Journal of The Korea Society of Computer and Information
Vol. 23 No. 5, pp. 9-14, May 2018
www.ksci.re.kr
https://doi.org/10.9708/jksci.2018.23.05.009

Cluster Analysis of Daily Electricity Demand with t-SNE

Yunhong Min*

Abstract

For an efficient management of electricity market and power systems, accurate forecasts for electricity demand are essential. Since there are many factors, either known or unknown, determining the realized loads, it is difficult to forecast the demands with the past time series only. In this paper we perform a cluster analysis on electricity demand data collected from Jan. 2000 to Dec. 2017. Our purpose of clustering on electricity demand data is that each cluster is expected to consist of data whose latent variables are same or similar values. Then, if properly clustered, it is possible to develop an accurate forecasting model for each cluster separately. To validate the feasibility of this approach for building better forecasting models, we clustered data with t-SNE. To apply t-SNE to time series data effectively, we adopt the dynamic time warping as a similarity measure. From the result of experiments, we found that several clusters are well observed and each cluster can be interpreted as a mix of well-known factors such as trends, seasonality and holiday effects and other unknown factors. These findings can motivate the approaches which build forecasting models with respect to each cluster independently.

▸Keyword: Electricity demand, Forecasting, Clustering, t-SNE, Dynamic Time Warping

I. Introduction

한국 전력거래소(KPX: Korea Power Exchange)는 국내 전력산업에서 전력시장의 운영, 전력계통(Power System)의 운영, 실시간 급전운영, 전력수급기본계획 수립 등의 기능을 담당하고 있다 [1]. 전력거래소의 주요 사업 사업으로는 시장운영, 계통운영, 전력수급으로 나뉠 수 있으며 이들 사업에서 전력수요에 대한 예측은 매우 중요한 역할을 하고 있다.

우리나라 전력시장의 시장가격은 1시간 단위로 결정되며 해당 시간에 발생할 것으로 추정되는 수요와 공급의 균형점에서 결정이 된다. 이 때 사용 되는 수요와 공급 곡선은 전력거래 당일 하루 전에 결정이 되며, 하루 전에 예측된 전력수요곡선과 공급입찰에 참여한 발전기들로부터 정의되는 공급 곡선이 가격결정에 사용된다. 따라서 시간 단위의 정확한 전력수요 예측은 정확한 시장가격 형성을 위해서 매우 중요하다.

뿐만 아니라 전기는 일반적으로 저장이 불가능하므로 필요한 전력수요를 예측하고 이를 공급하기 위해 충분한 공급능력을 확보 해야만 안정적인 전력공급이 가능하다. 이를 위해서 전력거래소는 연간·월간·일간의 전력수급 운영계획을 수립하여 운영하고 있으며 이러한 계획수립을 위해서는 수요예측이 필수적이다.

전력 수요예측은 과거의 수요 데이터를 기반으로 현재 혹은 미래의 수요를 예측하는 형태를 갖는다. 이 때 사용 되는 수요 데이터는 시계열(time series) 데이터로 시계열 데이터에 대한 예측 문제는 과거부터 현재까지 활발히 연구되고 있는 분야이다 [2]. 많은 시계열 데이터는 공통적으로 추세성(trend)과 계절성(seasonality)을 갖는데 전력 수요에 대한 시계열 데이터에서는 추세성은 전기제품의 사용이 증가함에 따라 꾸준히 전력

• First Author: Yunhong Min, Corresponding Author: Yunhong Min
*Yunhong Min (yunhong.min@inu.ac.kr), Graduate School of Logistics, Incheon National University
• Received: 2018. 03. 29, Revised: 2018. 04. 19, Accepted: 2018. 05. 25.
• This work was supported by research fund of Incheon National University in 2017.

사용량이 증가하는 것을 의미하며 계절성은 여름 혹은 겨울에 냉난방 기기의 사용량이 많아짐에 따라 해당 계절에 대한 수요가 증가하는 것을 의미한다. 하지만, 전력 수요는 단순히 추세성과 계절성으로만 결정되지 않으며, 날씨, 요일 및 휴일과 같은 사회적 요소, 그리고, 소비자들의 전력사용 패턴과 같은 다양한 요소에 의해 결정된다. 따라서 정확한 전력 수요예측을 위해서는 이들 요소들의 상호 연관 관계와 이들이 전력 수요에 미치는 영향을 알아야 한다.

