집분석 결과 교차분석

(단위: 명,%)

군집 분류 / 집단 분류		군집 class				합
		1	2	3	4	
2개 집단	n_{11}	22	**38**	21	10	91
	n_{12}	2	1	**4**	2	9
합		24	39	**25**	12	100

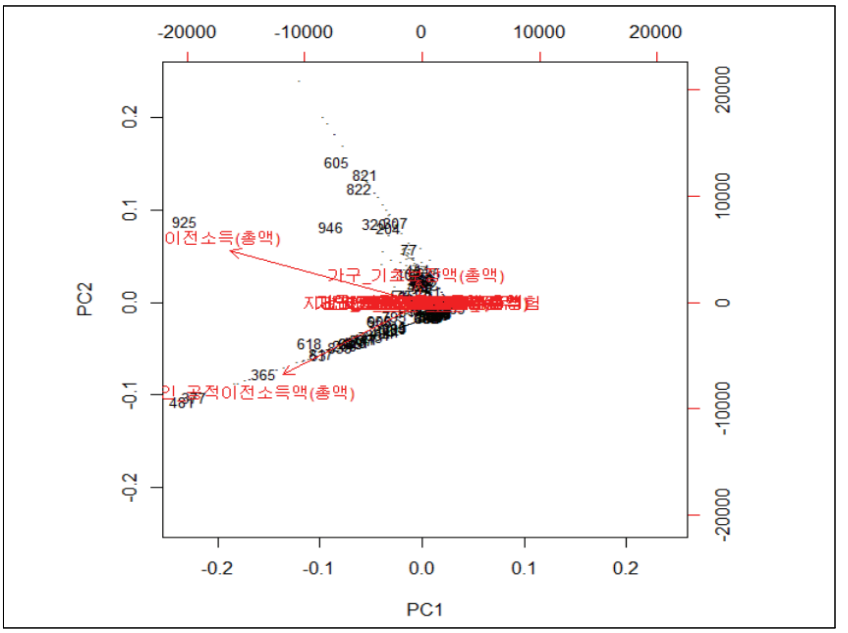
상위 100 이상값의 특성을 집단 분류와 군집 분류로 나누어 살펴보면, 학대 경험이 없는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{11})에 대부분(100명 중 91명) 속하는데, LOF에서 연속형 변수인 공적이전소득이 높은 케이스들을 이상값으로 판단하고 있어 군집 2 class의 특성값을 많이 반영하고 있다. 학대 경험이 없는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{12})은 군집에서 3 class에 많이 속해 있다. 이러한 특성은 혼합 모수적 방법을 이용한 탐색적 자료 분석과 크게 다르지 않다.

다음으로, LOF를 학대 경험이 있는 대상자 데이터셋에서 분석하였다. LOF는 값의 크기가 상위 100개의 이상값을 각각 추출하였으며, 그 인덱스를 표시해 보면 다음과 같다.

그림에서 보면 가구 공적이전소득(총액) 변수와 개인 공적이전소득(총

액) 변수, 가구 기초보장액(총액)이 이상값을 분류하는 데 중요한 변수라고 할 수 있다.

[그림 4-8] 학대 경험이 있는 대상자 데이터에서의 LOF



〈표 4-17〉 학대 경험이 있는 대상자 데이터에서 LOF에 의한 상위 100 이상값의 2개 집단과 군집분석 결과 교차분석

(단위: 명)

집단 분류 \ 군집 분류		군집 class				합
		1	2	3	4	
2개 집단	n_{21}	**20**	**20**	19	8	67
	n_{22}	1	0	**31**	1	33
합		21	20	50	9	100

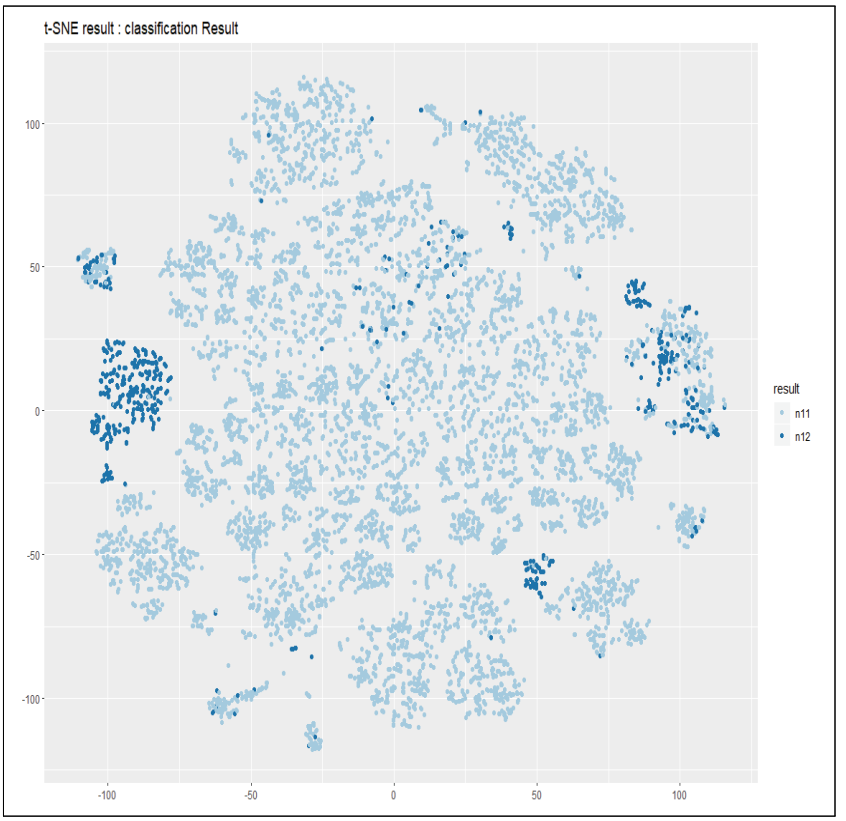
상위 100 이상값의 특성을 집단 분류와 군집 분류로 나누어 살펴보면, LOF값 상위 100명 중 67명이 여기에서 정의한 맥락적 이상에 속해 있다. 즉, 학대 경험이 있는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{21})에 대부분 속하며, 학대 경험이 있는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{22})은 대부분 군집에서 3 class에 속해 있다.

위 결과는 학대 경험 유무로 자료를 나누어 LOF값이 큰 100 대상자들의 특성을 앞서 정의한 2개 집단과 군집분석 결과로 사후적으로 살펴본 것으로, LOF값이 큰 대상자들의 숫자를 늘리면 혼합 모수적 방법을 이용한 탐색적 자료 분석 결과와 유사할 것이다.

다. t-SNE 및 DBSCAN 방법을 이용한 탐색적 자료 분석

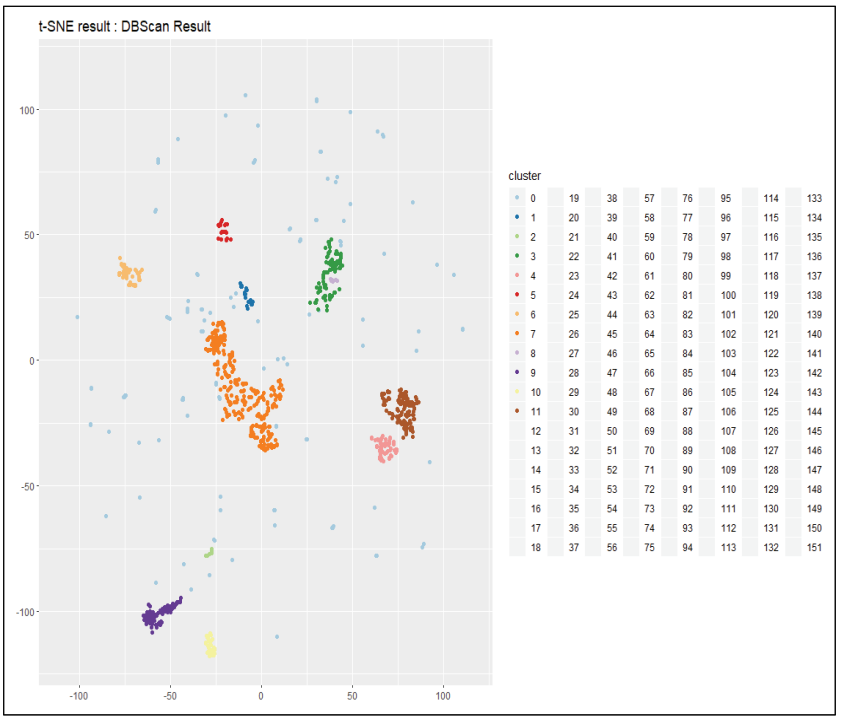
데이터 시각화 및 차원 축소 방법인 t-SNE은 원자료상의 자료 분포와 사영된 공간에서의 자료 분포 간 쿨백-라이블러 발산(Kullback-Leibler divergence 또는 KL divergence)을 최소화하여 저차원의 사영된 자료를 계산한다. t-SNE은 비모수적 알고리즘이고, 고차원 자료에서 많이 활용되는 방법이다. t-SNE 방법을 적용하여 학대 경험이 없는 대상자 데이터와 학대 경험이 있는 대상자 데이터를 나누어 탐색적 자료 분석을 실시하였다. 그리고, 군집화 기반 이상 탐지 기법으로 '정상값들은 하나 또는 몇 개의 군집에 모여 있고, 이상값은 군집에 속하지 않는다'고 가정하는 DBSCAN 알고리즘을 적용하여 t-SNE 방법 분석 결과 위에 DBSCAN 결과를 제시하여 세부적으로 그 특성들을 살펴보았다.

[그림 4-9] 학대 경험이 없는 대상자 데이터에서의 t-SNE 분석 결과(집단 분류 표시)



학대 경험이 없는 대상자 데이터에서 t-SNE 분석 결과를 학대 경험이 없는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{11})과 학대 경험이 없는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{12})으로 색깔로 구분하여 그려 보면 차이가 난다. 짙은 색으로 표시한 학대 경험이 없는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{12})은 군데군데 몰려 있는 것을 확인할 수 있다.

[그림 4-10] 학대 경험이 없는 대상자 데이터에서의 t-SNE 분석 결과(DBSCAN 알고리즘 결과 표시)



t-SNE 방법 분석 결과 위에 DBSCAN 결과를 그려 보면 총 152개의 cluster가 그려진다. 이 군집 결과를 맥락적 이상이라고 정의한 학대 경험이 없는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{12})의 순수도(purity)를 계산하고, 그 값이 높은 군집의 특성을 살펴볼 필요가 있다. 순수도는 특정 범주의 개체들이 포함되어 있는 정도를 의미하기에, 각 클러스터의 특징을 알 수 있다. 순수도(purity)값이 0.85가 넘는 군집은 16, 64, 100, 102, 128, 138이다.

