ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 20-02-15
104

j.inst.Korean.electr.electron.eng.Vol.24,No.2,475~485,june 2020
http://dx.doi.org/10.7471/ikeee.2020.24.2.475

크라우드센싱 시스템에서 머신러닝을 이용한 이상데이터 탐지

Anomaly Data Detection Using Machine Learning in Crowdsensing System

김 미 희*★, 이 기 훈*

Mihui Kim*★, Gihun Lee*

Abstract

Recently, a crowdsensing system that provides a new sensing service with real-time sensing data provided from a user's device including a sensor without installing a separate sensor has attracted attention. In the crowdsensing system, meaningless data may be provided due to a user's operation error or communication problem, or false data may be provided to obtain compensation. Therefore, the detection and removal of the abnormal data determines the quality of the crowdsensing service. The proposed methods in the past to detect these anomalies are not efficient for the fast-changing environment of crowdsensing. This paper proposes an anomaly data detection method by extracting the characteristics of continuously and rapidly changing sensing data environment by using machine learning technology and modeling it with an appropriate algorithm. We show the performance and feasibility of the proposed system using deep learning binary classification model of supervised learning and autoencoder model of unsupervised learning.

요 약

최근, 별도의 센서를 설치하지 않고 센서가 포함된 사용자의 기기로부터 제공되는 실시간 센싱 데이터를 가지고 새로운 센싱 서비스를 제공하는 크라우드센싱(Crowdsensing) 시스템이 주목받고 있다. 크라우드센싱 시스템에서는 사용자의 조작실수나 통신 문제로 인해 의미 없는 데이터가 제공되거나 보상을 얻기 위해 거짓 데이터를 제공할 수 있어 해당 이상 데이터의 탐지 및 제거가 크라우드센싱 서비스의 질을 결정짓는다. 이러한 이상데이터를 탐지하기 위해 제안되었던 방법들은 크라우드센싱의 빠른 변화 환경에 효율적이지 않다. 본 논문은 머신러닝 기술을 활용하여 지속적이고 빠르게 변화하는 센싱 데이터의 특징을 추출하고 적절한 알고리즘을 통해 모델링하여 이상데이터를 탐지하는 방법을 제안한다. 지도학습의 딥러닝 이진분류 모델과 비지도학습의 오토인코더 모델을 사용하여 제안 시스템의 성능 및 실현 가능성을 보인다.

Key words : Crowdsensing, Machine Learning, AutoML, Autoencoder, Anomaly Data Detection

* School of Comp. Eng. & Applied Math., Computer System Institute, Hankyong National University
★ Corresponding author
E-mail : mhkim@hknu.ac.kr, Tel : +82-31-670-5167
※ This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2018R1A2B6009620).
Manuscript received Jun. 1, 2020; revised Jun. 23, 2020; accepted Jun. 23, 2020.

Ⅰ. 서론

크라우드센싱(Crowdsensing)은 개인에게 널리 보급된 스마트폰과 같은 기기를 통하여 제공되는 센싱 데이터를 수집하고 정보를 추출하여 센싱 서비스를 제공하는 시스템이다[1]. 2018년 기준 대한민국의 스마트폰 보급률은 96%를 넘는다고 한다. 이러한 보급된 기기를 활용하여 크라우드센싱 시스템에서는 별도의 센서 설치 없이 방대한 양의 데이터를 수집할 수 있다. 크라우드센싱을 기반으로 공유된 정보를 이용하여 센싱 서비스를 제공하기 위해서는 제공된 데이터의 신뢰성이 중요하다. 서비스 제공자는 정보 제공을 유도하기 위해 참여자들에게 보상을 제공할 수 있다. 그러나 악의적 정보 제공자는 단순히 보상을 목적으로 무작위 또는 거짓 데이터(malicious false data)를 생성하여 제공할 수 있다[2]. 악의적인 의도가 없더라도 통신 문제, 조작 실수 등으로부터 오류 데이터가 제공될 시 공유된 데이터에서 추출한 정보의 정확도를 떨어트려 크라우드센싱 시스템의 서비스 질 저하로 이어지게 된다.

이러한 이상데이터 탐지를 위해 기존 무선 센서 네트워크(Wireless Sensor Network, WSN)에서는 관측 데이터를 기준으로 신뢰구간 추정 탐지 기법[3], 주변 이웃의 데이터를 수집하여 정상과 비정상 라벨링을 통한 이상탐지 방법[4], 통계학적 정상 데이터 범위 설정을 통한 이상 탐지 방법[5] 등이 있다. 그러나 기존 WSN에서의 이상 데이터(anomaly data) 탐지 기법들은 직접 설치한 센서의 데이터를 기준으로 설계하고 실험하여 데이터 다양성을 고려해야 하는 크라우드센싱 환경에 직접 적용이 적합하지 않다. 이에 크라우드센싱 환경에서의 연구들도 진행되었는데, 크라우드센싱 시스템의 참여자 평판을 중심으로 데이터를 평가하는 방법[6], 블록체인 스마트 컨트랙트를 이용하여 정보 사용자가 직접 정보의 질을 평가하는 이상데이터 탐지 시스템[7]이 있다. 그러나 현재 크라우드센싱 기반 데이터로는 소셜미디어로부터 얻는 데이터, 위치데이터와 연관된 데이터, 실시간 모니터링과 같이 시간 흐름에 따른 데이터가 주를 이루고 있다. 크라우드센싱 환경의 데이터의 다양성, 데이터의 시간과 위치 관련성, 다양한 데이터의 특성을 고려해야 하는 경우, 특정 데이터에 특화된 알고리즘으로 데이터의 질을 평가하는 것은 적합하지 않을 수 있다. 또한 정직한 정보 제공자 및 사용자에 의한 평판을 기반으로 하는 시스템에서는 그들의 계정이 탈취당하여 악의적으로 이용될 수도 있다.

본 논문에서는 크라우드센싱 시스템에 제공되는 이상데이터의 탐지를 위하여 머신러닝 기법을 이용하여 이상 데이터(즉, 노이즈 및 오류 데이터, 무작위 또는 거짓 데이터)를 탐지하는 방법을 제안한다. 이를 위해 첫째, 머신러닝 기법을 이용하여 데이터의 중요 요소 추출 및 전처리를 수행한다. 둘째, 이렇게 추출된 요소의 데이터를 통해 지도학습과 비지도학습 모델링을 통해 적절한 모델을 선택한다. 이렇게 선택된 모델을 통해 이상탐지를 수행한다.

