The Korean Journal of Applied Statistics
2021, Vol. 34, No. 6, 969–977

DOI: https://doi.org/10.5351/KJAS.2021.34.6.969
Print ISSN1225-066X / Online ISSN 2383-5818

Clustering load patterns recorded from advanced metering infrastructure

Hyojung Ann [a], Yaeji Lim [1,a]

[a]Department of Statistics, The Graduate School of Chung-Ang University

Abstract

We cluster the electricity consumption of households in A-apartment in Seoul, Korea using Hierarchical K-means clustering algorithm. The data is recorded from the advanced metering infrastructure (AMI), and we focus on the electricity consumption during evening weekdays in summer. Compare to the conventional clustering algorithms, Hierarchical K-means clustering algorithm is recently applied to the electricity usage data, and it can identify usage patterns while reducing dimension. We apply Hierarchical K-means algorithm to the AMI data, and compare the results based on the various clustering validity indexes. The results show that the electricity usage patterns are well-identified, and it is expected to be utilized as a major basis for future applications in various fields.

Keywords: advanced metering infrastructure, clustering algorithm, clustering validity index, Hierarchical K-means clustering, power consumption data

1. 서론

경제가 성장하면서 전력 사용량이 점점 많아지고 있으며 앞으로도 전력 사용량은 계속 증가할 것이라 예상된다. 전력 사용량이 증가함에 따라 경제, 사회 등 다양한 분야에서 여러 방법을 이용해 전력 사용량의 패턴을 파악하기 위해 노력하고 있으며, 전력 사용량 관련 연구들 또한 증가하고 있다.

전력 사용량 데이터를 이용한 몇 가지 연구 사례를 살펴보면 다음과 같다. Kim과 Park (2015)는 국내 전력 수요의 추세 변화의 여부와 발생 시기를 추정하고, 추세 변화가 일시적인 현상인지 구조적인 현상인지를 분석하였다. Yoo 등 (2019)은 수집된 전력 소비 데이터를 분석하여 발생할 수 있는 오류들을 정리하였으며, K-means 군집화 알고리즘을 사용해 소비 패턴을 월별로 분석하였다. Jung 등 (2017)은 K-means 군집화 알고리즘을 사용하여 복잡한 상업용 건물의 단위별 전력 소비 패턴 특성을 분석하고 군집하였다. Kim 등 (2015)은 2005년부터 2013년 동안의 분기별 평균기온 자료와 소득, 전력 가격, 전력 사용량 자료를 사용하여 전력 수요함수를 추정하였다.

K-means 방법론은 간단한 이론에 기반되어 있고 그 수렴속도가 빠르므로 전력 사용량 데이터 분석에 자주 사용되어 왔다. 하지만 기존의 K-means 알고리즘은 초기값에 민감하며, 대용량 자료의 경우 로컬 최적값 (local optimal solution)으로 수렴한다는 것이 알려져있다. 이를 보완하기 위해 많은 연구가 이루어졌으며 그 중 하나로서 Xu 등 (2015)이 Hierarchical K-means (H-K-means) 군집화 알고리즘을 제안하였다.

This research was supported by the National Research Foundation of Korea (NRF) funded by the Korean government (NRF-2021R1A2B5B01001790) and Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20199710100060).

[1] Corresponding author: Department of Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: yaeji@cau.ac.kr

Published 31 December 2021 / journal homepage: http://kjas.or.kr

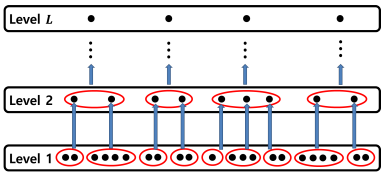

Figure 1: *Hierarchical structure in H-K-means algorithm.*

초기값에 크게 영향을 받는 K-means 군집화 알고리즘과는 달리 H-K-means 군집화 알고리즘은 초기값에 민감하지 않고, 차원축소를 통해 대용량 데이터를 적은 개수의 패턴으로 충분히 설명할 수 있는 방법론이다. Xu 등 (2015)에서는 다음의 이유로 H-K-means 군집화 알고리즘의 우수성을 설명하였다.