〈표 4-18〉 학대 경험이 없는 대상자 데이터에서 DBSCAN_군집 결과

DBSCAN_군집번호	n11 대상자수	n12 대상자수	n12_purity
0	172	17	0.089947
1	18	6	0.25
2	8	0	0
3	113	0	0
4	72	0	0
5	27	0	0
6	54	0	0
7	369	0	0
8	9	0	0
9	94	7	0.069307
10	44	2	0.043478
11	149	0	0
12	21	0	0
13	342	0	0
14	234	0	0
15	158	0	0
16	6	277	0.978799
17	134	4	0.028986
18	73	0	0
19	5	0	0
20	347	1	0.002874
21	54	0	0
22	506	1	0.001972
23	5	0	0
24	310	0	0
25	213	2	0.009302
26	25	0	0
27	60	0	0
28	191	0	0
29	342	1	0.002915
30	145	2	0.013605
31	8	0	0
32	69	52	0.429752

DBSCAN_군집번호	n11 대상자수	n12 대상자수	n12_purity
33	146	0	0
34	195	0	0
35	12	0	0
36	6	0	0
37	41	0	0
38	85	0	0
39	46	0	0
40	104	7	0.063063
41	177	23	0.115
42	12	0	0
43	124	0	0
44	31	0	0
45	32	0	0
46	89	0	0
47	133	0	0
48	6	0	0
49	35	0	0
50	19	0	0
51	61	0	0
52	74	5	0.063291
53	42	0	0
54	43	1	0.022727
55	49	3	0.057692
56	39	0	0
57	50	0	0
58	70	38	0.351852
59	17	0	0
60	121	68	0.359788
61	7	0	0
62	37	0	0
63	29	0	0
64	3	24	0.888889
65	44	0	0
66	149	0	0

DBSCAN_군집번호	n11 대상자수	n12 대상자수	n12_purity
67	28	0	0
68	32	0	0
69	5	0	0
70	45	0	0
71	27	9	0.25
72	11	14	0.56
73	5	0	0
74	18	0	0
75	1	4	0.8
76	4	6	0.6
77	8	0	0
78	62	0	0
79	40	2	0.047619
80	45	0	0
81	11	0	0
82	6	0	0
83	45	0	0
84	17	0	0
85	27	0	0
86	85	0	0
87	126	0	0
88	23	0	0
89	26	0	0
90	18	0	0
91	12	0	0
92	9	0	0
93	4	2	0.333333
94	7	0	0
95	0	40	1
96	41	0	0
97	33	0	0
98	39	0	0
99	92	0	0
100	4	43	0.914894

DBSCAN_군집번호	n11 대상자수	n12 대상자수	n12_purity
101	44	0	0
102	1	11	0.916667
103	30	0	0
104	9	3	0.25
105	54	0	0
106	12	0	0
107	11	0	0
108	68	0	0
109	74	0	0
110	26	0	0
111	26	0	0
112	14	0	0
113	20	0	0
114	6	0	0
115	36	0	0
116	46	0	0
117	7	0	0
118	16	1	0.058824
119	18	0	0
120	1	4	0.8
121	9	0	0
122	7	0	0
123	12	0	0
124	10	0	0
125	50	0	0
126	15	0	0
127	43	0	0
128	1	18	0.947368
129	9	0	0
130	5	0	0
131	14	0	0
132	7	0	0
133	22	0	0
134	13	0	0

DBSCAN_군집번호	n11 대상자수	n12 대상자수	n12_purity
135	14	0	0
136	10	0	0
137	30	0	0
138	1	10	0.909091
139	24	0	0
140	17	0	0
141	8	0	0
142	20	0	0
143	10	1	0.090909
144	7	0	0
145	7	0	0
146	5	0	0
147	5	0	0
148	6	0	0
149	13	0	0
150	5	0	0
151	6	0	0

Purity값이 0.85가 넘는 군집인 16, 64, 100, 102, 128, 138에 대해 각각의 군집 특성을 살펴볼 필요가 있다. 이는 맥락적 이상이라고 정의한 케이스들을 세세하게 나누어서 특성들을 분류한 결과라고 볼 수 있다.

〈표 4-19〉 학대 경험이 없는 대상자 데이터에서 DBSCAN_일부 군집 특성

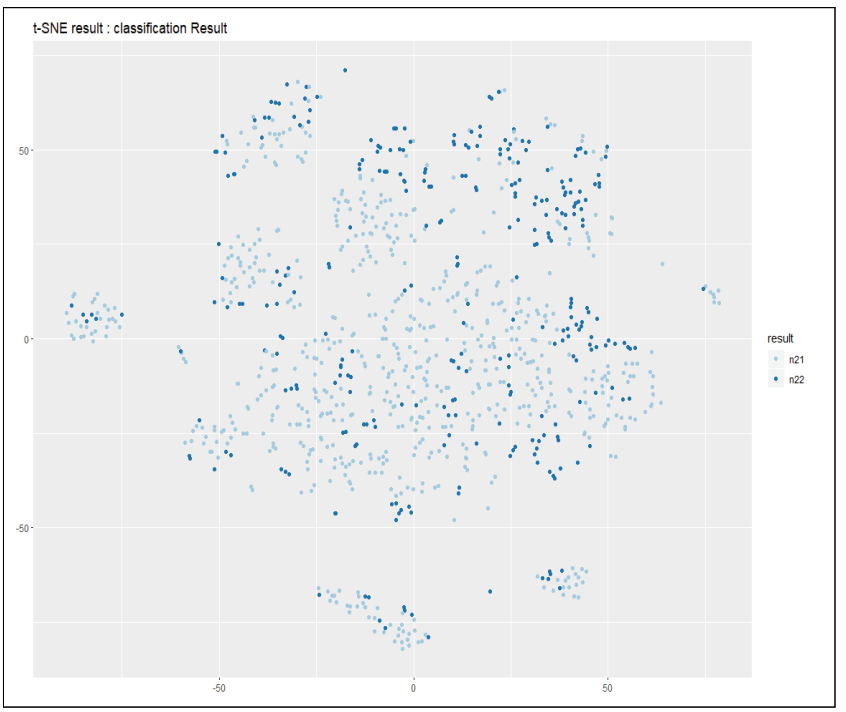
설명변수	DBSCAN_군집번호					
	16	64	100	102	128	138
리코딩 지역(동부 1, 읍면부 2)	.73	.67	.89	.50	.68	.64
리코딩_성	.36	.85	.30	0.00	.37	.91
리코딩_연령	2.66	2.04	1.96	1.25	2.58	1.45
리코딩_배우자 유무	.54	.74	.51	1.00	.74	.91
리코딩_독거가구 여부	.26	.26	.06	0.00	.11	.09
리코딩_교육 수준	3.06	3.37	3.85	3.50	4.00	4.55
리코딩_취업 여부	.32	.52	.30	.42	.11	.55

설명변수	DBSCAN_군집번호					
	16	64	100	102	128	138
리코딩_가구소득분위	2.72	3.15	3.49	3.00	3.21	3.73
리코딩_기능상태제한 여부	.33	.19	.23	.17	.05	.09
생존자녀_유무	1.00	1.00	1.00	1.00	1.00	1.00
생존손자녀_유무	1.00	1.00	0.00	.92	1.00	0.00
가까운 친인척_유무	.33	.37	.28	.50	.53	.36
친한 친구이웃지인_유무	.48	.33	.53	.50	.68	.55
평소의 건강상태	3.29	3.15	3.36	2.83	3.26	2.82
우울 유무	.33	.33	.45	.42	.32	.36
현재 흡연 여부	.01	1.00	0.00	.08	.05	1.00
공적이전 유무	1.00	1.00	.98	1.00	0.00	1.00
기초보장 유무	0.00	0.00	.06	0.00	0.00	.09
사적연금 유무	0.00	0.00	0.00	0.00	0.00	0.00
가구_공적이전소득(총액)	505.95	641.00	640.85	769.42	0.00	613.55
가구_기초보장액(총액)	0.00	0.00	25.72	0.00	0.00	44.82
가구_사적연금소득액(총액)	0.00	0.00	0.00	0.00	0.00	0.00
개인_공적이전 유무	1.00	1.00	.98	0.00	0.00	1.00
개인_사적연금 유무	0.00	0.00	0.00	0.00	0.00	0.00
개인_공적이전소득액(총액)	379.22	408.81	472.94	0.00	0.00	449.00
개인_사적연금소득액(총액)	0.00	0.00	0.00	0.00	0.00	0.00
지난 1년간 자녀와의 갈등 경험	1.00	1.00	1.00	1.00	1.00	1.00
거주 형태(가구)	1.95	1.85	1.60	1.92	1.47	1.73
본인_부채_유무	.26	.22	.26	.17	.79	.55

16번 군집은 가구 공적이전소득금액이 높으며, 기초보장액을 받고 있지 않은 집단이다. 64번 군집은 대부분 남자로 구성되어 있고, 가구 공적이전소득금액이 높으며, 기초보장액을 받고 있지 않은 집단이다. 100번 군집은 일부 케이스가 기초보장 수급을 받는 집단이고, 138번 군집도 마찬가지로 일부 케이스가 기초보장 수급을 받는다. 102번 집단은 모두 여성이고 개인의 공적이전이 없고 가구의 기초보장 수급도 받지 않는 특성이 있다. 128번 집단은 가구의 공적이전소득도 0인 특성을 가지고 있다.

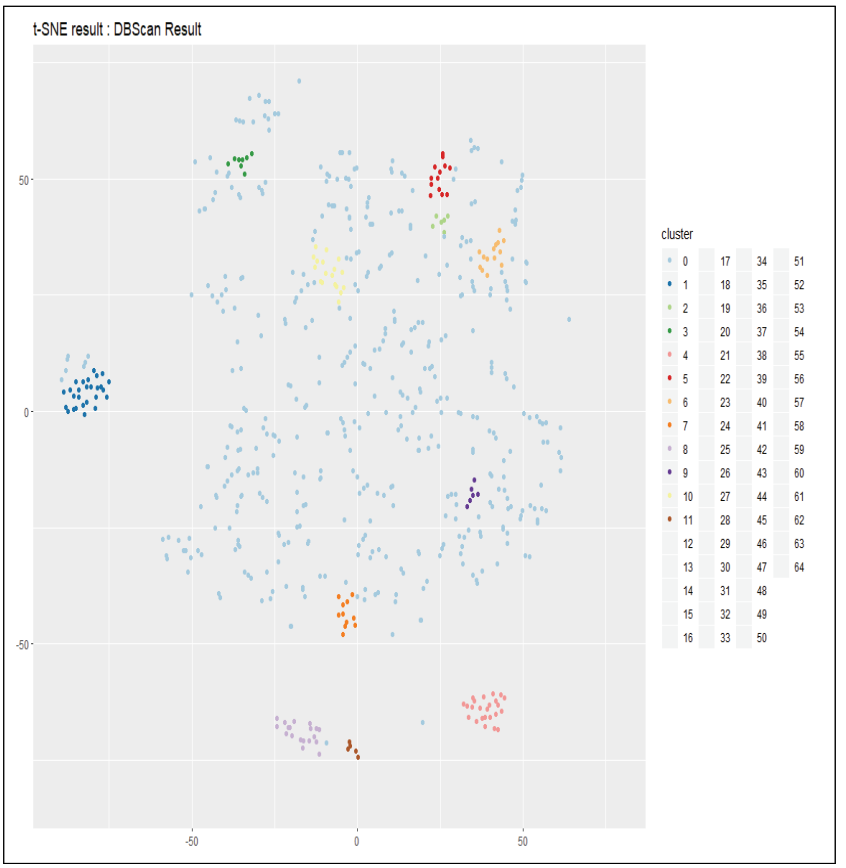
학대 경험이 있는 대상자 데이터에서 t-SNE 분석 결과를 학대 경험이 있는 대상자(참값)가 학대 경험이 없는 것으로 예측된 집단(n_{21})과 학대 경험이 있는 대상자(참값)가 학대 피해 경험이 있는 것으로 예측된 집단(n_{22})으로 색깔을 구분하여 그린 결과는 다음과 같다.

[그림 4-11] 학대 경험이 있는 대상자 데이터에서의 t-SNE 분석 결과(집단 분류 표시)



짙은 색으로 표시한 학대 경험이 있는 대상자(참값)가 학대 피해 경험이 없는 것으로 예측된 집단(n_{21})은 전반적으로 흩어져 있는 것으로 나타났다.