크라우드센싱은 주로 GPS센서를 이용한 위치데이터와 실시간 모니터링을 위해 시간 값과 함께 센싱 데이터를 수집하여 이용한다[1]. 광범위한 지역에서 실시간으로 공유되는 데이터에서 가짜 데이터(fake data)를 찾기 위해서는 정해진 알고리즘을 사용하는 것은 효율적이지 않다. 그래서 본 논문에서는 이상 데이터 탐지를 위하여 머신러닝을 이용하여 탐지 모델을 학습한다. 가장 많이 사용되는 머신러닝의 지도학습을 이용하기 위해서는 라벨링이 필요하다. 하지만 실생활에서 수집되는 데이터는 정리되어 있지 않으며 라벨링이 존재하지 않는다[8]. 또한 데이터 특성을 추출하여 머신러닝 모델 학습에 적용하여야 높은 예측율을 기대할 수 있다. 그래서 데이터의 특성을 찾는 것이 필수적이다. 본 논문에서는 이러한 작업을 위해 구글 클라우드 플랫폼에서 제공되는 자동화된 머신러닝 툴 AutoML[9]를 이용하여 자동으로 특성을 추출하고 그 특성을 이용하여 전처리한 데이터를 지도학습과 비지도학습을 통하여 이상 데이터 탐지를 비교하여 적절한 모델링 방법을 보인다.

2장에서는 본 논문의 기반이 되는 환경과 기술, 기존 방법을 소개한다. 3장에서는 제안하는 크라우드센싱 이상 데이터 탐지 방법을 설명한다. 4장에서는 실험성능을 분석하고, 5장에서는 결론을 맺는다.

Ⅱ. 기반 연구

본 장에서는 기반이 되는 크라우드센싱의 구조 및 서비스, 머신러닝, 지도학습, 비지도학습, 기존

방법을 소개하고 문제점을 설명한다.

2.1 크라우드센싱

크라우드센싱은 이미 널리 보급된 기기 스마트폰, 웨어러블 같은 기기를 활용해 데이터를 수집하고 정보를 추출하여 공유 및 서비스하는 시스템이다[1]. 크라우드센싱은 이미 보급된 기기들을 활용하기 때문에 추가적인 비용 없이 데이터를 수집할 수 있다. 서비스 제공자는 데이터를 수집하기 위해 참여자에게 보상을 지급할 수 있다. 보상은 재화, 포인트, 서비스 등이 될 수 있다[6].

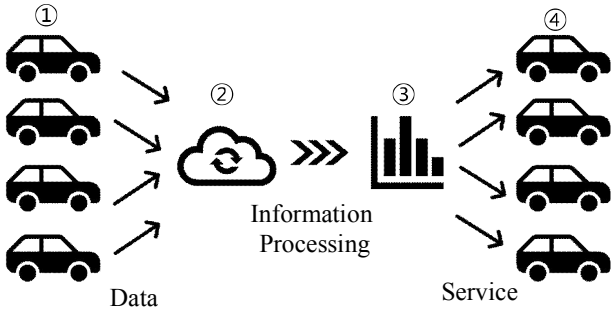

Fig. 1. Crowdsensing Navigation System.
그림 1. 크라우드센싱 내비게이션 시스템

크라우드센싱의 서비스 예로는 그림 1과 같은 내비게이션 서비스가 있다[10]. ① 사람들이 내비게이션 애플리케이션 서비스를 이용하며 이동할 때 스마트폰은 사용자의 위치를 서버에 전송한다. ② 서버는 수많은 내비게이션 사용자로부터 트래픽 상태 정보를 수집한다. ③ 수집한 정보를 바탕으로 도로 위 차량상황 정보를 추출한다. ④ 내비게이션 서비스에 추출한 정보를 표시하여 사용자들에게 보다 나은 서비스를 제공한다. 이 예시에서는 내비게이션 서비스가 정보 제공 참여자들에게 보상이 된다.

2.2 머신러닝

머신러닝은 주어진 데이터를 학습하여 원하는 결과를 도출하는 모델을 찾아내는 방법이다[11]. 모델을 찾아낸다는 것은 데이터간의 특성을 학습하여 특징 값을 찾고 학습된 모델을 사용하여 새로운 데이터를 입력하여 결과를 도출할 수 있음을 의미한다.

머신러닝의 단계는 1. 데이터 분석 및 이해, 2 .데이터 전처리, 3. 모델 수립, 4 .학습, 5. 성능평가로 구성된다[11]. 첫째, 데이터 분석 및 이해 과정은 머신러닝의 모델을 학습하기 위해서는 주어진 데이터 셋에 대하여 이해를 하는 것이다. 데이터를 이해하기 위해서 시각화, 다양한 변환 적용, 데이터 간의 상관관계 분석, 머신러닝 분류기 등을 사용하여 분석할 수 있다. 데이터 시각화는 많은 양의 데이터를 요약해서 보여주기 때문에 데이터의 구성 형태와 데이터의 변화를 파악할 수 있다. 이로부터 데이터의 특징을 파악, 가설을 구성할 수 있으며 데이터 자체의 문제점 파악 및 이해에 큰 도움이 된다.

둘째, 데이터 전처리란 원본 데이터를 머신러닝이 학습하기 좋은 형태로 변경하는 것이다. 실세계에서 얻은 데이터 셋은 복잡하고 지저분하다[8]. 통신 오류나 조작오류로 인해 비어있는 값이 생기거나 적절하지 않은 데이터가 발생할 수 있다. 이러한 데이터가 머신러닝 모델의 학습데이터로 사용되면 모델을 학습하는데 혼란이 오며 결과 값에 영향을 끼친다. 따라서 불필요한 데이터는 사전에 미리 제거해야 한다. 또한 데이터의 특성이 반영되어 있고 잡음이 없으며 편향되어 있지 않은 데이터로 학습을 하여야 적절한 머신러닝 모델을 만들 수 있다.

