RESEARCH

KEPCO Journal on Electric Power and Energy
p-ISSN 2465-8111, e-ISSN 2466-0124
Volume 7, Number 2, December 2021, pp. 301-307
DOI 10.18770/KEPCO.2021.07.02.301

# A Research on the Energy Data Analysis using Machine Learning

# 머신러닝 기법을 활용한 에너지 데이터 분석에 관한 연구

Dongjoo Kim, Seongchul Kwon, Jonghui Moon, Gido Sim, Moonsung Bae
김동주, 권성철, 문종희, 심기도, 배문성

## Abstract

After the spread of the data collection devices such as smart meters, energy data is increasingly collected in a variety of ways, and its importance continues to grow. However, due to technical or practical limitations, errors such as missing or outliers in the data occur during data collection process. Especially in the case of customer-related data, billing problems may occur, so energy companies are conducting various research to process such data. In addition, efforts are being made to create added value from data, which makes it difficult to provide such services unless reliability of data is guaranteed. In order to solve these challenges, this research analyzes prior research related to bad data processing specifically in the energy field, and propose new missing value processing methods to improve the reliability and field utilization of energy data.

*Keywords: Bad Data Management, Machine Learning, Energy Big Data*

## I. Introduction

최근 들어 지능형 검침 시스템(AMI)의 핵심 장치인 스마트미터의 보급이 확산되고 있다. 미국의 에너지 정보국(US. Energy Information Agency)에 따르면 2017년까지 약 7,890만 개의 스마트미터가 설치되었으며, 그중 소비자 가구용은 6,950만 개로 전체 가구 수의 88%에 이른다 [1]. 영국에서도 스마트미터의 보급이 매년 증가하여 2017년도에는 약 550만 개의 스마트미터가 설치되었다 [2]. 우리나라에서도 2017년 말 기준 약 520만 호에 스마트미터가 보급되었으며, 2020년까지 2,250만 호에 스마트미터를 확대 보급하겠다는 제2차 지능형전력망 기본계획을 수립하였다 [3].

스마트미터의 보급이 확대됨에 따라 많은 양의 데이터가 매일 생성되고 누적될 것이므로, 스마트미터 데이터의 전략적 분석 및 전력 서비스에서의 활용을 위한 빅데이터 분석의 필요성이 더욱 증대되고 있다. 우리나라에서도 스마트미터 데이터의 중요성을 인지하여, 제2차 지능형전력망 기본계획에 전력 빅데이터 기반 신사업모델 활성화를 목표로 전력 빅데이터 플랫폼을 구축하는 계획을 포함하였다 [3].

스마트미터 빅데이터의 주요 요소는 데이터 취득, 데이터 분석, 응용서비스 활용의 세 가지이다[5]. 그중 본 연구에서 다루는 대상은 데이터 분석 및 응용서비스 활용 부분이다. 먼저 데이터 분석기법에는 통계적 분석, 머신러닝, 빅데이터 분석으로 일컬어지는 방법들이 포함되고, 구체적으로는 시계열 분석(Time Series Analysis), 분류법(Classification), 클러스터링(Clustering), 학습기반 기법(Learning-based Methods) 등이 있다. 응용서비스 활용 부문은 이해관계자의 유형을 구분하고 유형별 목표를 파악하는 것이

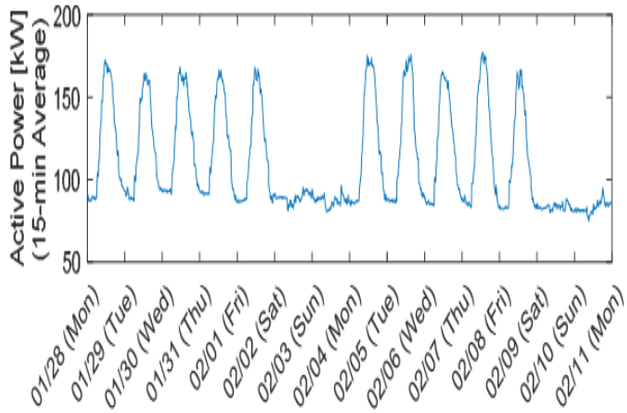


Fig. 1. 스마트미터 데이터 예시

**Article Information**

Manuscript Received Jul 12, 2021, Accepted September 27, 2021, Published online December 30, 2021

*The Authors are with KEPCO Research Institute, Korea Electric Power Corporation, 105 Munji-ro Yuseong-gu, Daejeon 34056, Republic of Korea.*

*Correspondence Author*: *Dongjoo Kim* (*djkim89@kepco.co.kr*)

*ORCID: 0000-0002-8924-257X (Dongjoo Kim); ORCID: 0000-0002-3961-5411 (Seongchul Kwon); ORCID: 0000-0002-2982-4009 (Jonghui Moon);*

*ORCID: 0000-0002-5685-043X (Gido Sim); ORCID: 0000-0043-4823-823X (*Moonsung Bae*)*