[그림 4-12] 학대 경험이 있는 대상자 데이터에서의 t-SNE 분석 결과(DBSsan 알고리즘 결과 표시)



t-SNE 방법 분석 결과 위에 DBSCAN 결과를 그려 보면 총 65개의 cluster가 그려진다. 이 군집 결과를 맥락적 이상이라고 정의한 학대 경험이 있는 대상자(참값)가 학대 피해 경험이 없는 것으로 예측된 집단(n_{21})의 purity를 계산하고, 그 값이 높은 군집의 특성을 살펴보고자 한다.

학대 경험이 있는 대상자 데이터에서 군집별 Purity값이 1인 집단이 23개이다. 반대로, purity값이 0인 집단은 전체 중 3개이다. 이는 맥락적

이상에 해당하는 학대 경험이 있는 대상자(참값)가 학대 피해 경험이 없는 것으로 예측된 집단의 특성이 여러 군집으로 흩어져 있음을 의미한다.

〈표 4-20〉 학대 경험이 있는 대상자 데이터에서 DBSCAN_군집 결과

DBSCAN 군집번호	n21 대상자수	n22 대상자수	n21_purity
0	294	159	0.649007
1	22	5	0.814815
2	2	4	0.333333
3	7	1	0.875
4	18	6	0.75
5	4	9	0.307692
6	2	12	0.142857
7	4	7	0.363636
8	16	3	0.842105
9	5	1	0.833333
10	18	0	1
11	2	3	0.4
12	7	1	0.875
13	14	2	0.875
14	4	1	0.8
15	7	0	1
16	6	1	0.857143
17	7	1	0.875
18	1	4	0.2
19	5	0	1
20	6	1	0.857143
21	10	0	1
22	22	5	0.814815
23	10	1	0.909091
24	18	3	0.857143
25	3	3	0.5
26	8	0	1
27	5	0	1
28	1	9	0.1
29	6	0	1
30	8	0	1
31	9	0	1
32	2	4	0.333333
33	10	3	0.769231
34	4	0	1

DBSCAN 군집번호	n21 대상자수	n22 대상자수	n21_purity
35	10	0	1
36	10	0	1
37	2	3	0.4
38	4	1	0.8
39	1	4	0.2
40	7	0	1
41	5	0	1
42	3	3	0.5
43	6	0	1
44	4	3	0.571429
45	7	0	1
46	4	1	0.8
47	2	3	0.4
48	3	1	0.75
49	0	6	0
50	5	0	1
51	1	4	0.2
52	0	5	0
53	6	0	1
54	6	1	0.857143
55	7	0	1
56	2	3	0.4
57	5	0	1
58	2	3	0.4
59	1	4	0.2
60	5	0	1
61	0	6	0
62	5	0	1
63	4	1	0.8
64	4	0	1

4. 시사점

여기에서는 노인 학대 노출 특성 분석을 위해 맥락적 이상(anomaly) 의 개념을 정의하였고, 이상 탐지의 통계적 기법의 혼합 모수적 방법, NN (nearest neighbor) 기반 이상 탐지 기법의 LOF, 스펙트럴 이상 탐지 기법의 t-SNE, 군집화 방법(clustering)의 DBSCAN 분석 기법을 적용하여 탐색적 자료 분석을 실시하였다. 우선, 이상 탐지의 통계적 기법 중 하나인 혼합 모수적 방법으로 맥락적 이상에 해당하는 케이스가 4개의 class에서 어디에 속하는지를 교차분석을 통해 살펴보았다. 그 다음, 학대 경험이 없는 대상자 데이터와 학대 경험이 있는 대상자 데이터를 나누어 NN(nearest neighbor) 기반 이상 탐지 기법의 LOF 방법을 적용하여 LOF값 상위 100명의 특성이 맥락적 이상에 해당하는 케이스들을 포함하고 있는지에 대한 분석을 실시하였다. 스펙트럴 이상 탐지 기법의 t-SNE 방법 역시, 학대 경험이 없는 대상자 데이터와 학대 경험이 있는 대상자 데이터를 나누어 맥락적 이상에 해당하는 케이스들을 각각 시각적으로 표현하였고 군집화 방법(clustering)의 DBSCAN 분석 기법을 적용하여 맥락적 이상에 해당하는 케이스들의 특성을 세부적으로 살펴보았다.

이상 탐지 기법을 적용하여 자료를 분류한다면, 혼합모형으로는 대분류 속성을 파악할 수 있고, 2단계의 중분류를 하면 LOF값으로, 소분류 속성으로 내려가면 DBSCAN 방법을 사용하여 특성을 파악할 수 있다. 즉, 대분류부터 정의해 나간가면 top-down 방식으로 혼합모형 -〉 LOF -〉 DBSCAN의 방식을 적용하고, 소분류부터 정의한다면 bottom-up 방식으로 DBSCAN -〉 LOF -〉 혼합모형을 적용하여 분석하는 것이 적절하다고 판단된다. 물론 이러한 분류 시, 각 방법의 특성 및 장단점도 충분히 고려해보아야 한다.

제 5 장

이상 탐지 기법 이슈 및 정책 제언

5 이상 탐지 기법 이슈 및 정책 제언

제1절 이상 탐지 기법 관련 이슈

다양한 분야에서 사용되는 현대적인 시스템들은 매우 복잡한 구조를 지니고 있으며, 따라서 비정상적인 작동의 종류가 매우 다양하게 나타나게 된다. 심지어는 기존에 관측되지 못한 형태의 비정상 자료가 새롭게 관측될 확률 또한 매우 높게 나타난다. 이러한 문제점 때문에 기존의 다중분류 기법으로 비정상 자료의 분류를 진행하는 것에는 많은 어려움이 따르게 된다.

이러한 문제점을 해결하기 위해 제시된 방법이 이상 탐지 기법이다.

추상적으로 이상값은 예상되는 행동을 따르지 않는 패턴으로 정의한다. 따라서 간단한 접근으로 정상을 나타내는 영역을 정하고 그 영역 밖에 있는 값을 이상값으로 하는 방식을 생각할 수 있다. 하지만 이를 매우 까다롭게 하는 다음과 같은 요인들이 있다.

- 모든 정상 개체를 포함하는 영역을 설정하기 어렵다. 또한 정상과 이상의 경계가 대체로 불분명하기 때문에 경계 근처의 값들이 잘못 판단될 수 있다.
- 만약 이상값이 의도적인 행동의 결과라면, 그 주체가 마치 정상값인 것처럼 꾸며 결과적으로 정상을 정의하는 것이 너무 어려워진다.
- 분야에 따라 정상 자료에서 어느 정도 떨어지면 이상값이라고 판단하는 크기(scale)가 다르다. 예시로, 체온은 정상값에서 조금만 달라져도 이상으로 취급하지만, 주식시장에서 그 정도의 변화는 무의

미하다. 따라서 특정 분야에서 학습된 이상 탐색 기준을 여러 분야에 동시에 적용하는 것은 쉽지 않다.

- 이상 탐지 분야에서 사용하는 학습/검증 자료 내 각 개체가 정상인지 이상값인지에 대한 라벨은 일반적으로 구하기 힘들다.
- 자료에 정상 자료 방향으로 치우친 잡음이 있으면 그것을 구별하고 제거하기 쉽지 않다.

이러한 문제 때문에 이상 탐지는 더 어려워진다. 실제로 대부분 제안된 기법들은 자료의 본질, 라벨의 존재 여부, 이상값의 종류 등 다양한 요인으로 정해지는 특정한 형태의 문제만을 해결한다. 현재까지 주어진 자료(또는 문제)의 형태에 따른 다양한 이상 탐지에 대한 연구가 진행되었고, 대부분의 방법론은 통계학, 기계학습, 데이터마이닝, 정보 이론, 스펙트럴 이론을 포함한 많은 분야의 방법론을 기반으로 한다.

또한 시간 자료에서 이상 탐지의 추가적인 난제는 다음과 같다.

- 특정 자료 형태와 문제에 따른 이상 탐지 모형이 존재하기 때문에 매우 다양한 모형이 존재하고, 이로 인해 규격화된 모형을 제시하기 쉽지 않다.
- 매 시간 새로운 자료가 얻어지기 때문에, 자료의 처리 및 저장 공간의 한계 등 어려움이 존재한다. 특히 스트리밍 데이터에서는 오직 한 번의 자료 검토 기회가 주어진다는 어려움이 있다.
- 자료가 분산된 상황에서 시간 자료의 이상 탐지는 노드 간의 상호작용을 최소화하면서 제한된 자원을 이용해 계산해야 하는 어려움이 존재한다.

이 연구에서는 이상 탐지 문제가 만들어지는 과정들을 논의했고, 다양한 기법에 대한 논문들을 개략적으로 설명하고자 했다. 각 (기반 이론) 분야에 대해 정상과 이상 자료에 대한 특수한 가정을 살펴보았다. 이 가정

들은 적용한 기법의 효율성을 평가하는 지침이 될 수 있다. 현재의 연구들은 통일된 이상값의 개념이 없는 채로 체계가 없이 이루어져 이상 탐지 문제에 대한 이론적인 이해가 매우 어려운 상황이다. 여러 기법의 이상값의 개념 및 가정을 통계적 틀이나 기계학습의 틀 안에서 통합하는 것이 앞으로의 가능성 있는 과제를 위해서도 필요하기에 제3장에서 최근 방법에 대해 소개하였다. 또한 Knorr and Ng(1997)이 이에 대해 제한된 시도를 했는데, 2차원 자료에 대한 거리 기반 이상과 통계적 이상의 관계를 보인 것이었다.

이상 탐지 기법에서는 여러 전도유망한 추후 연구 방향이 있다. 맥락적, 집단 이상 탐지 기법들이 몇몇 영역에서 자리를 잡아 가고 있고, 새로운 기법들이 개발, 발전할 여지도 많다. 다수의 위치에 퍼져 있는 자료를 위해서는 분산(distributed) 이상 탐지 기법이 있어야 한다(Zimmerman & Mohay, 2006). 이러한 기법들은 여러 장소에서 정보를 처리하면서도 각 장소의 사적인 정보는 보호해야 하는데, 여기서 개인정보 보호(privacy-preserving) 이상 탐지 기법이 필요하다(Vaidya & Clifton, 2004). 한편 센서 네트워크의 등장으로 자료를 즉각적으로 처리하는 것이 필수가 되었다. 본 연구에서 다룬 기법 중 다수는 이상값을 찾기 전에 테스트 자료를 온전히 갖추어야 하지만, 최근에는 온라인 방식으로 작동하는 기법들이 제안되고 있다. 이러한 기법들은 테스트 자료를 받자마자 이상 점수를 매길 뿐 아니라 그것을 이용해 점진적으로 모형을 개선해 나간다. 이외에도 항공 체계와 같은 복잡한 시스템에서의 이상 탐지의 활용성이 계속해서 늘어나고 있다. 이러한 시스템에 대한 이상 탐지 기법은 여러 성분 사이의 교호작용을 모형화할 수 있어야 한다.

제2절 이상 탐지 기법의 활용성과 정책 제언

이상 탐지 기법은 학습 자료를 기반으로 기존의 자료들과는 다른 특성을 갖는 자료를 찾는 모형을 만드는 방법으로, 대부분의 자료가 정상 분류이고 극소수의 자료가 비정상 자료인 경우 비정상 자료의 탐지를 위해 사용하는 방법이다. 하지만, 이 보고서에서도 기술했듯이 비정상 자료가 극소수가 아닐 수 있으며, 비정상의 개념을 문제에 따라 재정의할 수 있다. 보건사회 분야 자료의 이상 탐지 기법에 대한 탐색적 분석에서 살펴보았듯이, 정형 데이터뿐만 아니라 이미지 자료, 영상 자료의 비정형 데이터도 이상치를 탐지하는 데 정확도를 높일 수 있다. 또한, 여러 이상 탐지 기법을 사용하여 대분류부터 소분류까지 단계적으로 구분 지어 활용할 수 있다.