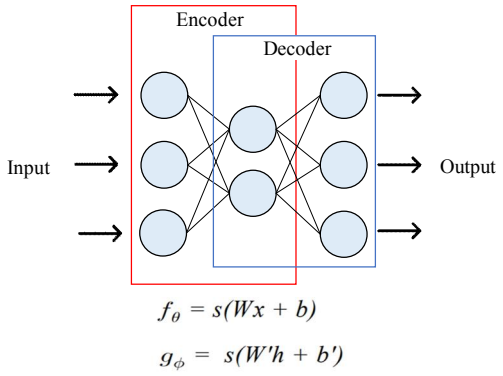

Fig. 2. Autoencoder Neural Network.
그림 2. 인공 신경망 구조

셋째, 모델 수립 과정에서는 데이터의 특성에 맞게 지도학습, 비지도학습, 강화 학습 중 효율적인 머신러닝 종류를 선택해야 한다. 지도학습은 데이터를 입력하여 발생하는 출력 값이 주어진 정답(라벨)과 일치하는 여부로 학습을 한다[12]. 대표적인 모델은 심층신경망(Deep Neurual Network)이 있다[13]. 비지도 학습은 입력만 존재하며 정답은 알려주지 않는 학습 방법이다. 비지도 학습은 데이터

의 특성을 군집화 하여 예측하는 모델을 학습하는 데 사용된다[12]. 대표적인 모델은 본 논문에서도 이용한 오토인코더(Autoencoder)가 있다. 오토인코더는 입력(Input)과 출력(Output)을 동일하게 맞추는 것이 목적인 신경망이다[14]. 그림 2는 오토인코더의 구성도를 보여준다. 입력층의 뉴런 수보다 적은 수의 뉴런을 은닉층에 배치한다. 이는 차원을 축소하는 의미를 가지고 있다. 다시 입력층과 동일한 노드의 수의 출력층에 배치하여 원본을 복원하도록 학습된다. 결과적으로 잡음을 제거하고 차원을 축소하는 효과를 갖게 된다. 그림 2의 인코더 식(f_θ)은 s는 활성 함수, W는 가중치(Weight), b는 바이어스(Bias)이다. 디코더 식(g_ϕ)은 s는 활성 함수, W'는 가중치, b'는 바이어스를 의미한다.

다섯째, 학습 과정은 앞서 만든 머신러닝 모델에 전 처리된 학습데이터를 입력하는 과정이다. 학습을 통하여 데이터 간 특징 값을 찾고 결과를 도출할 수 있다. 마지막으로 성능평가는 완성된 모델에 새로운 입력 데이터를 넣어 알맞은 결과 값을 확인하는 작업이다.

머신러닝 기술을 이용하는 시스템에서는 통상 앞서 소개한 다섯 단계를 거치가 되어 있으나, 데이터 분석을 통한 중요 요소 추출, 데이터 전처리의 완성도, 모델 수립 시 적용된 알고리즘과 파라미터 등에 따라 그 성능은 달라질 수 있다. 이에 본 논문에서는 해당 각 단계에서 최종 성능평가의 성능이 높아질 수 있도록 각 단계를 설계한다.

2.2 기존 방법

본 논문에서 가정하고 있는 크라우드센싱 환경과 유사한 환경으로 WSN 환경을 고려할 수 있다. WSN 환경에서 제안된 이상 데이터 탐지 기법으로는 이상 데이터 탐지를 위해 신뢰구간을 설정하는 방법[3]이 있다. 이 기법에서는 선박 내부에 여러 센서들이 설치되어 있다고 가정했으며 수집된 정보가 모여 전송될 때 지수평활법과 이동평균법을 이용하여 신뢰 구간을 설정한다. 측정된 데이터가 신뢰 구간을 벗어나면 이상 데이터로 감지한다. 또 다른 연구로서 이상 데이터 탐지 효율을 높이기 위해 센서로부터 얻은 데이터를 정상과 이상데이터로 라벨링 하는 방법이 제안되었다[4]. 수집한 데이터를 비슷한 값들끼리 묶는다. 그리고 구역을 설정한 뒤 정상 데이터와 이상 데이터 두 가지로 데이터를 바운더리 모델링(boundary modeling)을 한다. 통계적 방법을 사용하여 이상 데이터를 탐지하는 기법도 제안되었다[5]. 4분위 수를 이용하여 정상 데이터의 범위를 설정한다. 실험에서는 온도, 습도, 압력센서 데이터에 대해 4분위 수를 설정하고 오작동이나 통신 불량의 이상 데이터를 탐지한다.

하지만 [3]의 기법은 설치된 센서의 데이터의 값만을 기준으로 신뢰구간을 설정하였다. 이러한 방법은 다양한 기기의 센서로부터 얻어진 정보에 기반한 크라우드센싱 환경에서 적용하기 어렵다. [4]의 방법은 정상, 비정상 데이터로 라벨링을 위해 주변에 설치된 센서들과 값을 비교하여 라벨링을 진행하였다. 크라우드센싱 환경에서는 센서가 밀집되어 구축되어 있는 WSN 환경과 다르게 어디에서든 데이터가 제공될 수 있기 때문에 주변의 데이터와 비교하여 이상 데이터를 탐지하기에 적합하지 않을 수 있다. [5]의 통계학적 방법으로 4분위 수를 이용하여 정상 데이터를 설정하는 것은 크라우드센싱 데이터의 특성인 빠른 변화에 적절하게 대응하기에 어려움이 있다.

크라우드센싱 환경에서 이상 데이터 탐지를 위해 제안된 기법으로서 보상을 목적으로 하는 악의적인 공격을 막기 위하여 정보 제공자의 과거의 평판, 규칙의 준수여부 등을 반영하여 데이터의 질을 평가하여 비례하는 보상을 주는 방법이 있다[6]. 하지만 크라우드센싱의 데이터는 소셜미디어로부터 얻는 데이터, 위치데이터와 연관된 데이터, 실시간 모니터링과 같이 시간 흐름에 따른 데이터가 주를 이루고 있다. 크라우드센싱에서 특정 데이터에 특화된 알고리즘으로 다양한 데이터의 질을 평가하는 것은 적합하지 않을 수 있다. 또한 정직한 사용자의 계정이 탈취 당하여 악의적으로 이용될 수도 있다. 또 다른 연구로서 평판기반한 보상 지급 방법[6]의 문제점을 보완하고자 정보의 질을 블록체인의 스마트 컨트랙트를 이용하여 정보 사용자가 직접 정보의 질을 평가하는 시스템이 제안되었다[7]. 하지만 사용자가 정보의 질을 평가하기 위해서는 해당 정보가 옳은지 아닌지를 미리 알고 있어야 한다. 사용자는 정보를 얻기 위해 서비스를 이용하는 것이기 때문에 받은 정보가 실제로 맞는 정보인지는 실시간으로 알기 어려워 실시간 평가할 수 없다.

본 논문에서는 기존의 WSN 환경에서 제안된 이상 데이터 탐지방법[3][4][5]와 크라우드센싱에서 사

용되는 방법[6][7]의 문제점을 보완하고자 크라우드센싱 환경에서의 빠른 데이터 변화에 맞춰 신속한 탐지가 가능하고 탐지 성능을 향상시키기 위해 머신러닝과 오토인코더를 이용한 이상데이터 탐지 방법을 제안한다.

III. 제안방법

본 장에서는 다양한 데이터를 사용하는 크라우드센싱 환경에서 이상 데이터 탐지하는 방법을 제안한다.