1. K-means 군집화 알고리즘에 비해 군집결과의 질이 향상된다.
2. K-means 군집화 알고리즘에 비해 속도가 빠르다.
3. 대용량 자료에 적합하다. 특히, 대용량 전력 사용량 자료에 대해 효과적으로 군집 분석을 시행한다.

본 연구는 H-K-means 군집화 알고리즘을 사용하여 서울 A아파트 가구들의 전력사용패턴을 분석하고자 하였으며, 여름 저녁 피크 시간대의 전력 사용량 데이터를 사용하여 가구를 군집화 하고자 하였다. 이를 위해, AMI를 통해 얻어진 2020년 7월 1일부터 2020년 8월 31일 중 오후 5시부터 8시까지의 한 시간 간격의 전력 사용량 데이터를 사용하여 본 연구를 진행하였다.

논문의 구성은 다음과 같다. 2장에서는 본 연구에서 사용한 H-K-means 군집화 알고리즘에 대해서 간단히 설명한다. 3장에서는 서울의 A아파트에서 얻어진 AMI 자료에 군집화 알고리즘을 적용한 후 그 결과를 해석하고, 군집화 유효성 지수들을 이용해 다양한 군집 개수 및 level에 따른 군집화 알고리즘의 성능을 살펴본다. 4장에서는 결론 및 향후 연구에 대해 논의한다.

2. Hierarchical K-means 군집화 알고리즘

앞에서 언급한 바와 같이, K-means 군집화 알고리즘의 결과는 초기 중심값에 민감하며, local optimal solution 들로 인해 어려움이 있을 수 있다. 이러한 어려움을 극복하기 위해 Xu 등 (2015)는 H-K-means 군집화 알고리즘을 제안하였다. H-K-means의 주요 아이디어는 다음과 같다.

주어진 데이터셋에 대해서 서로 가까운 패턴들을 하나의 대표 패턴으로 처리한다면, 전체 데이터셋을 원래 패턴과 유사한 분포를 작은 크기의 데이터셋으로 표현할 수 있을 것이다. 이러한 관점에서 H-K-means의 계층 구조는 Figure 1과 같이 구축된다.

여기에서 level 1은 원본 데이터셋이며, level L은 사용자로부터 정의된 최종 level 값이다. 단계를 진행함에 따라 그림과 같이 원본 데이터셋보다 축소된 데이터셋을 얻을 수 있으며 최종 level L에서는 원본 데이터셋의 패턴을 반영한 축소된 최종 데이터셋을 얻게 된다. Level $l-1$에서 level l로 진행됨에 따라 다음을 만족하는 차원축소가 이루어져야 한다.

a. Level l의 자료는 level $l-1$의 자료와 비슷하여야 한다. 즉, 자료의 고유특성이 유지된 상태에서 차원축소가 이루어져야 한다.
b. Level이 높아질수록 자료의 갯수는 줄어들어야 한다.

위 두 조건을 만족하기 위해, H-K-mean 군집화 알고리즘에서는 군집 내 유사성을 나타내는 θ값을 정의한다. 아래 알고리즘의 수식 (2.1)로 정의되는 θ값은 그 값이 작을수록 잘 생성된 군집이며, 이 값이 미리 정한 threshold값인 t_l보다 작은 경우에는 그 중심값을 대표패턴으로 여겨 차원축소를 진행한다. 만약, θ값이 t_l 값보다 큰 경우에는 하위 군집으로 분할하여 얻은 중심값을 대표 패턴으로 사용한다.

보다 자세한 알고리즘 설명은 다음과 같다. 먼저, $\boldsymbol{x}_i$를 i번째 시계열 패턴이라고 정의하자. 이때 $\boldsymbol{x}_i$는 d 차원의 벡터이며, 원소인 x_{it}는 i번째 패턴의 t번째 시간에서의 값이다. 따라서, N개의 패턴을 군집화하기 위한 H-K-means 알고리즘의 입력값은 다음과 같은 행렬로 표현할 수 있다.