기계학습(Machine Learning)에 기반한 이상 탐지 기법 연구는 효과적인 정책 수립 및 집행으로 공공·행정 부문에서 효율성 증대가 가능하다. 아동의 권익 증진을 위하여 장기 결석, 건강검진 미실시 정보 등의 빅데이터를 활용하여 학대 등 위기 아동을 조기 발굴할 수 있는 데이터 분석에서 기계학습에 기반을 둔 이상 탐지 기법을 적용할 수 있다. 아동 학대의 사례처럼 아동 1,000명 중 1~2명이 학대 경험이 있다면 분류(classification) 문제로 접근하기 힘들다. 이런 불균형 자료(imbalanced data)에서는 이상 탐지 기법이 하나의 대안이 될 수 있다.

보건사회 분야에서 이상 탐지 기법이 가장 잘 활용될 수 있는 부분은 부정 수급 탐지이다. 부정 수급은 정부에서 지원하는 복지 혜택이나 복지 시설, 의료기관 등의 보조금을 더 받기 위해 수급 자격을 속이거나 입소자를 늘리는 등의 부정한 방법으로 복지 예산을 낭비하는 사례를 의미(보건복지부, 2018. 10. 31.)하는데, 2장 선행 연구에서 살펴본 바와 같이

국내·외 관련 사례가 존재한다. 행정 효율성을 높이고 기존과는 다른 새로운 유형의 부정 수급을 탐지하기 위해서는 부정 수급의 주요 유형별로, 시간적 흐름에 따른 패턴을 파악하고, 다른 자료와의 연계를 통해 통합적으로 자료를 살펴볼 필요가 있다. 연계되는 자료가 많아질수록, 새로운 이상 탐지 방법론을 적용할수록 부정 수급 탐지 기법은 발전할 것이다.

이상 탐지 기법은 개인정보 보호 기법과도 밀접한 연관이 있다. 개인정보는 민감정보이기 때문에 개인정보 보호 기법 활용은 보건사회 분야에서 특히 중요하다. 개인정보 보호 기법은 특정한 개인이 드러나지 않도록 통계적으로 마스킹(masking) 등의 처리를 해 주어야 하는데, 특정한 개인을 찾을 수 있는 방법으로 이상 탐지 기법을 활용할 수 있다. 개인정보 보호를 위한 통계적 기법은 데이터 활용 수요가 증가할수록 다양한 방법들이 개발될 것이다. 개인정보 보호 기법에 대한 이상 탐지 기법은 통계적 마스킹 외에 개인정보 유출 시도를 예측하는 데도 도움을 줄 수 있다. 이렇듯, 개인정보와 관련된 이상 탐지 기법의 활용성은 증가할 것으로 생각된다. 이를 위해서는 이상 탐지 기법의 다양한 방법론을 보건사회 분야의 정책 목적에 맞게 개발해야 하며, 정책 집행에 직접적으로 활용한 사례를 만들어서 정책 입안자 및 연구자들에게 공유할 필요가 있다. 공유의 확산 속도를 증가시키기 위해서는 데이터 및 분석 방법의 공개가 필수적일 것이다.

빅데이터 시대에 데이터의 활용가치는 증대되는 만큼, 최신 기법인 기계학습 기법에 기반한 이상 탐지 기법을 정책 대상 발굴이나 예산 효율성 제고에 접목시켜 활용한다면 예측 가능한 맞춤형 복지에 한층 가까이 다가설 수 있을 것이다.

참고문헌

관계부처 합동. (2017. 11). 4차 산업혁명 대응계획.

보건복지부. (2018. 10. 31.). Retrieved from http://bokjiro.go.kr/wrsd/guide1.do

안전보건공단(2012). 작업환경측정 제도안내. 2012-교육미디어-979정경희. (2017). 노인학대 현황 및 정책과제. 보건복지포럼, , 39-49.

정경희, 오영희, 강은나, 김경래, 이윤경, 오미애, 황남희, 김세진, 인선희, 이석구, 홍송이. (2017). *2017 년도 노인실태조사*. 세종: 보건복지부· 한국보건사회연구원.

정재윤. (2017). Novelty Detection-Overvier. Retrieved from https://jayhey.github.io/novelty%20detection/2017/10/18/Novelty_detection_overview/ 2018. 11.28.

정재윤. (2017). 로컬 아웃라이어 팩터. Retrieved from https://jayhey.github.io/novelty%20detection/2017/11/10/Novelty_detection_LOF/) 2018.11.28. 인출

차경엽, 오창석. (2015). *부정위험 탐지를 위한 데이터마이닝 적용방안 연구*. 감사원 감사연구원.

한국정보화진흥원, 건강보험심사평가원. (2016). *환자안전 조기 이상감지 시스템 구축*.

한송원. (2018). 치매 환자 年 4만명 느는데…'성년 후견인제' 이용률은 바닥. Retrieved from http://news.tvchosun.com/site/data/html_dir/2018/09/25/2018092590095.html 2018 .9.25. 인출.

행정안전부. (2017). 공공분야 빅데이터 5편- 실업급여 부정수급 방지 Retrieved from https://www.mois.go.kr/video/bbs/type019/commonSelectBoardArticle.do%3Bjsessionid=cXTLeCYJeMZWKayvsDTviTEF.node30?bbsId=BBSMSTR_000000000255&nttId=57816 2018.11.28. 인출

Abraham, B., &Chuang, A. (1989). Outlier detection and time series modeling. *Technometrics, 31*(2), 241-248. doi:10.2307/1268821

Abraham, B., & Box, G. E. P. (1979). Bayesian analysis of some outlier problems in time series. *Biometrika, 66*(2), 229-236. doi: 10.2307/2335653

Agarwal, D. (2005). An empirical bayes approach to detect anomalies in dynamic multidimen-sional arrays. *In Proceedings of the 5th IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA,* 26-33.

Agarwal, D. (2006). Detecting anomalies in cross-classified streams: A bayesian approach. *Knowledge and Information Systems, 11*(1), 29-44.

Aggarwal, C. (2005). On abnormality detection in spuriously populated data streams. *Proceedings of the 2005 SIAM international conference on data mining* (pp. 80-91) Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611972757.8 Retrieved from https://doi.org/10.1137/1.9781611972757.8

Agovic, A., Banerjee, A., Ganguly, A. R., &Protopopescu, V.Anomaly detection in transportation corridors using manifold embedding. Paper presented at the *Proceedings of the 1st International Workshop on Knowledge Discovery from Sensor Data,* 435-455.

Agrawal, R., &Srikant, R.Mining sequential patterns. Paper presented at the *Data Engineering, 1995. Proceedings of the Eleventh International Conference On,* 3-14.

Alzheimers Disease Neroimaging Initiative. (2017). About Biomarkers. Retrive from http://adni.loni.usc.edu/study-design/#background-container 2018. 9. 2.

Albrecht, S., Busch, J., Kloppenburg, M., Metze, F., &Tavan, P. (2000).

Generalized radial basis function networks for classification and novelty detection: Self-organization of optimal bayesian decision. *Neural Networks, 13*(10), 1075-1093.

Aleskerov, E., Freisleben, B., &Rao, B.Cardwatch: A neural network based database mining system for credit card fraud detection. Paper presented at the *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997,* 220-226.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., &Yang, Y.Topic detection and tracking pilot study: Final report. Paper presented at the *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, , 1998* 194-218.

Anderson, D., Frivold, T., Tamaru, A., &Valdes, A. (1994). Next generation intrusion detection expert system (NIDES), software users manual.

Anderson, D., Lunt, T. F., Javitz, H., Tamaru, A., &Valdes, A. (1995). *Detecting unusual program behavior using the statistical component of the next-generation intrusion detection expert systems (NIDES)* SRI International. Computer Science Laboratory.

Ando, S.Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. Paper presented at the *Icdm,* 13-22.

Angiulli, F., &Pizzuti, C.Fast outlier detection in high dimensional spaces. Paper presented at the *European Conference on Principles of Data Mining and Knowledge Discovery,* 15-27.

Anscombe, F. J., &Guttman, I. (1960). Rejection of outliers. *Technometrics, 2*(2), 123-147. doi:10.2307/1266540

Arning, A., Agrawal, R., &Raghavan, P.A linear method for deviation

detection in large databases. Paper presented at the *Kdd, , 1141*(50) 972-981.

Augusteijn, M. F., &Folkert, B. A. (2002). Neural network classification and novelty detection. *International Journal of Remote Sensing, 23*(14), 2891-2902.

Baker, L. D., Hofmann, T., McCallum, A., &Yang, Y.A hierarchical probabilistic model for novelty detection in text. Paper presented at the *Proceedings of International Conference on Machine Learning,*

Barbará, D., Couto, J., Jajodia, S., &Wu, N. (2001a). ADAM: A testbed for exploring the use of data mining in intrusion detection. *ACM Sigmod Record, 30*(4), 15-24.

Barbara, D., Wu, N., &Jajodia, S. (2001b). Detecting novel network intrusions using bayes estimators. Paper presented at the *Proceedings of the 2001 SIAM International Conference on Data Mining,* 1-17.

Basic medical key. (2018). Retrive from https://basic medicalkey.com/dementia-and-its-treatment/ 2018. 9. 2.

Basu, S., &Meckesheimer, M. (2007). Automatic outlier detection for time series: An application to sensor data. *Knowledge and Information Systems, 11*(2), 137-154.

Basu, S., Bilenko, M., &Mooney, R. J.A probabilistic framework for semi-supervised clustering. Paper presented at the *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 59-68.

Bay, S. D., &Schwabacher, M.Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Paper presented at the *Proceedings of the Ninth ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, 29-38.

Beckman, R. J., &Cook, R. D. (1983). Outliers. *Technometrics, 25*(2), 119-149.

Bengio, Y., Courville, A., &Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798-1828.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM, 18*(9), 509-517.

Bianco, A. M., Garcia Ben, M., Martinez, E. J., &Yohai, V. J. (2001). Outlier detection in regression models with arima errors using robust estimates. *Journal of Forecasting, 20*(8), 565-579.

Bishop, C. M. (1994). Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal Processing, 141*(4), 217-222.

Blender, R., Fraedrich, K., &Lunkeit, F. (1997). Identification of cyclone-track regimes in the north atlantic. *Quarterly Journal of the Royal Meteorological Society, 123*(539), 727-741.

Bolton, R. J., &Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit Scoring and Credit Control VII,* , 235-255.

Boriah, S., Chandola, V., &Kumar, V.Similarity measures for categorical data: A comparative evaluation. Paper presented at the *Proceedings of the 2008 SIAM International Conference on Data Mining,* 243-254.

Box, G. E., &Tiao, G. C. (1968). A bayesian approach to some outlier problems. *Biometrika, 55*(1), 119-129.

Brause, R., Langsdorf, T., &Hepp, M.Neural data mining for credit card fraud detection. Paper presented at the *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference On,* 103-106.

Breunig, M. M., Kriegel, H., Ng, R. T., &Sander, J.Optics-of: Identifying local outliers. Paper presented at the *European Conference on Principles of Data Mining and Knowledge Discovery,* 262-270.

Breunig, M. M., Kriegel, H., Ng, R. T., &Sander, J.LOF: Identifying density-based local outliers. Paper presented at the *ACM Sigmod Record, , 29*(2) 93-104.

Brito, M. R., Chavez, E. L., Quiroz, A. J., &Yukich, J. E. (1997). Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics &Probability Letters, 35*(1), 33-42.

Byers, S., &Raftery, A. E. (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association, 93*(442), 577-584.

Candès, E. J., Li, X., Ma, Y., &Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM), 58*(3), 11.