최근 데이터 가용성(availability)이 증가함에 따라, 전력수요에 영향을 미치는 다양한 요소 또는 변수에 대한 정보를 얻는 것이 가능해졌다. 하지만, 전력 수요예측에 모든 요소들을 파악하는 것은 일반적으로 불가능한 일이며, 파악한 변수 중에서도 데이터 획득이 불가능하거나 불완전한 경우가 있을 수 있다. 이를 해결하는 한 가지 방법은 군집 기반의 예측(clustering based forecasting) [3]이다. 군집 기반의 예측에서는 전체 데이터를 몇 개의 군집으로 나눈 다음, 각 군집 별로 독립적인 예측 모델을 구축한다. 각 군집은 유사한 전력수요 패턴을 갖는 데이터들의 집합으로 이들 데이터들은 전력수요를 결정짓는 일부 변수들에 대해 동일한 값을 갖는다고 가정한다. 따라서 군집별로 예측모델을 만들 경우, 가용하지 않은 변수에 의한 문제에 대한 해결을 일부 기대할 수 있다.

군집 기반의 예측의 정확도를 높이기 위해서는 군집 분석의 정확도가 중요하다. 전력 수요예측에 대한 군집 분석에 대한 기존 연구로는 [4,5]가 있다. [5]에서는 k-평균 군집분석(k-means clustering), [4]에서는 함수적 군집분석 (functional clustering)을 이용하여 군집분석을 수행하였다. [3]은 k-평균 군집분석, 가우시안 혼합 모델 군집분석(Gaussian mixture model clustering), 함수적 군집분석을 비교 분석하기 위해 각 방법에서 얻은 클러스터에서 다양한 예측모델을 학습한 다음, 이들 모델의 정확성으로 군집의 정확도를 평가하였다. 즉, 군집에서 예측모델의 정확성이 높으면 예측 모델에서 사용하지 않은 변수들을 중심으로 군집이 잘 형성됐다고 평가하는 것이다. 하지만, 이들 연구에서 사용한 군집분석 방법은 사용자가 군집의 개수를 미리 결정을 해야 하기 때문에 군집 개수를 결정하기 위해서 사전실험을 해야 한다는 단점이 있다. 뿐만 아니라, 시계열 데이터를 요인분석(factor analysis)으로 차원축소 한 다음 독립적으로 군집 분석을 수행하기 때문에 차원축소 방법과 군집 분석 사이의 상관관계에 대한 분석이 필요하다.

본 논문에서는 군집분석을 위해 t-SNE (t-Stochastic Neighbor Embedding) [6]를 사용한다. 이 방법은 고차원 데이터를 낮은 차원의 데이터로 차원 축소를 수행하는데 고차원에서의 거리와 저차원에서의 거리가 최대한 일치하는 방향으로 차원 축소를 수행한다. 이러한 유형의 다른 차원 축소 방법과의 차이점은 거리 계산 시 확률적인 개념을 사용한다는 것이다. 우리는 고차원 시계열 데이터를 2차원으로 축소하여 가까운 데이터끼리 군집으로 묶을 것이다. 기존에 t-SNE는 시계열 데이터가 아닌 이미지와 같은 정적 데이터에 적용해 왔다. 이 논문은 시계열 데이터에 t-SNE를 사용하기 위해 동적시간 워핑(DTW: Dynamic Time Warping)를 사용하여 시계열 데이터 사이의 거리를 계산했다. 따라서, 이 논문은 전력 수요예측을 위한 군집분석 이외에도 시계열데이터에 t-SNE를 적용했고 시계열 데이터를 위한 거리를 사용할 경우 t-SNE를 성공적으로 적용 가능하다는 것을 확인했다는 점에서도 의의를 찾을 수 있다.

이 논문은 먼저 수요 데이터 분석에 사용한 t-SNE와 DTW를 리뷰한 다음, 이 방법을 적용한 전력 수요 데이터와 분석결과를 설명할 것이다. 그리고 마지막으로 연구 결과에 대한 요약정리 및 한계점, 향후 연구방향을 제시할 것이다.