TABLE 1
스마트미터 데이터 관련 이해관계자 유형 및 응용서비스 [4]

| 이해관계자 유형 | 응용서비스 활용 |
|---|---|
| 소비자 (Consumer) | 개별적 수요예측 (HEMS의 입력정보로서 전력요금 절감에 활용, P2P 에너지 거래에 활용) |
| 판매자 (Retailer) | 수요예측 (수익증대를 위한 전력거래에 활용)<br>요금제 설계 (고객 유인에 활용)<br>소비자 특성분석 (서비스 품질 향상에 활용)<br>Abnormal Detection (데이터의 정확성 유지 및 electricity theft로 인해 손실 방지에 활용) |
| 중개자 (Aggregator) | 수요예측 및 수요반응 자원 평가 (수익 창출에 활용) |
| 운영자 (DSO) | 배전망 토폴러지 식별, 고장 식별 (효율적이고 안정적인 배전망 운영에 활용) |
| 데이터 서비스 제공자 (Data Service Provider) | 데이터 취득, 관리 및 분석 (다른 이해관계자에 데이터 서비스 편의 제공을 통한 수익 창출에 활용) |

필요하다. 아래의 표는 이해관계자를 5가지의 유형으로 구분하고, 유형별 목표에 따라 스마트미터 데이터를 활용할 수 있는 응용서비스를 정리한 것이다 [4].

따라서 본 논문에서는 에너지 데이터로 한정하여 오류 데이터를 처리하는 연구들에 대한 연구 목표, 사용 데이터의 유형, 데이터 분석기법 및 분석 결과를 분석하고, 스마트시티 연구과제를 통해 생성되는 데이터에 대한 분석 및 예측 알고리즘 개발에 대한 내용을 제안하고자 한다.

## II. 스마트미터 데이터 분석을 통한 오류 데이터 처리 방안

### A. 시계열 데이터 분석을 통한 누락데이터의 추정 [5]

#### 1) 연구 목표

다수의 스마트미터로부터 다량의 데이터가 통신망을 통해 전송되다 보면 누락된 데이터가 발생할 수 있다. 이는 스마트미터 데이터 분석 시에 문제를 발생시키게 되므로, 데이터 분석을 위한 전처리 과정으로서 누락된 데이터를 추정하여 완결된 데이터로 만들 필요가 있다.

#### 2) 사용 데이터

일정한 시간 간격으로 생성되는 시계열 데이터를 대상으로 한다. 검증용 데이터로는 Georgia Tech의 스마트미터 약 400개로부터 15분마다 생성되는 유효전력 데이터를 사용하였다. 해당 데이터의 일부분을 아래 그림에 나타내었다. 유효전력 사용량이 일정한 패턴을 보이는데, 학교라는 장소의 특성으로 인해 토요일과 일요일에는 매우 작은 값을 갖고 있음을 알 수 있다.

#### 3) 분석 방법

a) Linear Interpolation (LI)

Linear Interpolation은 누락 데이터의 앞쪽 및 뒤쪽 데이터로부터 선형(Linear) 관계식을 구성하여 누락된 데이터를 추정하는 것으로, 수식으로 표현하면 다음과 같다.

$$\hat{y}_i^{LI} = y_h + \frac{y_j - y_h}{x_j - x_h}(x - x_h),\ x_h < x_i < x_j \quad (1)$$

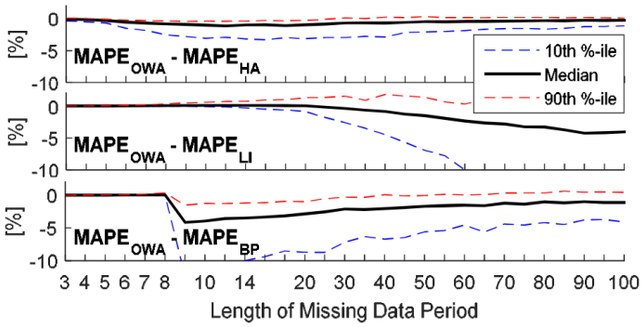


Fig. 2. 누락데이터의 추정 성능 비교

여기서 $\hat{y}_i^{LI}$는 누락된 데이터이고, $y_h$와 $y_j$는 각각 앞쪽 데이터와 뒤쪽 데이터를 의미한다. Linear Interpolation은 누락된 데이터의 시간 구간이 짧을 때 유용하게 사용할 수 있는 방법이다.

b) Historical Average Imputation (HA)

Historical Average Imputation은 누락된 데이터와 시간대가 동일한 데이터의 평균으로 누락 데이터를 추정하는 것으로, 수식으로 표현하면 다음과 같다.

$$\hat{y}_i^{HA} = \frac{1}{N_h}\sum_{j \in H} y_j \quad (2)$$

여기서 $\hat{y}_i^{HA}$는 누락된 데이터이고, $y_j$와 $N_h$는 각각 과거 데이터 및 그 개수를 의미한다. 본 연구사례에서는 앞뒤로 8일, 즉 16개의 과거 데이터를 사용하였다. 여기서 8일은 달력 상의 일자가 아니라 일정한 형태를 보이는 요일을 의미한다. 예를 들어, 누락 데이터가 월요일이면 앞뒤로 1일은 이전 주 월요일과 다음 주 월요일을 의미하는 것이다. Historical Average Imputation은 데이터가 일정한 주기로 유사한 형태를 가지는 경우에 긴 시간구간에서 통째로 누락된 데이터를 추정하는 용도로 사용할 수 있다.

c) Optimally Weighted Average Imputation (OWA)

LI와 HA의 가중평균으로 누락 데이터를 추정하는 것으로, 다음과 같이 수식으로 표현된다.