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)' = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1T} \\ x_{21} & x_{22} & \ldots & x_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{NT} \end{pmatrix}$$

이를 입력값으로 사용하여 다음의 단계가 진행된다.

- *Stage I* – 계층 구조 설정

Step 1. 원본 데이터셋을 level 1 데이터셋, D_1,으로 설정하고, $l = 2$로 정한다.

Step 2. 아래의 Algorithm을 이용하여 level $(l-1)$ 데이터셋, D_{l-1}를 기반으로 level l 데이터셋, D_l,를 생성한다. 이 때, 필요한 D_l에 대한 추정 크기(M_l) 및 한계값 (t_l)은 사용자로부터 설정된 값을 사용한다.

$D_l = \phi$라 하자.
K-means를 사용하여 D_{l-1}의 패턴을 M_l개의 군집으로 분리한다.
for $j = 1$ ***to*** $j = M_l$ **do**
  j번째 군집에 대한 θ_j를 다음과 같이 계산한다.

$$\theta_j := \max_{p=1,\ldots,n_j} \left(\frac{\left\| \boldsymbol{x}_p^j - \boldsymbol{c}^j \right\|^2}{\left\| \boldsymbol{c}^j \right\|^2} \right) \tag{2.1}$$

  여기서 $\boldsymbol{x}_p^j$는 j번째 군집의 p번째 패턴, n_j는 j번째 군집의 크기 (군집에 속하는 패턴의 갯수)이며, $\boldsymbol{c}^j$는 j번째 군집의 중심값이다. $\|\cdot\|$은 L^2 norm이다.
  if $\theta_j \le t_l$ **then**
    j번째 군집의 중심값을 D_l에 추가한다.
  else
    $k = n_j$, $\theta_{\max} = 0$이라 설정하고, j번째 군집의 모든 k개 패턴을 $\boldsymbol{C}^j := \{\boldsymbol{C}_1^j, \ldots, \boldsymbol{C}_k^j\}$라 한자.
    while $\theta_{\max} \le t_l$ **do**
      1. $\boldsymbol{R}^j := \boldsymbol{C}^j$이라 하고, $\boldsymbol{C}^j$에서 가장 가까운 두 패턴 중 하나를 제거한다. $k = k - 1$로 업데이트한다.
      2. $\boldsymbol{C}^j$의 패턴을 초기 중심값으로하여 K-means로 j번째 군집의 패턴을 k개의 하위 군집들로 군집화 한다.
      3. k개의 하위 군집들로 $\boldsymbol{C}^j$를 업데이트한다.
      4. k개의 하위 군집들에 대해 수식 (2.1)로 $\{\theta_{j1}, \ldots, \theta_{jk}\}$를 계산하고, $\theta_{\max} := \max\{\theta_{j1}, \ldots, \theta_{jk}\}$를 설정한다.
    end while
    $\boldsymbol{R}^j$의 패턴을 D_l에 추가한다.
end for

Step 3. $l = L$인 경우 Stage II로 이동하고, 그렇지 않은 경우 $l = l + 1$로 두고 Step 2로 이동한다.

- *Stage II* – 가중 군집화

Step 4. $l = L$인 경우 level l 데이터셋 중에서 K개의 패턴을 무작위로 선택하고, 아닌 경우에는 아래의 Step 5를 통해 얻어진 K개의 중심값 패턴을 사용한다.

Step 5. Step 4에서 얻은 K개의 패턴들을 초기 중심값으로 사용해 level l 데이터셋에 대해서 K-means 군집화를 시행하며, 각 군집의 중심값은 다음의 $\boldsymbol{w}_k$로 업데이트 시킨다.

$$\boldsymbol{w}_k = \frac{\sum_{p=1}^{n_k} \left(r_p \times \boldsymbol{x}_p^k\right)}{\sum_{p=1}^{n_k} r_p}, \quad k = 1, \ldots, K.$$

여기서 n_k는 level l 데이터셋에 있는 k번째 군집의 크기이고, $\boldsymbol{x}_p^k$는 k번째 군집에서의 p번째 패턴이다. $l > 1$인 경우 r_p는 level $(l-1)$ 데이터셋에서 $\boldsymbol{x}_p^k$가 속했던 군집의 크기이고, 그렇지 않은 경우 $r_p = 1$로 둔다.