Capelleveen, V. G., Poel, M., Mueller, R. M., Thornton, D., Hillegersberf, J. (2016). Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International Journal of Accounting Information Systems* 21 (2016) p. 23.

Chakrabarti, S., Sarawagi, S., &Dom, B.Mining surprising patterns using temporal description length. Paper presented at the *Vldb, 98* 606-617.

Chan, P. K., Mahoney, M. V., &Arshad, M. H. (2003). A machine

learning approach to anomaly detection.

Chandola, V., Banerjee, A., &Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR), 41*(3), 15.

Chandola, V., Boriah, S., &Kumar, V. (2008). Understanding categorical similarity measures for outlier detection. *Technical Report,*

Chaudhary, A., Szalay, A. S., &Moore, A. W.Very fast outlier detection in large multidimensional data sets. Paper presented at the *Dmkd,*

Chawla, S., &Sun, P. (2006). SLOM: A new measure for local spatial outliers. *Knowledge and Information Systems, 9*(4), 412-429.

Chen, D., Shao, X., Hu, B., &Su, Q. (2005). Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. *Analytical Sciences, 21*(2), 161-166.

Chiu, A. L., &Fu, A. W.Enhancements on local outlier detection. Paper presented at the *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International,* 298-307.

Chow, C.Parzen-window network intrusion detectors. Paper presented at the *Proceedings of the 16 Th International Conference on Pattern Recognition (ICPR'02) Volume 4-Volume 4,* 385-388.

Crook, P. A., Marsland, S., Hayes, G., &Nehmzow, U.A tale of two filters-on-line novelty detection. Paper presented at the *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference On, , 4* 3894-3899.

Crook, P., &Hayes, G.A robot implementation of a biologically inspired method for novelty detection. Paper presented at the *Proceedings of Towards Intelligent Mobile Robots Conference,*

Das, K., &Schneider, J.Detecting anomalous records in categorical

datasets. Paper presented at the *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 220-229.

Dasgupta, D., &Nino, F.A comparison of negative and positive selection algorithms in novel pattern detection. Paper presented at the *Systems, Man, and Cybernetics, 2000 IEEE International Conference On, , 1* 125-130.

Davy, M., &Godsill, S.Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. Paper presented at the *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference On, , 2* 1316.

De Stefano, C., Sansone, C., &Vento, M. (2000). To reject or not to reject: That is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30*(1), 84-94.

Denning, D. E. (1987). An intrusion-detection model. *IEEE Transactions on Software Engineering,* (2), 222-232.

Desforges, M. J., Jacob, P. J., &Cooper, J. E. (1998). Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 212*(8), 687-703.

Diehl, C. P., &Hampshire, J. B.Real-time object classification and novelty detection for collaborative video surveillance. Paper presented at the *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference On, , 3* 2620-2625.

Duda, R. O., Hart, P. E., &Stork, D. G. (2012). *Pattern classification*

John Wiley &Sons.

Dutta, H., Giannella, C., Borne, K., &Kargupta, H.Distributed top-k outlier detection from astronomy catalogs using the demac system. Paper presented at the *Proceedings of the 2007 SIAM International Conference on Data Mining,* 473-478.

Emamian, V., Kaveh, M., &Tewfik, A. H.Robust clustering of acoustic emission signals using the kohonen network. Paper presented at the *Icassp,* 3891-3894.

Endler, D.Intrusion detection. applying machine learning to solaris audit data. Paper presented at the *Computer Security Applications Conference, 1998. Proceedings. 14th Annual,* 268-279.

Ertöz, L., Steinbach, M., &Kumar, V. (2004). Finding topics in collections of documents: A shared nearest neighbor approach. *Clustering and information retrieval* (pp. 83-103) Springer.

Escalante, H. J.A comparison of outlier detection algorithms for machine learning. Paper presented at the *Proceedings of the International Conference on Communications in Computing,* 228-237.

Eskin, E.Anomaly detection over noisy data using learned probability distributions. Paper presented at the *In Proceedings of the International Conference on Machine Learning,*

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., &Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. *Applications of data mining in computer security* (pp. 77-101) Springer.

Eskin, E., Lee, W., &Stolfo, S. J.Modeling system calls for intrusion detection with dynamic window sizes. Paper presented at the *DARPA Information Survivability Conference &Exposition II,*

2001. DISCEX'01. Proceedings, , 1 165-175.

Eskin, E., Portnoy, L., &Stolfo, S.Intrusion detection with unlabeled data using clustering. Paper presented at the *Proceedings of ACM CSS Workshop on Data Mining Applied to Security,*

Ester, M., Kriegel, H., Sander, J., &Xu, X.A density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at the *Kdd, , 96*(34) 226-231.

Fan, W., Miller, M., Stolfo, S., Lee, W., &Chan, P. (2004). Using artificial anomalies to detect unknown and known network intrusions. *Knowledge and Information Systems, 6*(5), 507-527.

Fawcett, T., &Provost, F.Activity monitoring: Noticing interesting changes in behavior. Paper presented at the *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 53-62.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society.Series B (Methodological),* , 350-363.

Fujimaki, R., Yairi, T., &Machida, K.An approach to spacecraft anomaly detection problem using kernel feature space. Paper presented at the *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining,* 401-410.

G. G. Helmer, J. S. K. Wong, V. Honavar, &L. Miller.Intelligent agents for intrusion detection. Paper presented at the *1998 IEEE Information Technology Conference, Information Environment for the Future (Cat. no.98EX228),* 121-124. doi:10.1109/IT.1998.713396

Galeano, P., Peña, D., &Tsay, R. S. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the*

American Statistical Association, 101(474), 654-669.

Ghoting, A., Parthasarathy, S., &Otey, M. E. (2008). Fast mining of distance-based outliers in high-dimensional datasets. *Data Mining and Knowledge Discovery, 16*(3), 349-364.

Gibbons, R. D., Bhaumik, D. K., &Aryal, S. (2009). *Statistical methods for groundwater monitoring* John Wiley &Sons.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1-21.

Guha, S., Rastogi, R., &Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems, 25*(5), 345-366.

Günter, S., Schraudolph, N. N., &Vishwanathan, S. (2007). Fast iterative kernel principal component analysis. *Journal of Machine Learning Research, 8*(Aug), 1893-1918.

Guttormsson, S. E., Marks, R. J., El-Sharkawi, M. A., &Kerszenbaum, I. (1999). Elliptical novelty grouping for on-line short-turn detection of excited running rotors. *IEEE Transactions on Energy Conversion, 14*(1), 16-22.

Gwadera, R., Atallah, M., &Szpankowski, W.Markov models for identification of significant episodes. Paper presented at the *Proceedings of the 2005 SIAM International Conference on Data Mining,* 404-414.

Harris, T. (1993). Neural network in machine health monitoring. *Professional Engineering,* , 8.

Hartigan, J. A., &Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society.Series C (Applied Statistics), 28*(1), 100-108.

Hawkins, D. M. (1974). The detection of errors in multivariate data

using principal components. *Journal of the American Statistical Association, 69*(346), 340-344.

Hawkins, D. M. (1980). *Identification of outliers* Springer.

Hawkins, S., He, H., Williams, G., &Baxter, R.Outlier detection using replicator neural networks. Paper presented at the *International Conference on Data Warehousing and Knowledge Discovery,* 170-180.

Hazel, G. G. (2000). Multivariate gaussian MRF for multispectral scene segmentation and anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing, 38*(3), 1199-1211.

He, Z., Deng, S., &Xu, X.Outlier detection integrating semantic knowledge. Paper presented at the *International Conference on Web-Age Information Management,* 126-131.

He, Z., Deng, S., &Xu, X.An optimization model for outlier detection in categorical data. Paper presented at the *International Conference on Intelligent Computing,* 400-409.

He, Z., Deng, S., Xu, X., &Huang, J. Z.A fast greedy algorithm for outlier mining. Paper presented at the *Pacific-Asia Conference on Knowledge Discovery and Data Mining,* 567-576.

He, Z., Huang, J. Z., Xu, X., &Deng, S.Mining class outliers: Concepts, algorithms and applications. Paper presented at the *International Conference on Web-Age Information Management,* 589-599.

He, Z., Xu, X., &Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters, 24*(9-10), 1641-1650.

He, Z., Xu, X., Huang, J. Z., &Deng, S.A frequent pattern discovery method for outlier detection. Paper presented at the *International Conference on Web-Age Information Management,* 726-732.

Heller, K. A., Svore, K. M., Keromytis, A. D., &Stolfo, S. J.One class support vector machines for detecting anomalous windows registry accesses. Paper presented at the *Proc. of the Workshop on Data Mining for Computer Security, , 9*

Helman, P., &Bhangoo, J. (1997). A statistically based system for prioritizing information exploration under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 27*(4), 449-466.

Hickinbotham, S. J., &Austin, J.Novelty detection in airframe strain data. Paper presented at the *Pattern Recognition, 2000. Proceedings. 15th International Conference On, , 2* 536-539.

Hinton, G. E., &Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504-507.

Ho, L. L., Macey, C. J., &Hiller, R.A distributed and reliable platform for adaptive anomaly detection in ip networks. Paper presented at the *International Workshop on Distributed Systems: Operations and Management,* 33-46.

Hollier, G., &Austin, J.Novelty detection for strain-gauge degradation using maximally correlated components. Paper presented at the *Esann,* 257-262.

Hollmén, J., &Tresp, V.Call-based fraud detection in mobile communication networks using a hierarchical regime-switching model. Paper presented at the *Advances in Neural Information Processing Systems,* 889-895.

Horn, P. S., Feng, L., Li, Y., &Pesce, A. J. (2001). Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry, 47*(12), 2137-2145.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward

networks. *Neural Networks, 4*(2), 251-257.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics,* , 435-475.

Huber, P. J. (2011). Robust statistics. *International encyclopedia of statistical science* (pp. 1248-1251) Springer.

Idé, T., &Kashima, H.Eigenspace-based anomaly detection in computer systems. Paper presented at the *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 440-449.

Ihler, A., Hutchins, J., &Smyth, P.Adaptive event detection with time-varying poisson processes. Paper presented at the *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 207-216.

Ilgun, K., Kemmerer, R. A., &Porras, P. A. (1995). State transition analysis: A rule-based intrusion detection approach. *IEEE Transactions on Software Engineering, 21*(3), 181-199.

Ismo, K.Outlier detection using k-nearest neighbour graph. Paper presented at the *Null,* 430-433.

Jain, A. K., &Dubes, R. C. (1988). Algorithms for clustering data.

Janakiram, D., Reddy, V. A., &Kumar, A. P.Outlier detection in wireless sensor networks using bayesian belief networks. Paper presented at the *Communication System Software and Middleware, 2006. Comsware 2006. First International Conference On,* 1-6.

Javitz, H. S., &Valdes, A.The SRI IDES statistical anomaly detector. Paper presented at the *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium On,* 316-326.

Jiang, M., Tseng, S., &Su, C. (2001). Two-phase clustering process for

outliers detection. *Pattern Recognition Letters, 22*(6-7), 691-700.

Jin, W., Tung, A. K., &Han, J.Mining top-n local outliers in large databases. Paper presented at the *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 293-298.

Jin, J. (2018). Implementing Model-Agnosticism in Uber's Real-Time Anomaly Detection Platform. Retrive from http://eng.uber.com/anomaly-detection/ 2018. 6. 7.

Joachims, T.Training linear SVMs in linear time. Paper presented at the *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 217-226.

Jolliffe, I. (2011). Principal component analysis. *International encyclopedia of statistical science* (pp. 1094-1096) Springer.

Kadota, K., Tominaga, D., Akiyama, Y., &Takahashi, K. (2003). Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification. *Chem-Bio Informatics Journal, 3*(1), 30-45.

Karypis, G., &Kumar, V. (1998). Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed Computing, 48*(1), 96-129.