$$\hat{y}_i^{OWA} = w_i \hat{y}_i^{LI} + (1 - w_i)\hat{y}_i^{HA} \quad (3)$$

여기서 $w_i$는 두 방식으로 추정한 값에 대한 가중치를 의미한다. $w_i$의 값은 두 방식의 장단점을 고려하여, 누락된 시간 구간이 짧은 경우에는 LI에 대한 비중이 커지도록 하고, 누락된 시간 구간이 긴 경우에는 HA의 비중이 크게 할 수 있다. 따라서 $w_i$는 누락된 데이터의 시간 구간이 길어지면 값이 작아지는 지수함수 $e^{-a(\text{누락 시간구간})}$을 사용할 수 있다. 그러면 결국 지수함수의 파라미터 $a$를 적절하게 설정하는 것이 중요하게 되는데, 본 연구사례에서는 일종의 학습 방법으로서 과거 데이터로부터 가중치를 결정하는 방법을 제안하였다. 이는 일반적인 최적해 탐색 방식과 유사하게 누락된 시간 구간을 바꿔가면서 추정값과 실제값의 오차에 따라 $a$를 조정하여 최적해를 찾는 방법이다. 그러나 다양한 최적화 기법을 적용될 수 있을 뿐만 아니라, 단순히 $a$의 후보군에 대해 오차를 계산하여 가장 좋은 값을 선택하는 기본적인 방법을 사용해도 무방할 것으로 생각한다.

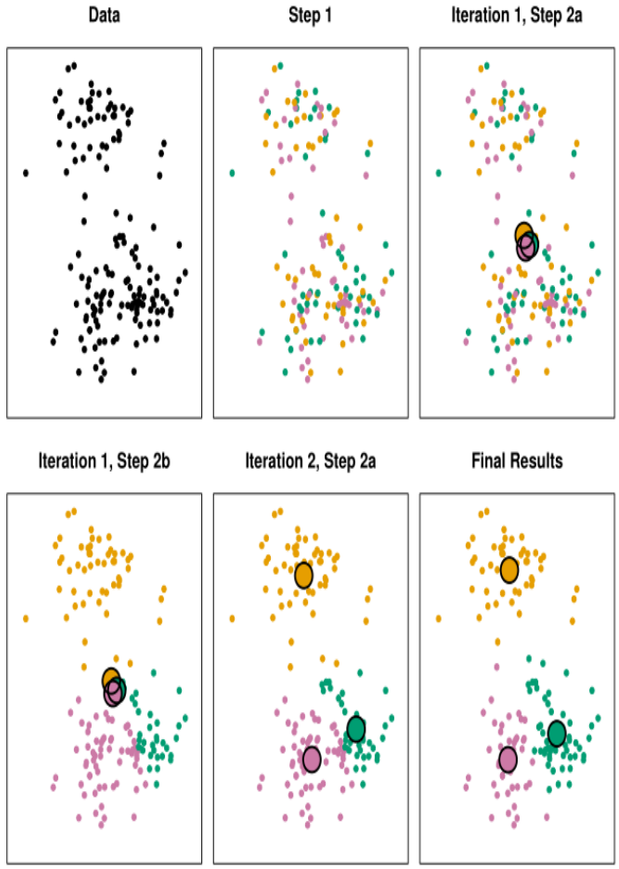


Fig. 3. K-means 클러스터링에서 클러스터가 나누어지는 과정 예시

4) 분석 결과

OWA과 LI, HA 및 업계의 Best Practice와 비교한 결과를 아래의 그림에 나타내었다. 3개(45분)에서 100개(25시간)까지의 데이터가 누락된 경우에 대하여 MAPE(Mean Absolute Percentage Error) 기준의 추정 성능을 비교하였고, OWA 방법이 다른 방법보다 더 좋음을 확인할 수 있다. 특히 LI와 비교한 결과를 보면 누락된 시간 구간이 길어질수록 LI 의 추정 성능이 매우 좋지 않음을 확인할 수 있다. HA가 전반적으로 좋은 성능을 보이는데, LI와 같은 단순한 방법과 조합함으로써 성능이 더 개선될 수 있다는 점을 눈여겨볼 만하다.

B. 클러스터링을 이용한 누락데이터의 추정 [6]

1) 연구 목표

시계열 데이터 분석을 통한 누락데이터의 추정과 마찬가지로 누락된 데이터를 추정하여 완결된 데이터로 만드는 것이다. K-means 클러스터링 방식에서 어떤 Distance 함수를 적용하는 것이 유용하고, 클러스터링을 위한 데이터 시간 구간을 얼마로 하면 좋을지에 대한 부가적인 평가를 수행하였다.

2) 사용 데이터

Irish Smart Metering Customer Behaviour Trials (CBT) 데이터 (2009년 7월 1일부터 2010년 12월 31일까지 18개월 동안 4,200여 개의 스마트미터에서 30분 간격으로 생성된 유효전력 데이터)를 사용하였다. 14일 동안의 데이터를 사용하였고, 그중 7일간의 데이터는 학습 데이터로서 K-means 클러스터링 방법에 사용하고, 나머지 7일 동안의 데이터는 테스트에 사용하였다.