3.1 제안된 이상데이터 감지 방법

본 논문에서 제안하는 방법은 그림 3과 같이 크게 네 단계로 구성된다. 첫째, 자동화된 중요 특징 추출과 전처리, 둘째 이상 데이터 분류를 위한 머신러닝을 이용한 모델 학습이다. 각 단계를 설명하면, ① 크라우드센싱 사용자들은 크라우드센싱 시스템에서 요청한 데이터를 크라우드센싱 서버에 제공한다. ② 서버는 자동으로 중요 특징을 추출하여 데이터를 전처리한 후, 머신러닝 모델 학습 모듈로 추출된 중요 요소의 데이터를 전달한다. ③ 여러 모델링 기법으로 머신러닝 모델을 생성하고 평가를 통해 적절한 모델을 선정한다. ④ 전처리한 데이터로 학습된 머신러닝 모델은 이상 데이터 탐지를 위해 사용된다.

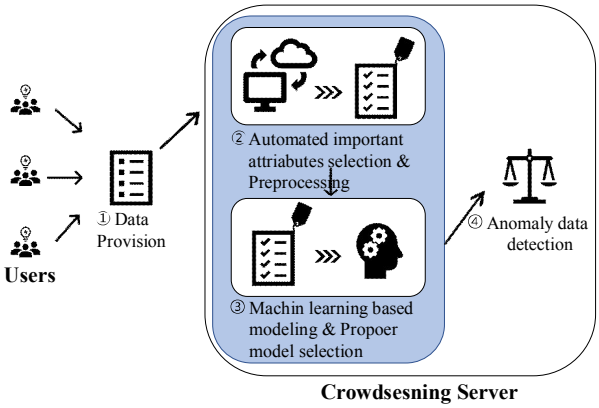

Fig. 3. Proposed Detection System.
그림 3. 제안된 탐지 기법

3.2 자동화된 중요 특징 추출 및 전처리

크라우드센싱 데이터를 머신러닝 모델에 효율적으로 학습시키기 위해서는 중요 특징 추출이 필수적이다. 하지만 크라우드센싱 데이터는 다양하고, 이러한 데이터에 대한 분석, 그리고 머신러닝에 적합한 데이터에 관련한 지식도 필수적이다. 따라서 관리자의 지식과 경험에 따라 크게 달라질 수 있다[15]. 이러한 방법은 많은 시도와 금전적인 낭비가 발생한다. 그래서 본 논문은 자동화된 특징 분석을 위해 AutoML 툴 사용을 제안한다. AutoML은 자동화된 머신러닝을 의미한다. 본 논문은 구글에서 제공하는 클라우드 플랫폼을 이용한다. 데이터를 AutoML에 입력하여 분석하면 분류기준에 따라 머신러닝에 적합한 특징 중요도를 자동으로 분석해 준다. AutoML의 결과 값은 머신러닝 모델의 학습 데이터를 만드는데 사용된다. AutoML을 이용하기 전 데이터에 기본적인 이산화, 널 값 제거, 오류 값 제거 등 전처리를 시행한다. 그 후 AutoML에 데이터를 입력한다. AutoML이 추출한 특성 중요도를 기반으로 데이터를 다시 수정한다.

3.3 머신러닝 모델 학습

머신러닝 모델을 이용한 이상데이터 탐지는 빠르게 변화하는 데이터에 맞춰 학습하여 데이터의 질을 평가하는데 적합하다. 본 논문은 구현을 위하여 지도학습 모델로는 딥러닝 이진 분류 모델(Binary classification model)을 사용한다. 그리고 비지도학습 모델은 오토인코더(Autoencoder) 모델을 사용한다.

3.3.1 지도학습

지도 학습을 위해서는 크라우드센싱에서 얻은 데이터 셋을 라벨링하는 작업이 필요하다. 대부분의 데이터는 어떠한 데이터가 잘못된 이상 데이터인지 판단되어 있지 않다. 본 논문은 지도학습 학습 데이터를 만들기 위한 라벨링 과정에 이상치 탐지(Novelty detection) 방법을 사용한다. 이상치 탐지는 학습 데이터를 모두 정상적인 데이터로 가정하고 이를 정상과 비정상을 판별하는 기준으로 삼는다[16]. 비지도학습에서 자주 사용되는 이상치 탐지를 지도학습을 위한 라벨링 과정에 적용 가능하다. 실제 환경에서 수집되는 데이터들은 대부분 정상이며, 머신러닝의 학습 데이터로 이용하기 전에 전처리로 잡음을 제거한다. 따라서 정상 데이터만을 이용하여 정상, 비정상의 범주를 설정하여 이상치를 탐지하는 방법은 정상 데이터들 간 공통된 특징이 있다면 좋은 성능을 보일 수 있다. 우선 정상적

인 데이터로 라벨링한 데이터셋을 만든다. 그리고 AutoML을 통해 중요 특성을 추출한다. 추출된 특성을 중심으로 데이터를 전처리한다. 최종적으로 이 데이터가 학습 데이터로 사용된다. 머신러닝 모델 생성과정은 전처리된 데이터에 적합하게 구조를 만든다. 본 논문에서는 입력 정보를 바탕으로 해당 입력이 어느 클래스에 속하는지 분류한다. 이를 위해 오차전역파(Backpropagation)를 이용하여 학습하여 바이너리 분류 모델에 높은 성과를 보이는 이진 분류 딥러닝 모델을 사용한다.

그림 4는 딥러닝 이진 분류 모델의 구조이다. 여기서 입력 값은 전처리된 데이터의 속성 수를 의미한다. 출력 값은 클래스를 분류하는 0, 1로 구성되어 있다. 중간층인 은닉층의 개수와 각 층의 노드의 개수는 다양하게 설정할 수 있다.

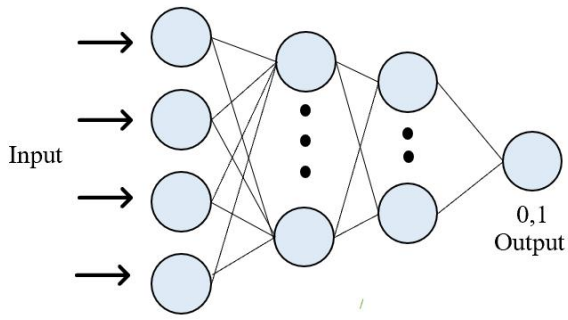

Fig. 4. Binary Classification Deep Neural Network.
그림 4 이진 분류 심층 신경망 구조

3.3.2 비지도학습

비지도학습은 번거로운 라벨링 과정을 생략할 수 있다는 장점이 있다. 본 논문에서는 오토인코더모델을 사용하여 이상데이터 탐지를 구현한다. 실생활에서 얻는 데이터는 대부분 정상 범주의 정보이다. 오토인코더는 정상 클래스의 데이터만으로 이상 데이터 감지와 아직 관측되지 않은 정상 데이터를 분류할 수 있다[14].