II. Preliminaries

1. t-Stochastic Neighbor Embedding

확률적 임베딩(SNE: Stochastic Neighbor Embedding) [7]은 차원축소(dimensionality reduction) 방법의 일종이다. 기본적인 아이디어는 고차원에서의 거리와 축소된 차원에서의 데이터 사이의 거리 혹은 유사도가 호환이 되도록 고차원 데이터를 축소하는 것이다. SNE에서는 유클리디안 거리(Euclidean distance)를 고/저차원에서의 거리 계산에 사용하는 대신 다음과 같은 조건부 확률 $p_{j|i}$와 $q_{j|i}$로 유사도를 표현한다. 식 (1)에서 x_i와 x_j가 고차원의 데이터이고 식 (2)에서 y_i와 y_j가 이에 대응하는 저차원 데이터라고 할 때,

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)} \quad (1)$$

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i}\exp\left(-\|y_i - y_k\|^2/2\sigma_i^2\right)} \quad (2)$$

σ_i는 x_i를 중심으로 하는 가우시안 분포의 표준편차로 분석에 사용되는 데이터에 의해 결정된다. 현재 많은 SNE의 구현에서는 이를 분석자가 직접 입력하는 대신 혼란도(perplexity)를 입력하면 이를 기반으로 주어진 데이터에서 적절한 수준의 표준편차를 결정한다. 혼란도는 $2^{H(i)}$로 $H(i) = -\sum_j p_{j|i}\log p_{j|i}$는 $p_{j|i}$의 새넌 엔트로피(Shannon entropy)이다. 일반적으로 혼란도가 증가하면 표준편차도 증가한다.

SNE에서는 주어진 고차원 데이터 x_i들에 대해서 $p_{j|i}$와 $q_{j|i}$

를 유사하게 만드는 저차원 임베팅 y_i를 찾기 위해서 두 분포 사이의 쿨백-라이블러 발산 (Kullback-Leigbler Divergence)을 최소화한다. y_i는 쿨백-라이블러 발산에 대해 경사강하법 (gradient descent method)을 사용하여 찾는다.

[6]은 SNE가 비대칭적인 조건부 확률을 사용한다는 점과 고차원의 데이터를 저차원으로 임베딩할 경우 고차원에서 멀리 떨어진 데이터를 저차원에서 충분히 멀리 떨어지도록 임베딩하지 못하는 과밀문제 (crowding problem)를 갖고 있다고 지적하였다. 그리고 이에 대한 대안으로 식 (3)과 같은 대칭 구조의 유사도를 고차원에서 사용했다.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2} \tag{3}$$

그리고 과밀문제를 해결하기 위해 저차원에서 식 (4)와 같은 자유도(degree of freedom)가 1인 Student t-distribution을 유사도로 사용하였다.

$$q_{ij} = \frac{\left(1 + \| y_i - y_j \|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \| y_i - y_l \|^2\right)^{-1}} \tag{4}$$

y_i를 찾는 과정에서는 SNE와 유사하게 식 (3)과 식 (4)의 쿨백-라이블러 발산을 최소화한다.

2. Dynamic Time Warping

동적시간 워핑(DTW: Dynamic Time Warping) [8]은 두 개의 시계열 데이터의 거리 혹은 유사도를 측정하는 방법이다. 흔히 두 데이터 사이의 거리를 계산하는데 사용되는 유클리디안 거리(Euclidean distance)를 이용하여 두 시계열 데이터의 거리를 계산할 경우 입력으로 주어지는 두 시계열 데이터의 길이가 일치해야 되는 제약이 있다. 하지만 DTW에서는 이러한 제약이 필요가 없기 때문에 시계열 분석에서 널리 사용되고 있다. DTW의 기본 아이디어는 길이가 다른 두 시계열 데이터를 같게 만드는데 드는 비용을 두 시계열 데이터 사이의 거리로 사용하는 것이다. 두 시계열 데이터를 같게 만들기 위해서는 특정 시점의 데이터를 증가시키거나 감소시켜야 할 뿐만 아니라, 주어진 시계열 데이터들의 길이를 늘리거나 줄일 필요가 있다. DTW에서는 특정 시점의 데이터를 증가, 감소시키는 것과 시계열 데이터의 길이를 증가시키는 것은 허용하지만, 길이를 감소시키는 것은 허용하지 않는다.