3) 분석 방법

a) K-means 클러스터링

값 또는 패턴이 유사한 데이터를 묶는 방식으로서, k개의 데이터 집합(클러스터)를 만든 이후에 집합의 평균값인 클러스터 센터를 추정값으로 결정하는 방법이다. 데이터 간 유사 정도를 파악하기 위해 Distance를 계산하는데, 다음과 같이 일반적인 수식으로 표현할 수 있다.

$$J = \sum_{j=1}^{k} \sum_{i=1, i\in j}^{n} \left\| LP_i - CC_j \right\|^2 \tag{4}$$

여기서 $LP_i$는 데이터이고, $CC_j$는 클러스터 센터이다. 최종적으로 $J$를 최소화하는 클러스터 센터를 구할 때까지 군집화하는 데이터로 달라지고 그에 따라 클러스터 센터의 값도 달라진다. 아래의 Fig. 3 은 클러스터가 구해지는 과정을 보여주는 예시이다[8]. 처음에는 임의로 클러스터를 배정하기 때문에 클러스터 센터 간에 차이가 없지만, 과정이 반복될수록 Distance가 가까운 점들이 모이기 때문에 마지막에는 클러스터 센터가 서로 구분되어 데이터를 나눌 수 있게 된다. K-means 클러스터링은 반복적인 과정을 통해 해를 구하는 방법이기 때문에 초기값에 따라 결과가 달라지고, 그 결과가 Global Optimum이라고 단정지을 수 없다.

수식 (4)에서 ‖ ‖는 Distance 함수를 나타내고, 주로 사용하는 Euclidean Distance 이외에도 다음과 같이 여러 가지 방식을 사용할 수 있다.

Average Euclidean Distance:

$$d(LP_i, LP_j) = \left[ \frac{\sum_{t=1}^{T} (lp_i(t) - lp_j(t))^2}{T} \right]^{1/2} \tag{5}$$

Average Manhattan (city block) Distance:

$$d(LP_i, LP_j) = \frac{\sum_{t=1}^{T} |lp_i(t) - lp_j(t)|}{T} \tag{6}$$

Average Canberra Distance:

$$d(LP_i, LP_j) = \begin{cases} 0 & for\ \ lp_i(t) = lp_j(t) = 0 \\ \frac{\sum_{t=1}^{T} |lp_i(t) - lp_j(t)|}{|lp_i(t)| + |lp_j(t)|} & for\ \ lp_i(t) \neq 0\ \ or\ \ lp_j(t) \neq 0 \end{cases} \tag{7}$$

Average Pearson Correlation Distance:

$$d(LP_i, LP_j) = \frac{1 - \frac{\sum_{t=1}^{T} (lp_i(t) - \overline{LP_i})(lp_j(t) - \overline{LP_j})}{\left[\sum_{t=1}^{T} (lp_i(t) - \overline{LP_i})^2 \sum_{t=1}^{T} (lp_j(t) - \overline{LP_j})^2\right]^{1/2}}}{T} \tag{8}$$

b) K-means 클러스터링을 이용한 누락데이터 추정 구조

일단 학습(Training) 데이터로부터 클러스터 센터가 결정되면, 누락된 데이터는 가장 distance가 작은 클러스터 센터의 값을 선택함으로써 간단하게 추정값을 구할 수 있다. 아래의 Fig. 4는 K-means 클러스터링부터 누락 데이터 추정까지 과정을 도식화한 것이다.

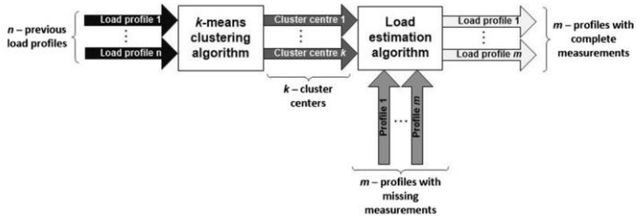


Fig. 4. 클러스터링을 이용한 누락 데이터 추정방식의 구조

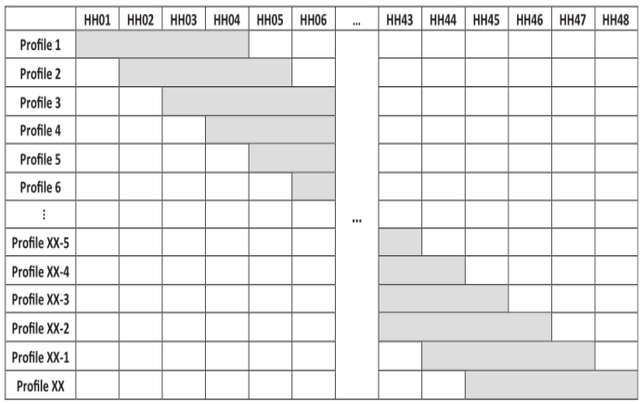


Fig. 5. Rolling Segmentation으로 변형된 데이터를 생성하는 방법

c) Rolling Segmentation 방식의 데이터 변형

본 연구사례에서는 24시간 동안의 데이터를 하나의 프로파일로 사용하는 것 외에도 시간 간격을 줄여 변형된 프로파일을 만들어 클러스터링하는 방법을 제안하였다. Fig. 5는 변형된 프로파일을 생성하는 Rolling Segmentation을 도식화한 것이다. 예를 들어, 첫 번째 Segmented 프로파일은 첫 번째 온전한 프로파일에서 생성하고, 두 번째 Segmented 프로파일은 두 번째 온전한 프로파일에서 시작 시점을 한 구간 이동하여 생성하는 방식이다. 전형적인 24시간 단위의 데이터에서 벗어나는 방법을 통해 성능개선을 할 수 있다는 예시로 참고할 수 있다.

d) 누락 데이터의 추정

누락 데이터가 프로파일의 가장 마지막에 있도록 하고, 가장 Distance가 작은 클러스터 센터를 찾아 누락 데이터를 추정한다. 누락된 데이터가 연속으로 있는 경우, 앞서 추정한 데이터를 정상 데이터로 간주하고 차례로 복구해 나간다. Fig. 6으로부터 이와 같은 추정 방법을 쉽게 이해할 수 있다.