Kearns, M. J. (1990). *The computational complexity of machine learning* MIT press.

Keogh, E., Lonardi, S., &Ratanamahatana, C. A.Towards parameter-free data mining. Paper presented at the *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 206-215.

King, S. P., King, D. M., Astley, K., Tarassenko, L., Hayton, P., &Utete, S.The use of novelty detection techniques for monitoring

high-integrity plant. Paper presented at the *Control Applications, 2002. Proceedings of the 2002 International Conference On, , 1* 221-226.

Knorr, E. M., &Ng, R. T.A unified approach for mining outliers. Paper presented at the *Proceedings of the 1997 Conference of the Centre for Advanced Studies on Collaborative Research,* 11.

Knorr, E. M., &Ng, R. T.Algorithms for mining distancebased outliers in large datasets. Paper presented at the *Proceedings of the International Conference on very Large Data Bases,* 392-403.

Knorr, E. M., Ng, R. T., &Tucakov, V. (2000). Distance-based outliers: Algorithms and applications. *The VLDB Journal—The International Journal on very Large Data Bases, 8*(3-4), 237-253.

Kohonen, T. (1997). Self-organizing maps.

Kou, Y., Lu, C., &Chen, D.Spatial weighted outlier detection. Paper presented at the *Proceedings of the 2006 SIAM International Conference on Data Mining,* 614-618.

Kruegel, C., &Vigna, G.Anomaly detection of web-based attacks. Paper presented at the *Proceedings of the 10th ACM Conference on Computer and Communications Security,* 251-261.

Krügel, C., Toth, T., &Kirda, E.Service specific anomaly detection for network intrusion detection. Paper presented at the *Proceedings of the 2002 ACM Symposium on Applied Computing,* 201-208.

Kumar, V. (2005). Parallel and distributed computing for cybersecurity. *IEEE Distributed Systems Online, 6*(10)

Labib, K., &Vemuri, R. (2002). NSOM: A real-time network-based intrusion detection system using self-organizing maps. *Networks and Security,* , 1-6.

Lakhina, A., Crovella, M., &Diot, C.Mining anomalies using traffic

feature distributions. Paper presented at the *ACM SIGCOMM Computer Communication Review, 35*(4) 217-228.

Losangeles Department of Public social Services. (2018). Welfare Fraud Prevention and Investigations. (http://dpss.lacounty.gov/wps/portal/dpss/main/programs-and-services/welfare-fraud-prevention-and-investigation/ 2018.12.30. 인출)

Lauer, M.A mixture approach to novelty detection using training data with outliers. Paper presented at the *European Conference on Machine Learning,* 300-311.

Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., &Kavsek, B.Informal identification of outliers in medical data. Paper presented at the *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, , 1* 20-24.

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., &Srivastava, J.A comparative study of anomaly detection schemes in network intrusion detection. Paper presented at the *Proceedings of the 2003 SIAM International Conference on Data Mining,* 25-36.

Le, Q. V.Building high-level features using large scale unsupervised learning. Paper presented at the *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference On,* 8595-8598.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., &Jackel, L. D.Handwritten digit recognition with a back-propagation network. Paper presented at the *Advances in Neural Information Processing Systems,* 396-404.

Lee, W., &Stolfo, S. J.Data mining approaches for intrusion detection. Paper presented at the 79-93.

Lee, W., Stolfo, S. J., &Chan, P. K.Learning patterns from unix process

execution traces for intrusion detection. Paper presented at the *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management,* 50-56.

Lee, W., Stolfo, S. J., &Mok, K. W. (2000). Adaptive intrusion detection: A data mining approach. *Artificial Intelligence Review, 14*(6), 533-567.

Lee, W., &Xiang, D.Information-theoretic measures for anomaly detection. Paper presented at the *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium On,* 130-143.

Lin, J., Keogh, E., Fu, A., &Van Herle, H.Approximations to magic: Finding unusual medical time series. Paper presented at the *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium On,* 329-334.

Lin, S., &Brown, D. E. (2006). An outlier-based data association method for linking criminal incidents. *Decision Support Systems, 41*(3), 604-615.

Liu, F. T., Ting, K. M., &Zhou, Z.Isolation forest. Paper presented at the *2008 Eighth IEEE International Conference on Data Mining,* 413-422.

Liu, J., &Weng, C. (1991). Detection of outlying data in bioavailability/bioequivalence studies. *Statistics in Medicine, 10*(9), 1375-1389.

Lu, C., Chen, D., &Kou, Y.Algorithms for spatial outlier detection. Paper presented at the *Data Mining, 2003. ICDM 2003. Third IEEE International Conference On,* 597-600.

Lu, J. (2015). Anomaly Detection for Airbnb's Payment Platform.

Ma, J., &Perkins, S.Online novelty detection on temporal sequences. Paper presented at the *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data*

Mining, 613-618.

Ma, J., &Perkins, S.Time-series novelty detection using one-class support vector machines. Paper presented at the *Neural Networks, 2003. Proceedings of the International Joint Conference On, , 3* 1741-1745.

Maaten, L.Learning a parametric embedding by preserving local structure. Paper presented at the *Artificial Intelligence and Statistics,* 384-391.

Maaten, L. v. d., &Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(Nov), 2579-2605.

Macdonald, J. W., &Ghosh, D. (2006). COPA—cancer outlier profile analysis. *Bioinformatics, 22*(23), 2950-2951.

Mahoney, M. V., &Chan, P. K.Learning nonstationary models of normal network traffic for detecting novel attacks. Paper presented at the *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 376-385.

Mahoney, M. V., &Chan, P. K. (2003). Learning rules for anomaly detection of hostile network traffic.

Manson, G.Identifying damage sensitive, environment insensitive features for damage detection. Paper presented at the *Proceedings of the Third International Conference on Identification in Engineering Systems,* 187-197.

Manson, G., Pierce, G., &Worden, K.On the long-term stability of normal condition for damage detection in a composite panel. Paper presented at the *Key Engineering Materials, , 204*359-370.

Manson, G., Pierce, S. G., Worden, K., Monnier, T., Guy, P., &Atherton, K.Long-term stability of normal condition data for

novelty detection. Paper presented at the *Smart Structures and Materials 2000: Smart Structures and Integrated Systems, , 3985* 323-335.

Marceau, C. (2005). Characterizing the behavior of a program using multiple-length n-grams.

Marchette, D. J.A statistical method for profiling network traffic. Paper presented at the *Workshop on Intrusion Detection and Network Monitoring,* 119-128.

Mark B. (2017). Multi-dimensional Reduction and Visualisation with t-SNE. Retrieved from https://data scienceplus.com/multi-dimensional-reduction-and-visualisation-with-t-sne/ 2018 11. 30. 인출

Markou, M., &Singh, S. (2003a). Novelty detection: A review—part 1: Statistical approaches. *Signal Processing, 83*(12), 2481-2497.

Markou, M., &Singh, S. (2003b). Novelty detection: A review—part 2: Neural network based approaches. *Signal Processing, 83*(12), 2499-2521.

Marsland, S., Nehmzow, U., &Shapiro, J. (1999). A model of habituation applied to mobile robots. *Proceedings of Towards Intelligent Mobile Robots,*

Marsland, S., Nehmzow, U., &Shapiro, J. (2000a). Novelty detection for robot neotaxis. *arXiv Preprint Cs/0006005,*

Marsland, S., Nehmzow, U., &Shapiro, J. (2000b). A real-time novelty detector for a mobile robot. *arXiv Preprint Cs/0006006,*

McCallum, A., Nigam, K., &Ungar, L. H.Efficient clustering of high-dimensional data sets with application to reference matching. Paper presented at the *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining, 169-178.

Motulsky, H. (1995). Intuitive biostatistics: Choosing a statistical test, chapter-17.

Moya, M. M., Koch, M. W., &Hostetler, L. D. (1993). One-class classifier networks for target recognition applications. *NASA STI/Recon Technical Report N, 93*

Nair, V., &Hinton, G. E.Rectified linear units improve restricted boltzmann machines. Paper presented at the *Proceedings of the 27th International Conference on Machine Learning (ICML-10),* 807-814.

Nigam, K., &Ghani, R.Analyzing the effectiveness and applicability of co-training. Paper presented at the *Proceedings of the Ninth International Conference on Information and Knowledge Management,* 86-93.

Noble, C. C., &Cook, D. J.Graph-based anomaly detection. Paper presented at the *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 631-636.

Otey, M., Parthasarathy, S., Ghoting, A., Li, G., Narravula, S., &Panda, D.Towards nic-based intrusion detection. Paper presented at the *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 723-728.

Otey, M. E., Ghoting, A., &Parthasarathy, S. (2006). Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery, 12*(2-3), 203-228.

Palshikar, G. K.Distance-based outliers in sequences. Paper presented at the *International Conference on Distributed Computing and Internet Technology,* 547-552.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., &Faloutsos, C.Loci: Fast outlier detection using the local correlation integral. Paper presented at the *Data Engineering, 2003. Proceedings. 19th International Conference On,* 315-326.

Parra, L., Deco, G., &Miesbach, S. (1996a). Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation, 8*(2), 260-269.

Parra, L., Deco, G., &Miesbach, S. (1996b). Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation, 8*(2), 260-269.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*(3), 1065-1076.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., . . . Toga, A. W. (2010). Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology, 74*(3), 201-209.

Pires, A. M., &Santos-Pereira, C.Using clustering and robust estimators to detect outliers in multivariate data. Paper presented at the *Proceedings of the International Conference on Robust Statistics,*

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers, 10*(3), 61-74.

Pokrajac, D., Lazarevic, A., &Latecki, L. J.Incremental local outlier detection for data streams. Paper presented at the *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium On,* 504-515.

Porras, P. A., &Neumann, P. G.EMERALD: Event monitoring enabling

response to anomalous live disturbances. Paper presented at the *Proceedings of the 20th National Information Systems Security Conference,* 353-365.

Protopapas, P., Giammarco, J. M., Faccioli, L., Struble, M. F., Dave, R., &Alcock, C. (2006). Finding outlier light curves in catalogues of periodic variable stars. *Monthly Notices of the Royal Astronomical Society, 369*(2), 677-696.

Qin, M., &Hwang, K.Frequent episode rules for internet anomaly detection. Paper presented at the *Network Computing and Applications, 2004.(NCA 2004). Proceedings. Third IEEE International Symposium On,* 161-168.

Qu, D., Vetter, B. M., Wang, F., Narayan, R., Wu, S. F., Hou, Y. F., . . . Sargor, C.Statistical anomaly detection for link-state routing protocols. Paper presented at the *Network Protocols, 1998. Proceedings. Sixth International Conference On,* 62-70.

Ramadas, M., Ostermann, S., &Tjaden, B.Detecting anomalous network traffic with self-organizing maps. Paper presented at the *International Workshop on Recent Advances in Intrusion Detection,* 36-54.

Ramaswamy, S., Rastogi, R., &Shim, K.Efficient algorithms for mining outliers from large data sets. Paper presented at the *ACM Sigmod Record, , 29*(2) 427-438.

Ratsch, G., Mika, S., Scholkopf, B., &Muller, K. (2002). Constructing boosting algorithms from SVMs: An application to one-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(9), 1184-1199.

Ravi, S., &Rif, R. (2013). PET scans accurately identify amyloid deposition after traumatic brain injury. Retrieved from

https://www.2minutemedicine.com/pet-scans-accurately-identify-amyloid-deposition-after-traumatic-brain-injury/ 2018 .9.20.

Roberts, S. J. (1999). Novelty detection using extreme value statistics. *IEE Proceedings-Vision, Image and Signal Processing, 146*(3), 124-129.