Step 6. $l = 1$인 경우 process를 중단하고, Step 5에서 얻어진 결과를 최종 군집화 결과로 출력한다. 그렇지 않은 경우 $l = l - 1$로 업데이트한 후, Step 4로 돌아간다.

Remark

M_l과 t_l값은 사용자가 미리 선택해야하는 매개변수(parameter)들이며, Xu 등 (2015)에서는 computation efficienty를 고려하여 패턴의 갯수를 50% 감소해주는 매개변수를 선택하였다. 본 논문에서는 패턴의 갯수를 50% 감소시키는 상황 하에서 최적의 결과를 주는 t_l값을 trial-and-error 방식으로 정하였다.

3. Advanced metering infrastructure 자료 분석

3.1. 데이터

서울의 A아파트로부터 얻어진 $N = 29,068$개의 여름 저녁 피크 시간대의 전력 소비 패턴에 대해 H-K-means 군집화 알고리즘을 적용하였다. i번째 일일 전력 사용량 패턴을 $\boldsymbol{x}_i$라 하고, i번째 패턴에서의 t번째 시간의 전력 사용량을 x_{it}라 하자. 본 연구에서는 508가구에서 얻어진 총 62일 간의 자료를 사용하였으므로 $N = 31,496$이며, 저녁시간에 해당하는 오후 5시부터 8시까지의 1시간 간격의 자료이므로 $T = 4$이다.

분석에 앞서서 IQR rule를 사용해 이상치들을 선별하여 제거하였다. IQR rule은 값이 제 3사분위수(Q_3)와 제 1사분위(Q_1)로부터 $(1.5) \times IQR$만큼 떨어져있는 값들을 이상치로 정의한다. 따라서, 전력 사용량의 값이 이상치이거나 0인 패턴을 제외하고 남은 최종 표본 크기는 $N = 29,068$이다.

또한, 본 연구에서는 주어진 원 전력 사용량(raw load profile)이 아닌 cumulative load profile을 사용하였다. 이는 raw load profile 간의 Euclidean distance가 자료를 제대로 반영하지 못하는 경우가 발생하기 때문이다 (Kwon과 Park, 2020). 이 문제에 대한 간단한 예시가 Figure 2에 제시되어 있다. Figure 2의 상단 그림은 서로 다른 세 가구에 대한 raw load profile이며 이들은 서로 다른 패턴을 보인다. 하지만 세 가구의 pairwise Euclidean distance를 구해보면 모두 20~22의 값이 나온다. 즉, 서로 다른 패턴을 가지지만 Euclidean distance가 같으므로 이를 사용하여 군집화 방법론을 적용하면 패턴의 차이를 제대로 반영하지 못한 군집분석 결과를 얻게 된다.

이를 해결하기 위해 Satre-Meloy A 등 (2020)은 cumulative load profile $\boldsymbol{c}_i = \{c_{i1}, \ldots, c_{i4}\}$를 다음과 같이 정의하여 사용하였다.

$$c_{it'} = \sum_{t=1}^{t'} x_{it}, \quad i = 1, \ldots, N, \quad t' = 1, \ldots, 4.$$