4) 분석 결과

1시간에서 24시간까지의 데이터를 인위적으로 제거하여 시험하였다. Distance 함수에 따른 결과와 데이터 Segmentation 방식에 따른 결과를 각각 Fig. 7 및 Fig. 8에 나타냈었다. 우선 본 연구사례에서는 Canberra가 가장 좋은 결과를 보이는데, 이 결과를 모든 데이터에 적용하기는 어려울 것이다. 다만 일반적으로 사용하는 Euclidean 함수 이외에 다른 Distance 함수를 비교 평가할 필요가 있을 것이다. Fig. 8을 살펴보면 Segmentation 방식이 단순한 하루

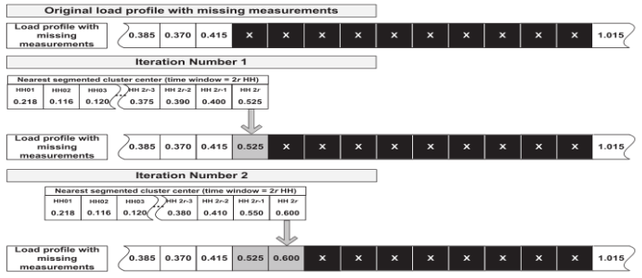


Fig. 6. 누락 데이터 추정 방법

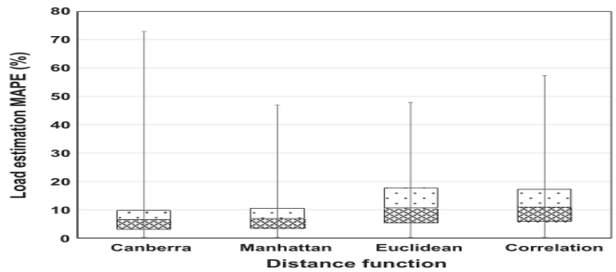


Fig. 7. Distance 함수 별 추정성능 비교

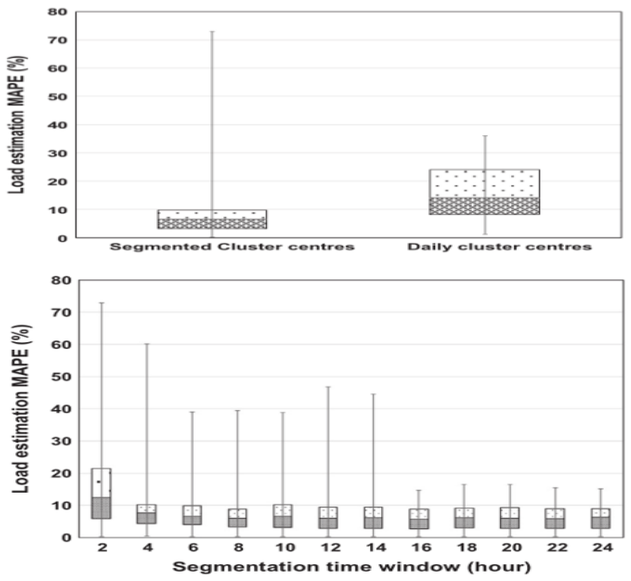


Fig. 8. Segmentation 방식 별 추정성능 비교

단위의 데이터보다 좋은 성능을 보인다. 따라서 정형화된 하루 단위 데이터에서 벗어나는 방식을 적용해보는 것이 좋은 선택이 될 수 있을 것이다. 본 연구사례에서는 16시간에서 24시간 구간으로 Segmentation 하는 것이 좋다는 결과를 도출하였는데, 이 결과도 일반적으로 적용되지는 않겠지만 일반적으로 예상할 수 있듯이 너무 작은 시간 구간의 Segmentation은 바람직하지 않음을 알 수 있다.