오토인코더에서 이상데이터를 탐지하기 위해서 정상적인 데이터를 이용하여 오토인코더 모델을 학습한다. 오토인코더는 데이터의 확률 분포, 즉 정상 클래스 데이터의 분포를 학습한다. 학습 후, 새로운 데이터를 넣어 입력 값과 출력 값의 오차의 크기로 정상 클래스 데이터와 비정상 클래스 데이터를 구분할 수 있다.

3.4 머신러닝 모델의 평가 및 선정

생성된 다양한 학습 모델을 데이터로 평가하여 해당 크라우드센싱 데이터에 적합한 모델을 선정하고 이를 탐지 모델로 사용한다.

Ⅳ. 성능평가

본 장에서는 실험을 통해 제안된 구조의 과정과 성능을 보인다.

4.1 실험 설계

본 논문에서 제안한 머신러닝을 이용한 이상데이터 탐지 방법의 성능을 평가하기 위해 본 논문은 Kaggle[17]에서 공개된 세 가지 유형의 데이터를 사용하였다: 택시 운행 데이터, 신용카드 사용 데이터, 구인구직 포스팅 데이터.

첫째, 뉴욕 도시의 택시 운행 데이터를 사용하였다. 택시 데이터는 이틀 동안의 9,883건의 정상 클래스 데이터를 포함한다. 비정상 클래스 데이터로 사용될 이상 데이터를 이상치 탐지(Novelty detection) 방법을 이용하여 정상 데이터 범위를 벗어나는 이상데이터를 약 10% 생성하였다. 제안된 구조대로 AutoML에서 중요 특성을 추출한다. 정상데이터는 0, 이상데이터는 1로 라벨링을 하였고, 이에 대한 분포는 그림 5와 같이 92%와 8%이다.

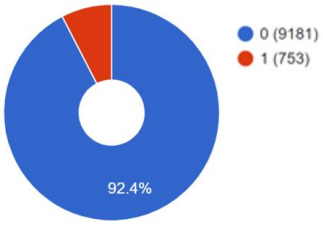

Fig. 5. Normal(0), Abnormal(1) data distribution.
그림 5. 정상(0)/비정상(1) 데이터 분포

그림 6은 AutoML을 이용하여 특성 중요도를 계산한 것이다. 특성 중요도를 중심으로 데이터의 속성 개수를 5개로 줄여 데이터를 간략화 하였다. 5개는 trip_duration, google_duration, google_distance, gc_distance, pickup_datetime이다. 각각은 실제 택시 탑승시간, 구글에서 제공하는 탑승 예상 시간, 예상 이동 거리, 실제 최단 거리, 탑승 시간 값을 담고 있다. 또한 머신러닝 학습 효율성을 향상시키

기 위한 전처리로서 타임스탬프 데이터는 0시~23시로 총 24가지의 카테고리로 이산화 작업을 진행하였다.

전체 데이터의 70%는 학습데이터로 30%는 검증 데이터로 분류하여 학습한다. 또한, 라벨링을 하지 않은 동일한 데이터를 가지고 오토인코더를 이용해 모델을 생성하였다. 마찬가지로 전체 데이터의 70%는 학습데이터로 30%는 검증 데이터로 사용하였다.

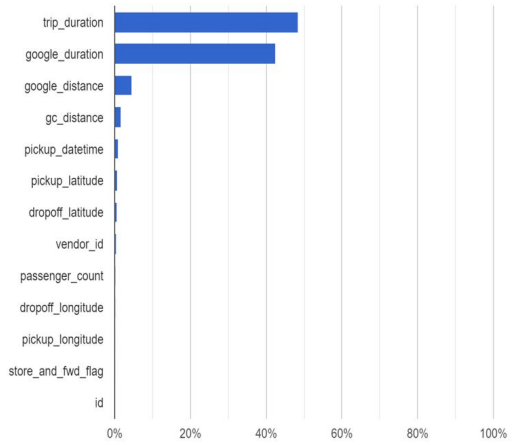

Fig. 6. Taxi data's important feature from AutoML.
그림 6. AutoML에서 계산된 택시 데이터의 특성 중요도

두 번째 실험 데이터는 2013년 9월에서 발생한 신용카드 사용에 대한 데이터 셋이다. 이 데이터는 284807개의 데이터 값, 32가지 속성으로 구성된다. 하지만 시간을 제외한 나머지는 V1~V30으로 표시되어 있다. 신용카드의 민감한 정보를 보호하기 위해서이다. 이러한 경우 관리자는 어떠한 속성이 중요한 데이터인지 확인할 수 없다. 그러므로 이러한 경우 AutoML을 사용하면 중요 특성 추출에 유용하다. 모든 실험조건은 첫 번째 데이터 셋을 이용한 실험과 동일하며 아래 그림 7은 AutoML을 이용하여 특성 중요도를 계산한 것이다. 특성 중요도에 따라 8개의 특성 즉, V14, V10, V12, V4, V7, V13, V3, V8의 데이터를 사용하여 모델링하였다.

세 번째 데이터는 17,838개의 구인구직 포스팅 글 중 정상데이터와 이상데이터를 모두 포함하는 데이터 셋이다. 총 18개의 속성을 가지고 있다. 그림 8은 AutoML을 이용한 특성 중요도이다. 이 중 description, company_profie, job_id, has_company_logo, requirements, required_education, benefits, location에 대한 데이터를 사용하여 모델링 하였다.

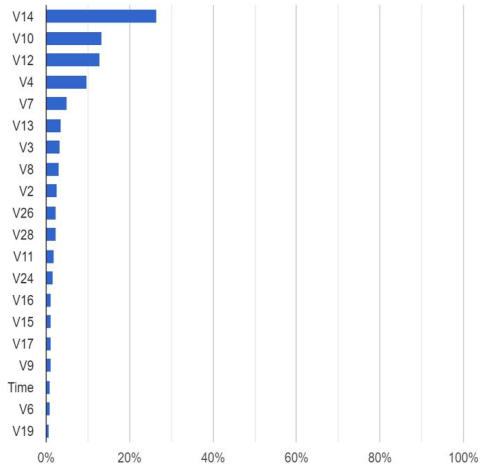

Fig. 7. Credit card data's important feature from AutoML.
그림 7. AutoML에서 계산된 신용카드 데이터의 특성 중요도