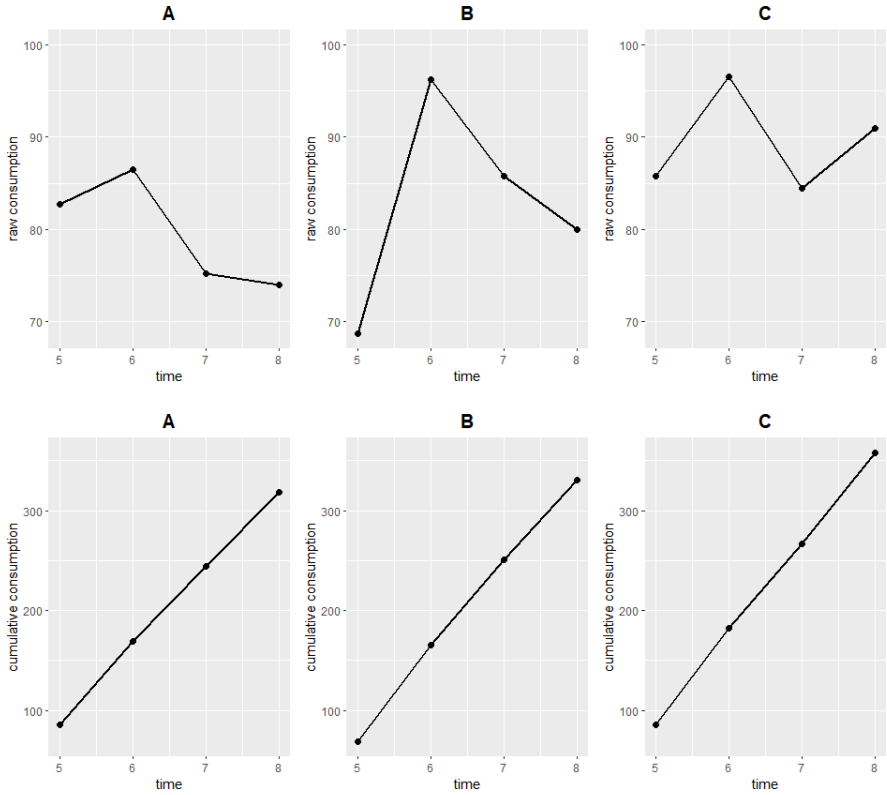

Figure 2: *(Top) Three raw load profiles with the same Euclidean distance between each pair, and (bottom) corresponding cumulative load profiles.*

Figure 2의 하단 그림은 세 가구의 cumulative load profile이다. 이들에 대해 Euclidean distance를 다시 구해보면, A 패턴과 B 패턴 간의 거리는 20, B 패턴과 C 패턴 간의 거리는 39, A 패턴과 C 패턴의 거리는 47로 계산이 되어, 세 가구의 서로 다른 전력 사용량 패턴이 잘 반영됨을 확인할 수 있다. 따라서 본 연구에서는 원자료의 패턴을 잘 반영하기 위해 raw load profile이 아닌 cumulative load profile을 고려하였다.

H-K-means 알고리즘에서 사용자가 정해야 하는 level L은 $L = 2, 3, 4$를 고려하였고, Step 2에서 필요한 추정 크기 (M_l) 및 한계값 (t_l)은 패턴의 갯수를 50% 감소해주도록 다음과 같이 설정하였다.

- Level 2 : $M_2 = 300, \quad t_2 = 0.0065$로 설정하여 얻은 최종 패턴의 갯수는 $14,755$개이다.
- Level 3 : $M_3 = 150, \quad t_3 = 0.014$로 설정하여 얻은 최종 패턴의 갯수는 $7,100$개이다.
- Level 4 : $M_4 = 75, \quad t_4 = 0.025$로 설정하여 얻은 최종 패턴의 갯수는 $3,453$개이다.

3.2. 군집화 결과

군집의 개수는 $K = 2, 3, 4$인 경우를 고려하여 결과를 정리하였다. Figure 3의 첫 번째 행 그래프들은 서울의 A아파트 가구들의 전력사용량을 H-K-means 군집화 알고리즘을 이용해 $K = 2$로 군집했을 때의 군집별 raw load profile의 평균 그래프이다. Level(L)이 $2, 3, 4$인 경우의 군집화 결과를 비교해보면, 모든 level에서 전력사용량에 따라 크게 두 군으로 구분되었다. 전력 사용량이 적은 첫 번째 군집은 오후 시간 동안 일정하게 낮은