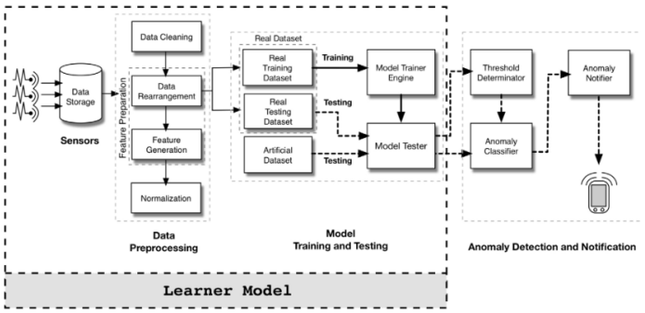


Fig. 9. 적용된 이상 데이터 검출 방법의 구조

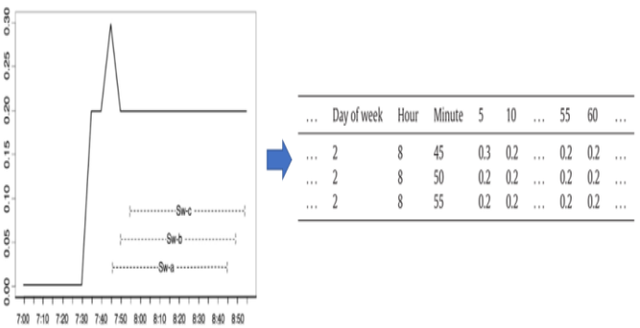


Fig. 10. Feature Preparation 데이터 전처리 과정의 예시

C. 클러스터링을 이용한 이상데이터 검출 [7]

1) 연구 목표

비정상적인 전력사용 패턴을 검출하여 빌딩 관리자에게 알림으로써 에너지를 절약하고자 한다.

2) 사용 데이터

Powersmiths라는 센서회사에서 제공하는 캐나다의 모 학교에서 2013년부터 2015년까지 5분 간격으로 생성된 HVAC 전력소비량 데이터를 사용하였다.

3) 분석 방법

a) 이상데이터 검출 방법의 구조

본 연구에 적용된 Collective Contextual Anomaly Detection with Sliding Window(CCAD-SW) 방법의 구조를 Fig. 9에 나타내었다. CCAD-SW 방법은 학습(Learning) 단계와 이상 데이터 검출 단계로 구성된다.

Data Cleaning은 완전한 오류 데이터를 제거하는 것으로, 본 연구에서는 0보다 작은 전력사용량 데이터를 제거하는 과정을 적용하였다. Feature Preparation은 하나의 전력사용량 대신에 일정한 시간 구간(Sliding Window)에서의 전력사용량 값들로 이루어지는 새로운 데이터를 만들고(Data Rearrangement), 겹치는 시간 구간에 대한 정보를 저장하기 위해 해당 시간 구간에 대한 시간 인덱스를 새로운 입력 항목으로 추가(Feature Generation)하는 것이다.

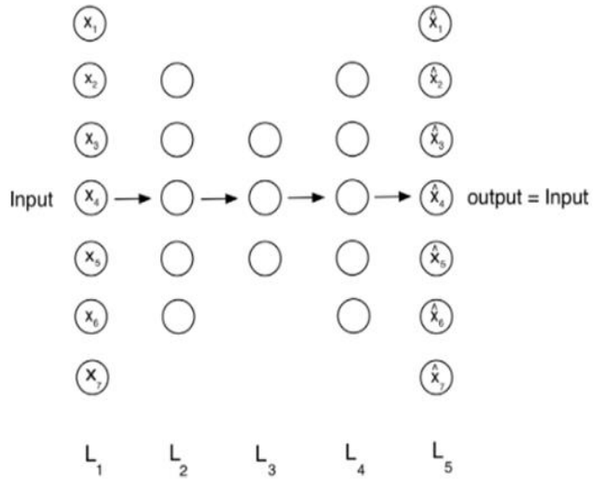


Fig. 11. Autoencoder 패턴 학습 방법의 구조

그러면 Fig. 10과 같이 전력사용량 시계열 데이터는 일련의 시간 구간 데이터로 변환된다.

학습(Training) 단계에서는 Autoencoder, Principal Component Analysis(PCA)와 같은 패턴 학습 알고리즘을 사용할 수 있고, 본 연구사례에서는 Autoencoder 방법을 사용하였다.

① Autoencoder (AE)

비지도(Unsupervised) 학습 방법의 하나로서, Fig. 11과 같이 차원(Dimension)이 작아졌다가 다시 커지는 은닉층(Hidden Layer)을 가지고 출력은 입력의 차원과 같게 하는 인공 신경회로망이다. 은닉층은 입력과 출력의 오차가 작아지도록 학습이 이루어지고, 학습이 완료되면 차원이 작은 은닉층(예를 들어, Fig. 11의 L3 계층)의 값을 입력에 대한 축소된 차원의 부호(Code)로 사용한다.

② Principal Component Analysis (PCA)

PC도는 입력 데이터의 차원을 줄이고 특징을 추출하는데 활용되는 비지도 학습 방법이다. 임의의 벡터를 고유벡터의 선형합(Linear Combination)으로 나타내듯이, 데이터의 분산 패턴을 잘 표현하는 특징 벡터를 찾아서 다음과 같이 원본 데이터를 특징 벡터와의 내적으로 변형하는 것이다.

$$Z_m = \emptyset_{1m} X_{i1} + \emptyset_{2m} X_{i2} + \cdots + \emptyset_{pm} X_{ip} \tag{9}$$

여기서 $[\emptyset_{1m}, \emptyset_{2m}, \cdots, \emptyset_{pm}]^T$는 $m$번째 Principal Component의 Loading이고, $i$번째의 차원이 $p$인 데이터 $[X_{i1}, X_{i2}, \cdots, X_{ip}]$는 식 (9)에 따라 $m$번째 Principal Component인 $Z_m$으로 표현되는 것이다. 일반적으로 Principal Component는 1개 또는 2개 정도를 구하므로, 입력 데이터보다 차원이 줄어든 새로운 데이터를 생성할 수 있을 뿐만 아니라, 차원이 축소된 데이터를 시각적으로 표현할 수 있게 된다. Principal Component는 다음과 같이 데이터를 잘 표현하는, 즉 변환시켰을 때 분산이 커지게 하는 패턴을 찾도록 하는 최적화 문제를 풀어서 구한다.