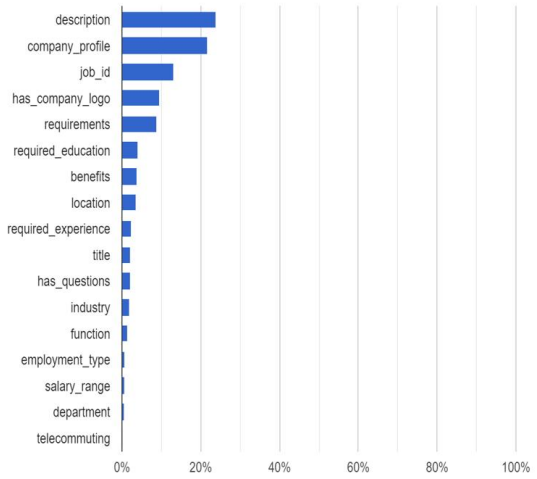

Fig. 8. Job posting's important feature from AutoML.
그림 8. AutoML에서 계산된 구인구직 데이터의 특성 중요도

4.2 결과 분석

그림 9는 첫 번째 데이터셋을 이용한 지도학습에서 데이터의 특성 중요도를 적용한 전, 후의 성능 차이를 나타낸다. 그림 9의 첫 번째 결과는 특성 중요도를 적용하지 않은 데이터 셋으로 학습 후에는 93.13%의 예측정확도를 보여준다. 반면에 그림 9의 두 번째 결과는 특성 중요도를 적용한 데이터 셋으로 98.33%로 데이터의 속성의 수가 줄었음에도 불구하고 오히려 높은 예측 정확도를 보여주고 있다.

머신 러닝 모델 평가는 실제 정답과 머신 러닝 모델의 답을 이용해 만들어 낼 수 있는 경우의 수를 이용한다. 가능한 경우는 표 1의 a(True Positives), b(False Negatives), c(False Positives), d(True Negatives)와 같다.

```
estimator = KerasClassifier(build_fn=create_baseline, epochs=10, batch_size=5, verbose=0)
kfold = StratifiedKFold(n_splits=10, shuffle=True)
results = cross_val_score(estimator, train_x, encoder_Y, cv=kfold)
print("Baseline: %.2f%% (%.2f%%)" % (results.mean()*100, results.std()*100))
```

```
Baseline: 93.13% (9.85%)
```

```
estimator = KerasClassifier(build_fn=create_baseline, epochs=10, batch_size=5, verbose=0)
kfold = StratifiedKFold(n_splits=10, shuffle=True)
results = cross_val_score(estimator, train_x, encoder_Y, cv=kfold)
print("Baseline: %.2f%% (%.2f%%)" % (results.mean()*100, results.std()*100))
```

```
Baseline: 98.33% (0.97%)
```

Fig. 9. Prediction rate comparison between original data and importanct feature data.
그림 9. 원본 데이터셋과 중요 특성 데이터셋의 모델링 예측정확도 비교

Table 1. Model evaluation metrics.
표 1. 모델 평가 측정값

		Prediction	
		True	False
Answer	True	a	b
	False	c	d

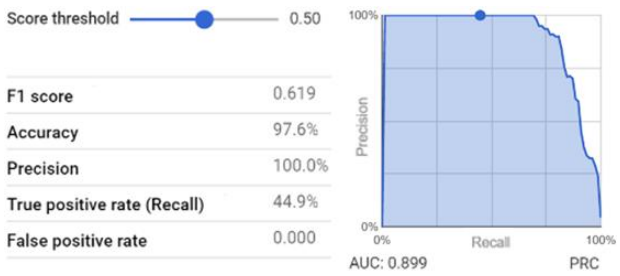

Fig. 10. First original data set's classification performance.
그림 10. 첫 번째 원본 데이터 셋의 분류 성능

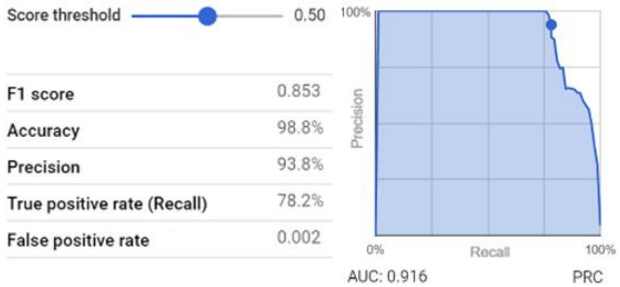

Fig. 11. First important feature data set's classification performance.
그림 11. 첫 번째 중요 특징 데이터 셋의 분류 성능

그림 10은 첫 번째 택시 데이터 실험에 대한 머신러닝 모델 평가 이다. 그림 10의 Precision은 머신러닝 모델이 True라고 평가한 것들 중 실제 True의 비율, Recall은 실제 정답 중에 모델이 맞춘 값의 비율, Accuracy는 전체 정답 중 몇 개를 맞췄는지, F1 Score는 Precision과 Recall을 적절하게 사용한 조화 평균 값이다. 그림 10 결과 값은 전처리 전, 그림 11 결과 값은 전처리 후이다. 전처리 전의 F1 score는 0.619, 전처리 후의 F1 score는 0.853로 중요 특성을 추출하여 데이터 셋을 중요 특성 위주로 변경하였을 때 더 높은 결과를 확인할 수 있다.

Fig. 12. Performance comparison between first data (1) Original and (2) Important feature.
그림 12. 첫 번째 데이터 셋((1)원본/(2)중요 특징)의 성능 비교

그림 12는 ROC(Receiver operating characteristic) curve와 AUC(Area under curve)를 이용한 분류모델 평가지표이다. ROC는 임계값(Threshold) 설정에 따른 True positive 와 False positive의 비율을 의미한다. 즉 임계값이 달라짐에 따라 분류모델의 성능 변화를 나타낸 곡선이다. AUC는 ROC 곡선 아랫부분의 넓이로 0~1 사이의 값을 갖고, 높을수록 통상적으로 좋은 모델을 의미한다. 중요 특징을 추출하여 데이터 셋을 변경하였을 때 기존의 0.899AUC보다 높은 0.916AUC를 확인할 수 있다.