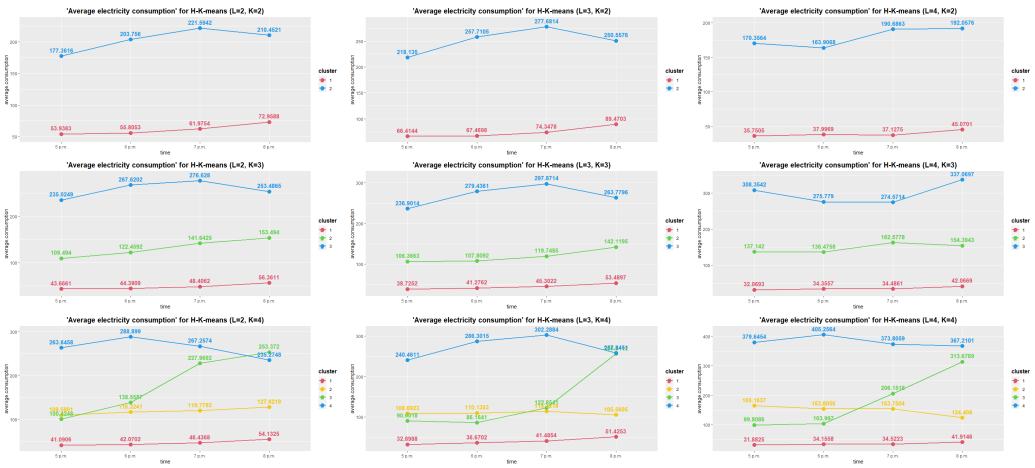

Figure 3: *Average electricity consumption for each cluster when the number of cluster group is* $K = 2$ *(top),* $K = 3$ *(middle), and* $K = 4$ *(bottom).*

소비량을 가졌으며 전력 사용량이 많은 두 번째 군집은 level에 따라 그 패턴의 차이를 보였다. $L = 2, 3$인 경우에는 오후 5시부터 7시까지 사용량이 많았으며, 상대적으로 7시 이후에는 사용량이 적었던 것에 반해, $L = 4$인 경우에는 오후 7시부터 8시까지 가장 많은 전력 사용량을 보였다. 두 번째 행 그래프들은 $K = 3$으로 군집했을 때의 결과이다. 모든 level에서 전력 사용량에 따라 크게 세 군으로 구분되었지만, $L = 4$인 경우에 한해서 세 번째 군집이 다른 level과는 상이한 패턴을 보였다. 세 번째 행 그래프들은 $K = 4$로 군집했을 때의 결과로, level에 따라 패턴의 차이가 조금씩 발생하였다. $L = 2$인 경우에는 오후 8시에 세 번째 군집에 속하는 패턴들의 전력 사용량이 네 번째 군집에 속하는 패턴들보다 평균적으로 많아졌다. 반면에, $L = 3, 4$에서는 이러한 역전현상은 보이지 않았으며, $L = 3$의 경우에는 세 번째 군집의 패턴들이 오후 7시까지 상대적으로 적은 전력사용을 보였다.

그래프와 함께 군집의 질을 비교하고자 군집화 유효성 지수인 Mean Index Adequacy (MIA), Davies Bouldin, Dunn index를 구해보았다. 각 지수들에 대한 간단한 설명은 다음과 같다.

- Mean Index Adequacy

  MIA는 각 군집 내 패턴과 해당 군집의 중심 사이의 거리의 평균으로, 다음의 식을 이용해 구할 수 있다 (Chicco G 등, 2006).

$$\text{MIA} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{x\in S_k} d^2(c_k, x)}.$$

  여기서 K는 군집의 개수, d^2은 Euclidean distance의 제곱, n_k는 군집 k의 크기, c_k는 군집 k의 중심점, S_k는 k번째 군집을 의미한다. MIA index의 값이 작을수록 성능이 좋은 군집화 알고리즘이라 평가되고, 값이 클수록 성능이 좋지 않은 군집화 알고리즘이라 평가된다.