Roberts, S. J.Extreme value statistics for novelty detection in biomedical signal processing. Paper presented at the *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. no. 476),* 166-172.

Roberts, S., &Tarassenko, L. (1994). A probabilistic resource allocating network for novelty detection. *Neural Computation, 6*(2), 270-284.

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics, 25*(2), 165-172.

Rousseeuw, P. J., &Leroy, A. M. (1987). Robust regression and outlier detection. John Wiley &Sons, Inc.

Roth, V.Outlier detection with one-class kernel fisher discriminants. Paper presented at the *Advances in Neural Information Processing Systems,* 1169-1176.

Roth, V. (2006). Kernel fisher discriminants for outlier detection. *Neural Computation, 18*(4), 942-960.

Salvador, S., Chan, P., &Brodie, J.Learning states and rules for time series anomaly detection. Paper presented at the *FLAIRS Conference,* 306-311.

Sarawagi, S., Agrawal, R., &Megiddo, N.Discovery-driven exploration of OLAP data cubes. Paper presented at the *International Conference on Extending Database Technology,* 168-182.

Saunders, R., &Gero, J. S.The importance of being emergent. Paper

presented at the *Proceedings of Artificial Intelligence in Design,*

Scarth, G., McIntyre, M., Wowk, B., &Somorjai, R.Detection of novelty in functional images using fuzzy clustering. Paper presented at the *Proceedings of the 3rd Meeting of International Society for Magnetic Resonance in Medicine,*

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., &Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*(7), 1443-1471.

Scott, S. L. (2001). Detecting network intrusion using a markov modulated nonhomogeneous poisson process. *Submitted to the Journal of the American Statistical Association,*

Sebyala, A. A., Olukemi, T., Sacks, L., &Sacks, D. L.Active platform security through intrusion detection using naive bayesian network for anomaly detection. Paper presented at the *London Communications Symposium,* 1-5.

SECU N CCTV News. (2018). Retrived from 보안 관제, 인공지능으로 효과적인 대응방안 구현해 나갈 것. http://www.cctvnews.co.kr/news/articleView.html?idxno=78306 2018 .6.7.

Sheikholeslami, G., Chatterjee, S., &Zhang, A.Wavecluster: A multi-resolution clustering approach for very large spatial databases. Paper presented at the *Vldb, , 98* 428-439.

Shekhar, S., Lu, C., &Zhang, P.Detecting graph-based spatial outliers: Algorithms and applications (a summary of results). Paper presented at the *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 371-376.

Shyu, M., Chen, S., Sarinnapakorn, K., &Chang, L. (2003). A novel anomaly detection scheme based on principal component

classifier.

Siaterlis, C., &Maglaris, B.Towards multisensor data fusion for DoS detection. Paper presented at the *Proceedings of the 2004 ACM Symposium on Applied Computing,* 439-446.

Singh, S., &Markou, M. (2004). An approach to novelty detection applied to the classification of image regions. *IEEE Transactions on Knowledge and Data Engineering, 16*(4), 396-407.

Smith, R., Bivens, A., Embrechts, M., Palagiri, C., &Szymanski, B. (2002). Clustering approaches for anomaly based intrusion detection. *Proceedings of Intelligent Engineering Systems through Artificial Neural Networks,* , 579-584.

Smyth, P. (1994). Markov monitoring with unknown states. *IEEE Journal on Selected Areas in Communications, 12*(9), 1600-1612.

Solberg, H. E., &Lahti, A. (2005). Detection of outliers in reference distributions: Performance of horn's algorithm. *Clinical Chemistry, 51*(12), 2326-2332.

Song, X., Wu, M., Jermaine, C., &Ranka, S. (2007). Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering, 19*(5), 631-645.

Soule, A., Salamatian, K., &Taft, N.Combining filtering and statistical methods for anomaly detection. Paper presented at the *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement,* 31.

Spence, C., Parra, L., &Sajda, P.Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. Paper presented at the *Mmbia,* 3.

Stefansky, W. (1972). Rejecting outliers in factorial designs. *Technometrics, 14*(2), 469-479.

Sun, H., Bao, Y., Zhao, F., Yu, G., &Wang, D.CD-trees: An efficient index structure for outlier detection. Paper presented at the *International Conference on Web-Age Information Management,* 600-609.

Sun, J., Qu, H., Chakrabarti, D., &Faloutsos, C.Neighborhood formation and anomaly detection in bipartite graphs. Paper presented at the *Data Mining, Fifth IEEE International Conference On,* 8 pp.

Sun, J., Qu, H., Chakrabarti, D., &Faloutsos, C.Neighborhood formation and anomaly detection in bipartite graphs. Paper presented at the *Data Mining, Fifth IEEE International Conference On,* 8 pp.

Sun, J., Xie, Y., Zhang, H., &Faloutsos, C.Less is more: Compact matrix decomposition for large sparse graphs. Paper presented at the *Proceedings of the 2007 SIAM International Conference on Data Mining,* 366-377.

Sun, P., &Chawla, S. (2004). *On local spatial outliers* IEEE.

Sun, P., Chawla, S., &Arunasalam, B.Mining for outliers in sequential databases. Paper presented at the *Proceedings of the 2006 SIAM International Conference on Data Mining,* 94-105.

Surace, C., &Worden, K.A novelty detection method to diagnose damage in structures: An application to an offshore platform. Paper presented at the *The Eighth International Offshore and Polar Engineering Conference,*

Surace, C., Worden, K., &Tomlinson, G.A novelty detection approach to diagnose damage in a cracked beam. Paper presented at the *Proceedings-SPIE the International Society for Optical Engineering,* 947-953.

Symantec. (2018). Symantec Anomaly Detection for Automotive. https://www.symantec.com/content/dam/symantec/docs/data-sheets/anomaly-detection-for-automotive-en.pdf. (2018. 6. 7. 인출).

Tan, P., Steinbach, M., &Kumar, V. (2005). Introduction to data mining. 1st.

Tang, J., Chen, Z., Fu, A. W., &Cheung, D. W.Enhancing effectiveness of outlier detections for low density patterns. Paper presented at the *Pacific-Asia Conference on Knowledge Discovery and Data Mining,* 535-548.

Tao, Y., Xiao, X., &Zhou, S.Mining distance-based outliers from large databases in any metric space. Paper presented at the *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 394-403.

Tarassenko, L., Hayton, P., Cerneaz, N., &Brady, M. (1995). Novelty detection for the identification of masses in mammograms.

Teng, H. S., Chen, K., &Lu, S. C.Adaptive real-time anomaly detection using inductively generated sequential patterns. Paper presented at the *Research in Security and Privacy, 1990. Proceedings., 1990 IEEE Computer Society Symposium On,* 278-284.

Theiler, J. P., &Cai, D. M.Resampling approach for anomaly detection in multispectral images. Paper presented at the *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery IX, , 5093* 230-241.

Thottan, M., &Ji, C. (2003). Anomaly detection in IP networks. *IEEE Transactions on Signal Processing, 51*(8), 2191-2204.

Tibshirani, R., &Hastie, T. (2006). Outlier sums for differential gene expression analysis. *Biostatistics, 8*(1), 2-8.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra,

R., Sun, X., . . . Kuefer, R. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science, 310*(5748), 644-648.

Torr, P., &Murray, D. W. (1993). Outlier detection and motion segmentation. sensor fusion VI volume: 2059, pages: 432-44. robotics research gorup, department of engineering science, university of oxford parks road.

U.S. Centers for Disease Control. (2018). U.S. burden of Alzheimer's disease, related dementias to double by 2060. CDC Newsroom.

Tsay, R. S., Peña, D., &Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika, 87*(4), 789-804.

Vaidya, J., &Clifton, C.Privacy-preserving outlier detection. Paper presented at the *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference On,* 233-240.

Valdes, A., &Skinner, K.Adaptive, model-based monitoring for cyber attack detection. Paper presented at the *International Workshop on Recent Advances in Intrusion Detection,* 80-93.

Vapnik, V. (1995). The nature of statistical learning theory springer new york google scholar.

Vilalta, R., &Ma, S.Predicting rare events in temporal domains. Paper presented at the *Null,* 474.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., &Manzagol, P. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research, 11*(Dec), 3371-3408.

Vinueza, A., &Grudic, G. (2004). Unsupervised outlier detection and semi-supervised learning.

Wei, L., Qian, W., Zhou, A., Jin, W., &Jeffrey, X. Y. (2003). Hot: Hypergraph-based outlier test for categorical data. Paper presented at the *Pacific-Asia Conference on Knowledge Discovery and Data Mining,* 399-410.

Weigend, A. S., Mangeas, M., &Srivastava, A. N. (1995). Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting. *International Journal of Neural Systems, 6*(04), 373-399.

Weiss, G. M., &Hirsh, H.Learning to predict rare events in event sequences. Paper presented at the *Kdd,* 359-363.

Wikipedia.Alzheimer's disease. Retrieved from https://en.wikipedia.org/wiki/Alzheimer%27s_disease 2018 .9.2. 인출.

Wikipedia.Autoencoder. Retrieved from https://en.wikipedia.org/wiki/Autoencoder 2018 .9.1. 인출.

Williams, G., Baxter, R., He, H., Hawkins, S., &Gu, L.A comparative study of RNN for outlier detection in data mining. Paper presented at the *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference On,* 709-712.

Wong, W., Moore, A. W., Cooper, G. F., &Wagner, M. M.Bayesian network anomaly pattern detection for disease outbreaks. Paper presented at the *Proceedings of the 20th International Conference on Machine Learning (ICML-03),* 808-815.

Wong, W., Moore, A., Cooper, G., &Wagner, M.Rule-based anomaly pattern detection for detecting disease outbreaks. Paper presented at the *Aaai/Iaai,* 217-223.

Wu, M., &Jermaine, C.Outlier detection by sampling with accuracy guarantees. Paper presented at the *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining, 767-772.

Wu, N., &Zhang, J.Factor analysis based anomaly detection. Paper presented at the *Information Assurance Workshop, 2003. IEEE Systems, Man and Cybernetics Society,* 108-115.

Xie, J., Girshick, R., &Farhadi, A.Unsupervised deep embedding for clustering analysis. Paper presented at the *International Conference on Machine Learning,* 478-487.

Yairi, T., Kato, Y., &Hori, K.Fault detection by mining association rules from house-keeping data. Paper presented at the *Proceedings of the 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space, , 18* 21.

Yamanishi, K., &Takeuchi, J.Discovering outlier filtering rules from unlabeled data: Combining a supervised learner with an unsupervised learner. Paper presented at the *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 389-394.

Yamanishi, K., Takeuchi, J., Williams, G., &Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery, 8*(3), 275-300.

Yang, B., Fu, X., Sidiropoulos, N. D., &Hong, M. (2017). Towards k-means-friendly spaces: Simultaneous deep learning and clustering. *arXiv Preprint arXiv:1610.04794,*

Ye, N., &Chen, Q. (2001). An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International, 17*(2), 105-112.

Yi, B., Sidiropoulos, N. D., Johnson, T., Jagadish, H. V., Faloutsos, C.,

&Biliris, A.Online data mining for co-evolving time sequences. Paper presented at the *Data Engineering, 2000. Proceedings. 16th International Conference On,* 13-22.

Ypma, A., &Duin, R. P. (1997). Novelty detection using self-organizing maps. *Progress in Connectionist-Based Information Systems, 2*, 1322-1325.

Yu, D., Sheikholeslami, G., &Zhang, A. (2002). Findout: Finding outliers in very large datasets. *Knowledge and Information Systems, 4*(4), 387-412.

Yu, J. X., Qian, W., Lu, H., &Zhou, A. (2006). Finding centric local outliers in categorical/numerical spaces. *Knowledge and Information Systems, 9*(3), 309-338.