그림 13, 14, 15는 두 번째 실험 데이터로 나타난 결과 값이다. F1 score는 0.763, 전처리 후의 F1 score는 0.833로 중요 특성을 추출하여 데이터 셋을 중요 특성 위주로 변경하였을 때 더 높은 결과를 확인할 수 있다. 중요 특징을 추출하여 데이터 셋을 변경하였을 때 기존의 AUC인 0.623보다 높은 0.768를

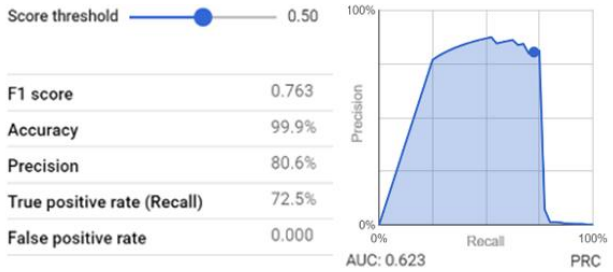

Fig. 13. Second original data set's classification performance.
그림 13. 두 번째 원본 데이터 셋의 분류 성능

확인할 수 있다. 두 번째 데이터셋 같은 경우는 결과 값의 차이가 크다. 이유는 중요하지 않은 데이터들이 학습 과정에서 많은 잡음을 발생했기 때문이다. 이러한 데이터에는 중요 특성 추출을 통하여 이를 제거했을 경우 높은 성능향상을 보인다.

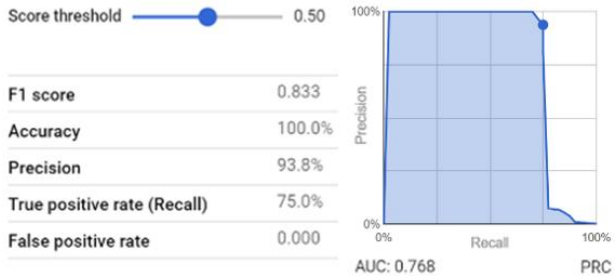

Fig. 14. Second important feature data set's classification performance.
그림 14. 두 번째 중요 특징 데이터 셋의 분류 성능

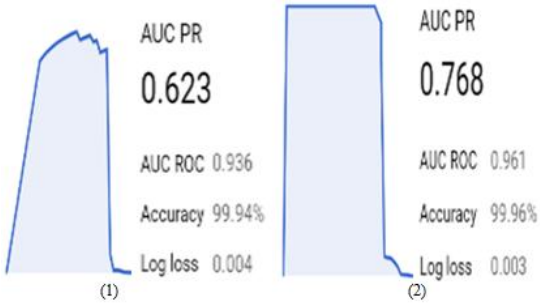

Fig. 15. Performance comparison between second data (1) Original and (2) Important feature.
그림 15. 두 번째 데이터 셋((1)원본/(2)중요 특징)의 성능 비교

그림 16, 17, 18은 세 번째 실험 데이터로 나타난 결과 값이다. F1 score는 0.958, 전처리 후의 F1 score는 0.992로 중요 특성을 추출하여 데이터 셋을 중요 특성 위주로 변경하였을 때 더 높은 결과를 확인할 수 있다. 중요 특징을 추출하여 데이터 셋을 변경하였을 때 기존의 AUC 0.987보다 높은

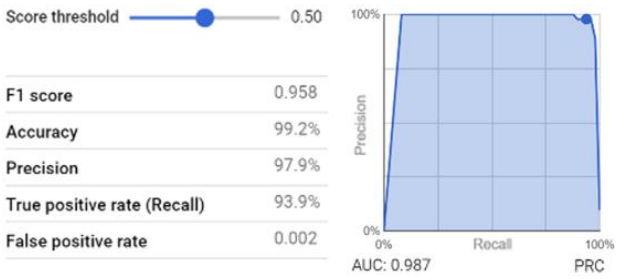

Fig. 16. Third original data set's classification performance.
그림 16. 세 번째 원본 데이터 셋의 분류 성능

0.993AUC를 확인할 수 있다.

세 가지 실험에서 자동 특징 추출을 기반으로 전처리한 데이터 값이 지도학습에 더 높은 성능을 가진 것을 확인할 수 있다.

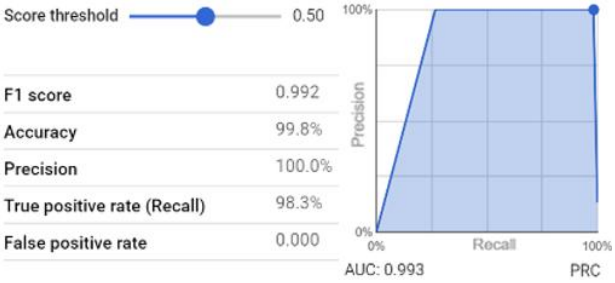

Fig. 17. Third important feature data set's classification performance.
그림 17. 세 번째 중요 특징 데이터 셋의 분류 성능

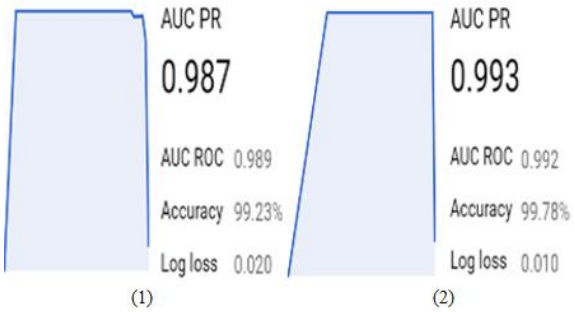

Fig. 18. Performance comparison between third data (1) Original and (2) Important feature.
그림 18. 세 번째 데이터 셋((1)원본/(2)중요 특징)의 성능 비교

그림 19는 세 가지 데이터셋을 이용하여 비지도 학습 방법을 적용한 오토인코더 모델의 AUC 값을

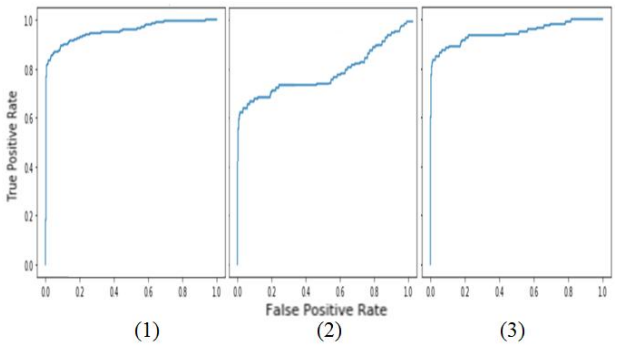

Fig. 19. Autoencoder AUC results (False positive rate vs. True positive rate): (1) First data, (2) Second data, and (3) Third data.
그림 19. 오토인코더 AUC 결과
(False positive rate vs. True positive rate)
(1) 첫 번째 (2) 두 번째 (3) 세 번째 데이터 셋

나타낸다. 첫 번째 데이터는 0.91 두 번째 데이터는 0.82 세 번째 데이터는 0.92 값을 갖는다. 지도학습과 비교하였을 때, 두 번째 데이터는 AUC값이 더 높았으며 첫 번째와 세 번째는 지도 학습이 더 높은 AUC값을 나타내었다. 이러한 차이가 나타나는 이유는 데이터 특성 때문이다. 두 번째 데이터 같은 경우는 데이터의 중요 특성이 명확히 클러스터링이 가능하여 지도보다 비지도에서 높은 AUC값을 나타낸다. 표 2에서는 모델별 데이터셋의 AUC 값을 비교하였다. 표 2에서는 모델별 데이터셋의 AUC 값을 비교하였다. 이러한 평가 결과에 따라 모델을 선택하여 실제 탐지 모델로 사용하면 된다.