$$\max_{\emptyset_{11},\cdots,\emptyset_{p1}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\emptyset_{j1}x_{ij}\right)^2 \qquad s.t. \sum_{j=1}^{p}\emptyset_{j1}^2 = 1 \tag{10}$$

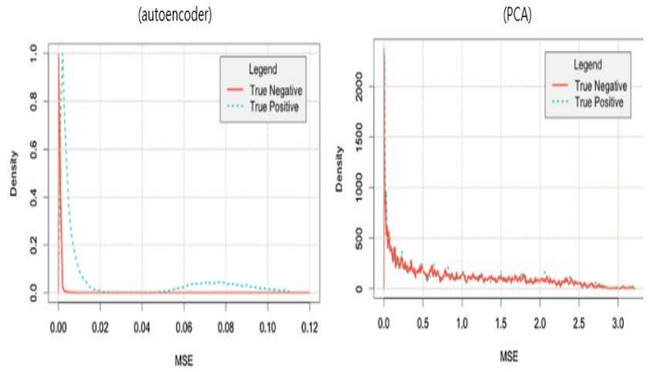


Fig.12. 이상데이터 검출 기준을 설정하기 위한 오차 분포

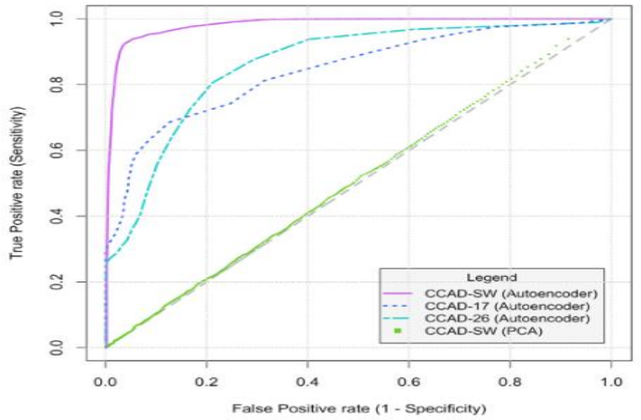


Fig. 13. 이상 데이터 검출 성능 및 다른 방법과의 비교

b) 이상 데이터 검출 방법

Fig. 9의 테스트 단계에서 이상 데이터(True Positive)의 오차 분포 및 정상 데이터(True Negative)의 오차 분포를 구한다. 오차는 원본 데이터와 학습된 Autoencoder로부터 복원된 데이터와의 차이를 계산한다. 어느 정도 성능을 보이는 방식이라면, 원본 데이터와의 오차가 작은 부분에 집중적으로 분포할 것이고, 이상 데이터에 관한 결과는 어떤 분포 형태를 띨 것이다. 이러한 오차 분포로부터 이상 데이터 검출을 위한 오차 기준을 설정한다. 오차 기준을 작게 하면 이상 데이터를 잘 검출하지만 (높은 True Positive Rate), 정상 데이터를 이상 데이터로 분류하는 오류(False Positive Rate)도 증가한다. 따라서 True Positive Rate와 False Positive Rate를 두 축으로 하는 Receiver Operating Characteristic(ROC) 곡선을 그려서 적절한 오차 기준을 설정하는 것이 중요하다.

4) 분석 결과

Fig. 12는 Autoencoder의 오차 분포(왼쪽)와 PCA의 오차 분포(오른쪽)를 나타낸 것이다. PCA 방식보다 Autoencoder 방식을 적용했을 때 이상 데이터 검출을 위한 오차 기준을 설정하는 것이 쉬움을 확인할 수 있다.

이상 데이터 검출 성능을 보여주는 ROC 곡선의 결과는 Fig. 13과 같다. 제안하는 검출 방법(CCAD-SW)이 다른 방법보다 성능이 좋음을 확인할 수 있으며, 최적의 결과는 검출 성능이 94.5% (True Positive Rate), 검출 오류는 5.7% (False Positive Rate)가 나왔다. 특히 흔히 사용되는 PCA의 경우 성능이 떨어지는 정도가 아닌 아예 검출 방법으로 사용할 수 없다는 결과가 나왔는데, 이는 실제 스마트미터 데이터를 바탕으로 새로운 기능 및 서비스를 구현할 때 확인해볼 필요가 있어 보인다.