Zeevi, A. J., Meir, R., &Adler, R. J.Time series prediction using mixtures of experts. Paper presented at the *Advances in Neural Information Processing Systems,* 309-318.

Zhai, S., Cheng, Y., Lu, W., &Zhang, Z. (2016). Deep structured energy based models for anomaly detection. *arXiv Preprint arXiv:1605.07717,*

Zhang, J., &Wang, H. (2006). Detecting outlying subspaces for high-dimensional data: The new task, algorithms, and performance. *Knowledge and Information Systems, 10*(3), 333-355.

Zimmermann, J., &Mohay, G.Distributed intrusion detection in clusters based on non-interference. Paper presented at the *Proceedings of the 2006 Australasian Workshops on Grid Computing and E-Research-Volume 54,* 89-95.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., &Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection.

[참고]

노인실태조사 주요 결과 통계표[20]

항목			2014	2017
연령 구성	구성별	65-69	31.7	32.4
		70-74	27.1	24.9
		75-79	20.6	21.1
		80세 이상	20.6	21.7
	평균연령		73.9	74.1
교육수준		글자 모름	9.6	6.6
		글자 해독	20.9	17.7
		초등학교	32.0	34.1
		중학교	13.2	16.9
		고등학교	16.6	17.3
		전문대학 이상	7.8	7.5
가구 형태		독거	23.0	23.6
		부부	44.5	48.4
		자녀 동거	28.4	23.7
		기타	4.0	4.4
자녀 동거 희망율			19.1	15.2
노인 단독거주 이유		경제적 능력	2.6	1.8
		건강	1.8	1.1
		개인(부부)생활 향유	15.5	18.8
		기존 거주지 거주 희망	11.1	11.0
		자녀 가사지원·손자녀 양육부담	-	-
		자녀의 결혼	32.7	36.0
		자녀의 별거 희망	10.1	8.7
		자녀가 타 지역에 있어서	20.6	18.8

20) 보건복지부 보도자료(2018. 5. 24.)

항목			2014	2017
	자녀의 경제적 형편이 어려워서		3.5	2.2
	자녀가 모두 딸이어서		-	-
	자녀가 없어서		-	-
	기타		2.2	1.7
단독가구 생활에서 경험하는 어려움	없음		12.7	44.5
	아플 때 간호		25.6	19.0
	일상생활 혼자 처리		10.4	7.5
	경제적 불안감		25.8	17.3
	안전에 대한 불안감		3.7	1.5
	심리적 불안감, 외로움		21.7	10.3
사회적 관계망	생존자녀 유무		97.7	97.1
	생존손자녀 유무		90.8	91.3
	형제·자매 유무		82.8	84.7
	가까운 친인척 유무		53.1	46.2
	친한 친구·이웃 유무		62.7	57.1
	비동거자녀 (주 1회 이상)	왕래 비율	37.7	38.0
		연락 비율	72.9	81.0
기혼자녀와의 동거 이유	기혼자녀 동거 당연		15.6	14.8
	외로움		6.4	6.9
	가사노동 부담		-	-
	본인/배우자 수발 필요		15.4	15.9
	경제적 능력 부족		24.4	19.5
	자녀에게 가사지원·손자녀 양육도움 제공		21.8	27.3
	장애·질병이 있는 자녀 보호 위해		-	-
	자녀의 경제적 능력 부족		16.0	14.8
	기타		0.3	0.7
연간 가구 총소득			2,305만 원	2,589만 원
연간 개인 총소득(노인)1)			959만 원	1,176만 원
소득원별 구성(개인)	근로소득		12.7	13.3
	사업소득		15.1	13.6

항목		2014	2017
	재산소득	11.5	12.2
	사적이전소득	23.8	22.0
	공적이전소득	35.0	36.9
	사적연금소득	0.4	0.8
	기타소득	1.3	1.2
소득원별 금액(개인)	근로소득	122.3만 원	156.2만 원
	사업소득	145.0만 원	160.4만 원
	재산소득	110.6만 원	143.3만 원
	사적이전소득	228.7만 원	258.4만 원
	공적이전소득	335.5만 원	434.7만 원
	사적연금소득	4.3만 원	9.1만 원
	기타소득	12.9만 원	14.2만 원
소득원별 소지 비율 (개인)	근로소득	14.3	17.1
	사업소득	13.8	14.8
	재산소득	27.6	23.5
	사적이전소득	92.9	93.4
	공적연금소득	31.9	34.6
노인가구의 자산 보유 유무(규모)2)	부동산	89.1 (2억 1342만 원)	91.3 (2억 4546만 원)
	금융자산	85.2 (3,142.2만 원)	91.6 (3,631.8만 원)
노인가구의 부채 유무(규모)3)		33.5 (2,630만 원)	29.0 (2,408만 원)
부담되는 지출 항목(가구)	식비	16.2	18.7
	월세	5.1	5.5
	주거관리비	35.4	24.9
	보건의료비	23.1	23.1
	경조사비	15.2	4.4
	기타	5.1	14.3
	없음	-	9.3
경제활동 유무	현재 일을 하고 있다	28.9	30.9
	일을 한 경험은 있으나 현재는 하지 않는다	60.4	59.3
	평생 일을 하지 않았다	10.7	9.8

항목		2014	2017
종사 직종	농어업	36.4	32.9
	단순노무직	36.6	40.1
	판매종사자	6.3	5.6
	기능원	2.6	3.8
	서비스근로자	5.5	5.2
	사무직원	1.5	0.9
	조립원	4.8	7.5
	전문가	2.7	2.2
	고위임직원관리자	3.7	1.8
근로희망	지금 일을 하지 않으나, 하고 싶다	9.7	9.4
경제활동 참여 사유	생계비 마련	79.3	73.0
	용돈 마련	8.6	11.5
	건강 유지	3.1	6.0
	친교·사교	0.4	0.7
	시간 보내기	3.6	5.8
	능력 발휘	3.0	1.3
	경력 활용	1.8	1.6
	기타	0.2	0.2
만성질환	1개 이상 비율	89.2	89.5
	3개 이상 비율	46.2	51.0
흡연율		11.9	10.2
음주율		27.6	26.6
운동실천율		58.1	68.0
건강검진수진율		83.9	82.9
치매검진수진율		-	39.6
우울증상		33.1	21.1
자살	생각비율	10.9	6.7
	자살시도 응답비율	12.5	13.2
인지기능 저하 비율		31.5	14.5
기능상태	IADL만 제한	11.3	16.6
	IADL+ADL 제한	6.9	8.7
수발률		81.7	71.4
가족수발자		91.9	89.4
여가활동 유형별 비율	TV시청	82.4	99.3

항목		2014	2017
	산책	17.8	27.5
	스포츠 참여	10.2	16.6
	화초 텃밭가꾸기	10.1	12.0
	종교활동	8.3	10.7
노인여가복지시설이용률	경로당 이용률	25.9	23.0
	노인복지관 이용률	8.9	9.3
경로당 이용 이유4)	친목 도모	85.5	91.4
	식사서비스	6.6	57.2
	건강증진프로그램 이용	0.5	9.0
	취미여가프로그램 이용	4.4	8.3
노인복지관 이용 이유	친목도모	14.9	42.3
	취미여가프로그램 이용	53.2	49.6
	식사서비스	17.6	27.5
	건강증진프로그램 이용	3.8	26.4
사회활동 유형별 비율	평생교육참여율	13.7	12.9
	자원봉사참여율	4.5	3.9
희망하는 사회참여활동5)	자원봉사활동	6.4	6.5
	학습활동	13.0	11.8
	취미여가활동	61.6	66.4
	종교활동	45.6	50.2
	정치사회단체활동	0.7	1.0
	친목단체활동	40.1	45.2
주거 만족도		-	78.9
주거불만족 사유	식사, 빨래 등 일상생활을 하기 불편한 구조라서	-	12.8
	주방, 화장실, 욕실 등이 사용하기에 불편해서	-	25.1
	냉난방 등 편의시설이 갖추어지지 않아서	-	13.0
	방음이나 채광에 문제가 있어서	-	16.2
	안전관리, 보수 등 관리가 힘들어서	-	13.5
	개보수 등 주거관리 비용이 많이 들어서	-	17.5
	기타	-	1.8

항목		2014	2017
주거지 생활편리성	생활하기 불편한 구조	17.3	9.9
	생활하기 불편한 구조는 아니지만, 노인 배려 설비 없음	78.1	84.0
	노인 배려 설비를 갖추고 있음	4.6	6.1
희망 거주 형태 (건강 유지 시)	현재 집에서 계속 산다	-	88.6
	거주 환경이 더 좋은 집으로 이사한다	-	11.2
	식사, 생활편의 서비스 등이 제공되는 주택에 들어간다	-	0.2
희망 거주 형태 (거동 불편 시)	(재가 서비스를 받으며) 현재 살고 있는 집에서 계속 산다	-	57.6
	배우자, 자녀 또는 형제자매와 같이 산다	-	10.3
	돌봄, 식사, 생활편의 서비스 등이 제공되는 노인요양시설 등에 들어간다.	-	31.9
주요 기관시설과의 도보 이동시간	노인(종합)복지관(30분 이상)	65.1	55.9
	(종합)사회복지관, 장애인복지관, 여성회관 등(30분 이상)	70.4	62.1
안전사고 경험 비율		3.0	0.6
노인학대 경험 비율		9.9	9.8
운전 경험	현재 운전하고 있음	16.1	18.8
	하다가 그만둠(그만둔 나이)	8.9 (59.7세)	10.5 (62.1세)
운전시 어려움 경험		12.2	11.1
연명치료 반대		88.9	91.8
희망하는 장례법6)	화장	19.7	26.4
	화장 후 자연장	9.6	14.8
	화장 후 산골	34.4	30.3
	매장	22.9	17.5
	시신 기증	2.2	2.0
	기타	0.0	0.1
	아직 생각해보지 않았다	11.3	8.9
노인연령에 대한 인지	69세 이하	21.7	13.8
	70-74세	46.7	59.4
	75-79세	16.3	14.8
	80세 이상	15.3	12.1

항목		2014	2017
지하철 무임승차 동의	매우 동의	-	11.7
	동의	-	55.9
지하철 무임승차 연령 상향 조정		-	86.6
지하철 운임 일부 본인부담		-	67.1
선호하는 노후생활비 마련 방법	스스로	31.9	34.0
	본인과 자녀	6.9	10.2
	자녀	7.9	7.6
	국가적 차원(사회보장제도)	18.6	14.1
	본인과 사회보장제도	34.3	33.7
	기타	0.5	0.4

간행물회원제 안내

▶ 회원에 대한 특전

- 본 연구원이 발행하는 판매용 보고서는 물론 「보건복지포럼」, 「보건사회연구」도 무료로 받아보실 수 있으며 일반 서점에서 구입할 수 없는 비매용 간행물은 실비로 제공합니다.
- 가입기간 중 회비가 인상되는 경우라도 추가 부담이 없습니다.

▶ 회원종류

- 전체간행물회원 : 120,000원
- 보건분야 간행물회원 : 75,000원
- 사회분야 간행물회원 : 75,000원
- 정기간행물회원 : 35,000원

▶ 가입방법

- 홈페이지(www.kihasa.re.kr) - 발간자료 - 간행물구독안내

▶ 문의처

- (30147) 세종특별자치시 시청대로 370 세종국책연구단지 사회정책동 1~5F
  간행물 담당자 (Tel: 044-287-8157)

KIHASA 도서 판매처

- 한국경제서적(총판) 737-7498
- 교보문고(광화문점) 1544-1900
- 영풍문고(종로점) 399-5600
- 서울문고(종로점) 2198-2307
- Yes24 http://www.yes24.com
- 알라딘 http://www.aladdin.co.kr