$X := (x_1, x_2, ..., x_N)$와 $Y := (y_1, y_2, ..., y_M)$가 DTW의 두 입력 시계열 데이터라고 하자. 그러면 행의 개수가 N, 열의 개수가 M인 행렬 A를 만들 수가 있는데 행렬의 i번째 행과 j번째 열에 해당하는 값 A_{ij}는 x_i와 y_j의 거리가 되며 $|x_i - y_j|$ 혹은 $\sqrt{(x_i - y_j)^2}$가 주로 사용된다. DTW는 A_{11}에서 A_{NM}으로 가는 경로 중 최소 비용의 경로를 계산한 다음 이를 두 시계열 데이터의 거리로 출력한다. 경로는 반드시 인접한 행렬의 엔트리(entry)로만 이동이 가능하며 이동 시 엔트리의 행 또는 열의 인덱스가 감소할 수 없다. 이러한 조건을 만족하는 경로의 비용은 경로가 지나는 엔트리의 값의 합으로 정의한다. 최소 비용 경로를 계산하기 위해서 동적계획법(dynamic programming)을 사용하는데 이 경우 $O(NM)$의 시간 및 공간 복잡도(Time complexity/Space complexity)가 요구된다.

[9]가 제안한 FASTDTW는 최소 비용의 경로를 계산하진 않지만 시간과 공간에 대한 선형 복잡도를 요구하기 때문에 더 빠르게 거리를 계산할 수 있다는 장점이 있기 때문에 매우 많은 시계열 데이터를 분석할 경우 많이 사용되고 있다. 본 논문이 사용한 DTW는 [9]이 사용한 FASTDTW를 사용했으며 특정 시점의 두 시계열 데이터의 값의 차이를 계산할 때 $\sqrt{(x_i - y_j)^2}$를 사용하였다.

III. Experimental Results

1. Electricity Demand Data

이 논문에서 사용한 데이터는 전력거래소에서 수집된 2000년 1월 1일부터 2017년 12월 13일까지 시간대별 전력 수요 데이터이다. 하루의 수요 데이터는 총 24개의 값으로 구성되어 있으며 이는 해당 일의 시간대별 전력 수요를 의미한다. 그림 1은 2000년 12월 3일부터 2000년 12월 9일의 일주일 간의 전력 수요에서 시간대별 수요를 요일별로 측정한 것이다. 이 그림을 보면 화요일부터 금요일까지의 수요는 비슷한 패턴을 보이고 있으며, 월요일, 토요일, 일요일의 패턴은 다른 패턴과 차이를 보이고 있다. 따라서, 요일별로 다른 수요 패턴을 갖는 것을 확인할 수 있다.

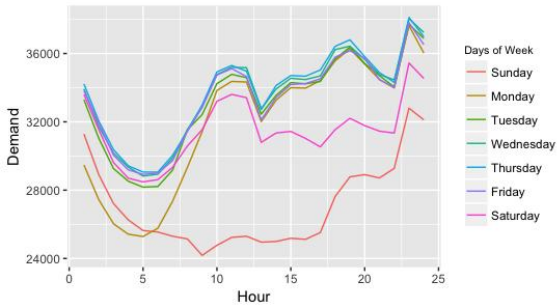

Fig. 1. Electricity demand with respect to days of week

그림 2는 여름과 겨울에 해당하는 일요일과 화요일의 수요이다. 그림에서 확인할 수 있듯이, 같은 요일이라도 계절에 따라 다른 수요 패턴을 보임을 확인할 수 있다.

마지막으로 전력수요에 존재하는 추세성(trend)를 분석하기 위해 서로 다른 해의 동일 날짜에 대한 수요를 그림 3에 표시했다. 그림에서 알 수 있듯이 2001년, 2006년, 2011년, 2016년 4년 동안 동일 날짜 (12월 3일)의 전력수요는 패턴이 다를 뿐만 아니라 증가하는 추세성도 보이고 있다.