TABLE 2
결측치 처리 사례연구 요약

| 구분 | DAE | HA |
|---|---|---|
| 랜덤 발생 | 32.8 | 44.6 |
| 블록 발생 | 43.2 | 51.1 |

## III. 머신러닝 기법을 활용한 에너지 데이터 분석 및 시각화 기술 개발

### A. Denoising Autoencoder (DAE) 기반의 결측치 처리 알고리즘 개발

1) 결측치 처리를 위한 DAE 모델 구조

선행연구 분석을 통해서 살펴본 Autoencoder는 x를 입력했을 때, 데이터 x를 그대로 출력(복원)하도록설계된 구조였다면, DAE는 입력으로 x 대신에 부분적으로 손상된 $\hat{x}$ 를 입력하여 손상되지 않은 데이터 x를 출력(복원)하도록 하는 구조이다. 따라서 DAE는 원본 데이터와 손상된 데이터 사이의 공통된 잠재적 특징을 학습하여 입력의 부분적 손상에 대해서 강건해지도록 한다. 데이터를 손상시키는 함수를 $h$ 라고 할 때, $\hat{x} = h(x)$ 라는 수식을 얻을 수 있다. 인코더함수를 $f$, 디코더 함수를 $g$라고 한다면 DAE는 $x$와 $g(f(\hat{x}))$ 사이의 오차를 최소화하도록 훈련한다.

따라서, 본 연구를 통해서 제안하고자 하는 DAE의 학습 방법은 다음과 같다. 15분 단위의 에너지사용량 데이터를 입력데이터 $x$로 사용하고, 이때 $x$는 결측치가 없는 데이터로 가정한다. 그리고, 임의의 결측치를 만들어서 손상된 데이터 $\hat{x}$를 구성한다. 이를 통해 DAE는 결측치가 포함된 데이터 $\hat{x}$를 입력받더라도 결측치가 없는 원래의 데이터 $x$를 출력하도록 학습할 수 있다. 이때 결측된 데이터는 0으로 처리하는데, 에너지 사용량이 0이 되는 경우는 없으므로, 0으로 표기하여 결측치임을 쉽게 확인할 수 있도록 표현하였다.

2) 기존모델 대비 DAE 모델의 성능

DAE의 성능평가를 위해 60개의 가정용 고객의 1년치(15분 단위) 데이터를 사용하였으며, 성능평가의 척도는 RMSE(Root Mean Squared Error)를 활용하였다. DAE의 비교군으로는 HA(Historical Average)를 사용하였으며, 이때 15분 단위 데이터 기준 (t-1)과 하루 전 같은 시간대의 값인 (t-96) 시간대의 평균치로 값을 구성하였다. 결측치는 임의의 확률로 5%, 10%, 30%, 50% 설정하여 발생시키는 랜덤 발생 형태와 특정 시간 간격(1시간, 3시간, 6시간, 12시간) 동안을 결측치로 발생시키는 블록 형태로 나누었으며, 그 결과는 다음과 같다.

결측치를 랜덤으로 발생시키는 경우와 블록 형태로 발생시키는 두 경우 모두에 대해서, DAE가 HA 알고리즘보다 우수한 성능을 나타냄을 확인할 수 있다.

## IV. 결론

스마트미터의 데이터를 분석하여 활용하는 연구는 다양하게 진행되고 있다. 본 연구에서 살펴본 결측치 처리 뿐 아니라, 수요 패턴 분석, 소비자 특성 분석, 수요반응 프로그램 설계, 요금제 설계 등 그 활용성이 다양하다. 그러나 결국 이러한 연구들이 원활하게 수행되기 위해서는 결측치에 대한 처리가 우선되어야 하기에, 본 연구에서는 스마트미터 데이터의 결측치를 처리하는 선행연구 분석을 통해 응용 사례와 적용방법들에 대해서 살펴보고, 새로운 결측치 처리 알고리즘을 제안하여 이에 대한 사례연구를 수행하였다. 사례연구 결과, 제안한 방법이 기존의 방법과 비교하여 결측치 처리에 있어서 우수한 성능을 나타냄을 알 수 있었다. 비록, 본 논문에서는 결측치를 처리하는 방안에 대해서만 다루었지만, 향후에는 예측 알고리즘과 결합하여 결측치에 강인한 예측 알고리즘의 개발도 진행할 계획이다.

## ACKNOWLEDGEMENT

This research was supported by Smart City R&D project of the Korea Agency for Infrastructure Technology Advancement(KAIA) grant funded Ministry of Land, Infrastructure and Transport (Grant 21NSPS-B149869-04)

본 연구는 국토교통부 국토교통과학기술진흥원의 스마트시티 혁신성자동력 프로젝트 지원으로 수행되었음 (과제번호 21NSPS-B149869-04)

## REFERENCES

[1] U.S. Energy Information Agency, “FAQ: How many smart meters are installed in the United States, and who has them? ”, Available: http://www.eia.gov/tools/faqs/faq.php?id=108&t=3

[2] “Smart Meters, Quarterly Report to end December 2017”, Department for Business Energy and Industrial Strategy, Great Britain, Technical Report, 2018.

[3] “제2차 지능형전력망 기본계획”, 산업통상자원부, 2018.

[4] Y. Wang, Q. Chen, T. Hong, C. Kang, “Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges,” IEEE Trans. Smart Grid, vol. PP, no. 99, pp. 1–1, 2018.

[5] J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, “Handling bad or missing smart meter data through advanced data imputation,” in IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5, 2016.

[6] A. Al-Wakeel, J. Wu, and N. Jenkins, “k-means based load estimation of domestic smart meter measurements,” Applied Energy, vol. 194, pp. 333–342, 2017.

[7] D. B. Araya, K. Grolinger, H. F. ElYamany, M. A. Capretz, and G. Bitsuamlak, “An ensemble learning framework for anomaly detection in building energy consumption,” Energy and Buildings, vol. 144, pp. 191–206, 2017.