- Davies-Bouldin Index

  Davies-Bouldin Index는 다음의 식을 이용해 구할 수 있다 (Desgraupes B, 2017).

$$\text{DB} = \frac{1}{K}\sum_{k=1}^{K}\max_{j\neq k}\left(\frac{\sigma_k + \sigma_j}{d(c_k, c_j)}\right).$$

Table 1: Results of the clustering validity index

		H-K-means (L = 2)	H-K-means (L = 3)	H-K-means (L = 4)	K-means
K = 2	MIA	142.2119	179.8579	250.6793	**132.1456**
	Davies Bouldin Index	**0.5609**	0.5679	0.9474	0.6196
	Dunn Index	0.0020	**0.0035**	0.0011	0.0009
K = 3	MIA	158.2713	174.4882	199.7740	**133.1554**
	Davies Bouldin Index	0.6015	0.6933	**0.5663**	0.6267
	Dunn Index	0.0011	0.0025	**0.0297**	0.0007
K = 4	MIA	153.3724	202.2675	203.1067	**134.6585**
	Davies Bouldin Index	0.6569	0.7190	**0.6036**	0.6343
	Dunn Index	0.0004	**0.0011**	0.0006	0.0008

여기서 K은 군집의 개수, c_k는 군집 k의 중심점, σ_k는 군집 k 내의 모든 패턴으로부터 중심점 c_k까지 거리의 평균값, $d(c_k, c_j)$는 중심점 c_k와 중심점 c_j 간의 Euclidean distance이다.

높은 군집 내 유사도 및 낮은 군집 간 유사도를 가지는 군집결과 에 대해서 매우 작은 Davies-Bouldin Index 값을 가지게 된다.

- Dunn Index

  Dunn Index는 군집 간 최소 거리와 군집 간 최대 거리의 비율로 정의된다. 다음의 식을 이용해 각 군집에 대하여 Dunn Index를 구할 수 있다 (Desgraupes B, 2017).

$$D = \frac{\min_{1 \le i < j \le K} d(i, j)}{\max_{1 \le k \le K} d'(k)}.$$

  여기서 $d(i, j)$는 군집 i와 군집 j 간의 Euclidean distance이고, $d'(k)$는 군집 k의 군집 내 거리이다. 본 논문에서는 군집 간 거리인 $d(i, j)$는 두 군집의 중심점 간의 거리로 정의하였으며, 군집 내 거리인 $d'(k)$는 군집 내 가장 멀리 떨어진 패턴들 간의 거리로 구하였다.

  높은 군집 내 유사도와 낮은 군집 간 유사도에 높은 점수를 주기 때문에 Dunn Index 값이 큰 결과는 군집화 성능이 좋은 것이라 판단할 수 있다.

Table 1은 군집한 결과들에 대해서 계산된 군집화 유효성 지수들을 정리한 것이다. 모든 경우에 대해 MIA 기준으로는 K-means 방법론이 가장 우수한 결과를 보였다. Davies Bouldin Index와 Dunn Index 기준으로는 K-means 방법론보다 H-K-means 방법론이 더 우수한 결과를 제공하였으며, K = 2인 경우에는 상대적으로 낮은 level에서 군집의 결과가 좋았으며, K = 3, 4인 경우에는 상대적으로 높은 level인 L = 4에서 대체적으로 좋은 결과를 얻었다.

4. 결론

본 연구에서는 차원을 축소해주면서 패턴을 파악할 수 있는 H-K-means 군집화 알고리즘을 적용하여 서울의 한 아파트 단지의 AMI를 통해 얻어진 전력 사용량을 분석해보았다. H-K-means 군집 분석은 기존의 K-means 군집화 방법론의 한계를 극복하기 위해 제시된 방법으로써 원본 데이터셋의 패턴을 반영하면서 동시에 차원 축소가 가능한 통계 방법론이다. 다양한 군집 개수 및 level에 따라 군집 결과를 정리하였으며, 이를 군집화 유효성 지수들을 이용해 비교해보았다.