Table 2. AUC result comparison of models.
표 2. 모델별 AUC 비교 표

		First Dataset	Second Dataset	Third Dataset
Supervised	Original	0.899	0.623	0.987
	Important Feature	0.916	0.768	0.993
Unsupervised	Autoencoder	0.91	0.82	0.92

Ⅴ. 결론

본 논문은 크라우드센싱 환경에서 이상 데이터 탐지를 위해 머신러닝 기술을 적용하는 방법을 제안하였다. 또한 자동 중요 특징 추출을 이용하여 관리자의 지식과 경험의 영향이 큰 영향을 끼치는 문제를 해결하는 방법을 제안했다. 본 연구에서 제안하는 방법은 현업에서 얻는 데이터 특성에 따라 지도학습과 비지도학습을 통해 모델링을 하고 평가 성능에 따라 선택하여 사용할 수 있다. 또한 크라우드센싱 데이터의 시간적, 공간적, 그리고 빠른 변화를 갖는 특징에 맞춰 기존 시스템보다 빠른 모델 업데이트에 도움이 될 것을 기대한다.

References

[1] R. Ganti, F. Ye and H. Lei, “Mobile Crowdsensing-Current State and Future Challenges,” *IEEE Communications Magazine*, vo.49, no.11, pp.32-39, Nov. 2011.

[2] B. Guo, Z. Yu, X. Zhou and D. Zhang, “From Participatory Sensing to Mobile Crowd Sensing,” in *Proc. of the 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pp.593-598, 2014.

[3] Y. J. Kim, Y. Y. He and J. K. Park, “Efficient Anomaly Detection Through Confidence Interval Estimation Based on Time Series Analysis,” *The Journal of Korean Institute of Communications and Information Sciences*, vo.39, no.1, pp.708-715, Aug. 2014.

[4] S. Suthaharan, C. Leckie, M. Moshtaghi and S. Karunasekera, “Sensor data boundary estimation for anomaly detection in wireless sensor networks,” in *Proc. of the IEEE International Conference on Mobile Ad-hoc and Sensor Systems (IEEE MASS)*, 2010, pp.546-551.

[5] A. Chirayil, R. Maharjan and C. Sehwu “Survey on Anomaly Detection in Wireless Sensor Networks (WSNs),” in *Proc. of the IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2019, pp.150-157.

[6] M. Musthag, A. Raij, D. Ganesan, S. Kumar and S. Shiffman, “Exploring micro-incentive strategies for participant compensation in high-burden studies,” in *Proc. of the Proceedings of the 13th international conference on Ubiquitous computing*, 2011, pp.435-444.

[7] D. Chatzopoulos, S. Gujar, B. Faltings and P. Hui, “Privacy Preserving and Cost Optimal Mobile Crowdsensing Using Smart Contracts on Blockchain,” in *Proc. of the 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2018, pp.442-450.

[8] S. H. Kwan, M. J. Ahnn and H. C. Lee, “Fault Detection and Classification of Process Cycle Signals using Density-based Clustering and Deep Learning,” *Korean Institute of Industrial Engineers*, vo.44, no.6, pp.475-482, Dec. 2018.

[9] A. Truong, A. Walters and J. Goodsitt, “Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools,” in *Proc. of the IEEE 31st International Conference*

on Tools with Artificial Intelligence (ICTAI), 2019, pp.1471-1479.
[10] L. Klopfenstein, S. Delpriori, P. Polidori and A. Sergiacomi, "Mobile crowdsensing for road sustainability: exploitability of publicly-sourced data," *International Review of Applied Economics*, vo.0, no.0, pp.1-22, Jul. 2019.
[11] L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys & Tutorials*, vo.18, no.0, pp.1153-1176, Oct. 2015.
[12] D. Bui, D. K. Nguyen and T. D. Ngo, "Supervising an Unsupervised Neural Network," in *Proc. of the First Asian Conference on Intelligent Information and Database Systems*, 2019, pp.307-312.
[13] Y. Sani, A. Mohamedou and K. Ali, "An overview of neural networks use in anomaly Intrusion Detection Systems," in *Proc. of the IEEE Student Conference on Research and Development (SCOReD)*, 2009, pp.89-92.
[14] B. K. Ko and J. G. Back, "Anomaly Detection With Variational Autoencoder To Prevent System Malfunctions," *Korean Institute of Industrial Engineers*, vo.0, no.6, pp.537-557, Nov. 2018.
[15] H. Cai, J. Lin, Y. Lin and Z. Liu, "AutoML for Architecting Efficient and Specialized Neural Networks," *IEEE Micro*, vo.40, no.1, pp.75-82, Jan. 2020.
[16] C. Wendl, D. Marcos and D. Tuia, "Novelty detection in very high resolution urban scenes with Density Forests", *Joint Urban Remote Sensing Event (JURSE)*, vo.0, no.0, pp.1-4, Aug. 2019.
[17] A. Goldbloom, "Kaggle", https://www.kaggle.com/

BIOGRAPHY

Mihui Kim (Member)

1997 : B.S. in Dept. of Computer Science and Engineering, Ewha Womans University, Korea
1999 : M.S. in Dept. of Computer Science and Engineering, Ewha Womans University, Korea
1999~2003 : Researcher, ETRI (Electronics and Telecommunication Research Institute), Korea
2007 : Ph.D in Dept. of Computer Science and Engineering, Ewha Womans University, Korea
2007~2009: Full Time Lecturer, Dept. of Computer Science and Engineering, Ewha Womans University, Korea
2009~2010 : Postdoctoral Researcher, Computer Science, North Carolina State University, USA
2011~Current : Professor, School of Computer Eng. & Applied Mathematics, Hankyong National University, Korea
Research interests : Security and efficient protocol design in IoT and crowdsensing system, Blockchain technologies

Lee Gihun (Member)

2013~2020 : B.S. in Dept. of Computer Science & Engineering, Hankyong National University, Korea
2020~Current : M.S. student, School of Computer Eng. & Applied Mathematics, Hankyong National University, Korea
Research interests : Security in IoT and crowdsensing system, Blockchain technologies