군집화 유효성 지수들에 대해서 일치된 결과가 나오지 않았으며, MIA 지수 기준으로는 K-means 방법론이 우수한 결과를 주었다. 다만, H-K means 군집화 알고리즘에서 사용되는 매개변수들을 통계적 기법을 통해 최적의 값으로 정할 수 있다면, H-K means 군집화 알고리즘을 보완할 수 있을 것으로 기대한다.

또한, 추가적인 가구 정보 등을 활용할 수 있다면 결과를 다방면에서 해석 가능할 것이다. 예를 들어, 가구원 수에 따른 전력 사용량의 패턴, 건조기 유무에 따른 전력 사용량의 차이, 가구원 나이에 따른 전력 사용량의 비교 등 추가적인 분석을 수행할 수 있을 것이다.

향후 연구에서는 군집화 결과에 따른 가구 특성의 차이, 새로 도입되는 계시별 요금제에 따른 군집별 요금 변동의 비교 등을 진행하고자 한다.

References

Chicco G, Napoli R, and Piglione F (2006). Comparisons among clustering techniques for electricity customer classification, *IEEE Transactions on power systems*, **21**, 933–940.

Desgraupes B (2017). *Clustering Indices*, University of Paris Ouest-Lab Modal'X, 1, 34.

Jung D, Yoon Y, and Moon H (2017). *Clustering of Energy Consumption Patterns from a Complex Commercial Building using K-means Algorithm*, Autumn Conference, 175–176.

Kim C and Park G (2015). *Analyze Structural Changes and Factors of Change of Domestic Power Consumption Pattern*, Korea Energy Economics Institute.

Kim HM, Kim IG, Park KJ, and Yoo SH (2015). The effect of temperature on the electricity demand: An empirical investigation, *Journal of Energy Engineering*, **24**, 167–173.

Kwon S and Park M (2020). Time-series data clustering based on the correlation of periodogram, *Journal of The Korean Data Analysis Society*, **22**, 1751–1766.

Satre-Meloy A, Diakonova M, and Grünewald P (2020). Cluster analysis and prediction of residential peak demand profiles using occupant activity data, *Applied Energy*, **260**, 114246.

Xu TS, Chiang HD, Liu GY, and Tan CW (2015). Hierarchical *K*-means method for clustering large-scale advanced metering infrastructure data, *IEEE Transactions on Power Delivery*, **32**, 609–616.

Yoo N, Lee E, Chung BJ, and Kim DS (2019). Analysis of apartment power consumption and forecast of power consumption based on deep learning, *Journal of IKEEE*, **23**, 1373–1380.

Received July 12, 2021; Revised August 5, 2021; Accepted August 6, 2021

AMI로부터 측정된 전력사용데이터에 대한 군집 분석

안효정[a], 임예지[1,a]

[a]중앙대학교 통계학과

요 약

본 연구에서는 Hierarchical K-means 군집화 알고리즘을 이용해 서울의 A아파트 가구들의 전력 사용량 패턴을 군집화 하였다. 차원을 축소해주면서 패턴을 파악할 수 있는 Hierarchical K-means 군집화 알고리즘은 기존 K-means 군집화 알고리즘의 단점을 보완하여 최근 대용량 전력 사용량 데이터에 적용되고 있는 방법론이다. 본 연구에서는 여름 저녁 피크 시간대의 시간당 전력소비량 자료에 대해 군집화 알고리즘을 적용하였으며, 다양한 군집 개수와 level에 따라 얻어진 결과를 비교하였다. 결과를 통해 사용량에 따라 패턴이 군집화 됨을 확인하였으며, 군집화 유효성 지수들을 통해 이를 비교하였다.

주요용어: 지능형 검침 인프라, 군집화 알고리즘, 군집화 유효성 지수, Hierarchical K-means 군집화 알고리즘, 전력 사용량 데이터

본 연구는 2021년도 정부의 재원으로 한국연구재단의 지원을 받았으며 (NRF-2021R1A2B5B01001790), 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다 (No. 20199710100060).

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: yaeji@cau.ac.kr