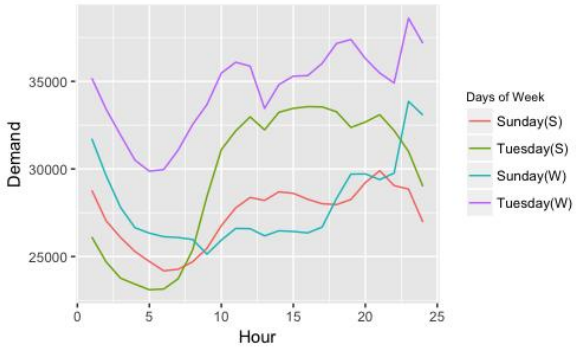

Fig. 2. Electricity demand with respect to seasons

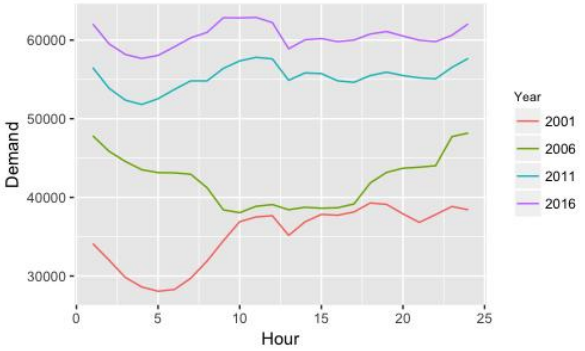

Fig. 3. Electricity demand with respect to years

2. Preprocessing and Experimental Setting

시계열 데이터 분석을 위해서 사용되는 많은 전처리 기법이 존재한다. 전처리의 목적은 분석 결과에 영향을 미치거나 분석의 목적과 무관한 요소를 미리 데이터에서 제거하여 분석결과의 정확성을 높이는 것이다. [3]은 전력 수요 시계열 데이터를 분석하기 위하여 로그변환을 사용하여 시간에 따른 데이터의 변동폭 증가 경향을 제거하였다. [10]과 같은 음성 시계열 데이터 분석에서는 시계열 데이터에서 평균을 빼고 표준편차로 나누어 주는 정규화 과정을 수행한다. 이런 정규화 과정은 모든 데이터를 동일한 스케일(scale)로 처리할 수 있다는 장점이 있기 때문에 특히 신경망을 이용한 지도학습 및 비지도학습 연구에서 널리 사용되고 있다. 이 논문에서는 매일의 시계열 데이터를 해당 일자의 수요의 평균과 표준편차를 계산하여 정규화시켰다. 정규화 과정으로 추세성과 시간에 따른 데이터 변동폭의 증가와 같은 요인을 제거하고 하루 내에 발생하는 수요 변화의 패턴만을 고려하는 것이 가능하다.

t-SNE로 분석을 하기 위해서는 혼란도(perplexity), 반복횟수(number of iteration), 학습율(learning rate)을 정의해야 한다. [6]에 의하면 5에서 50 사이의 혼란도를 권장하였지만 실험결과로 얻게 되는 클러스터의 개수에는 큰 영향을 미치지 않았으며 200의 혼란도를 사용할 경우에 클러스터들 사이의 거리가 가장 뚜렷하게 나타났다. 그리고 반복횟수는 1,000회, 학습율(learning rate)은 100으로 설정하였다. 그리고 DTW에서 특정 시점에서의 데이터의 차이는 유클리디언 거리를 사용하였다.

3. Experimental Results and Analysis

그림 4는 t-SNE를 수행하여 얻은 군집분석 결과이다. 4개의 군집을 관찰할 수 있었다.

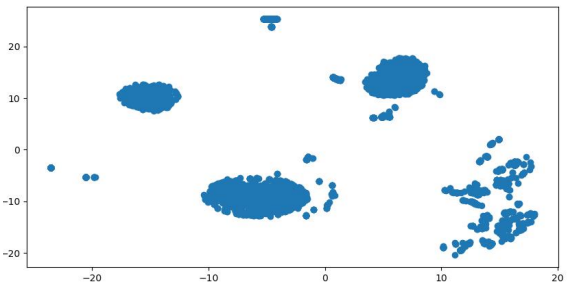

Fig. 4. Result of t-SNE

다음으로 각 군집이 어떤 데이터로 구성되어 있는지를 알아보자. 이를 위해 그림 4에서 나타나는 군집들을 그림 4의 좌표를 기준으로 표 1과 같이 정의하였다.

Table 1. Definitions of clusters

Cluster No.	x coordinate	y coordinate	Number of Data
Cluster 1	(-11, 3)	(-15,0)	1,874
Cluster 2	(10,20)	(-25,5)	1,745
Cluster 3	(0,10)	(0,20)	1,553
Cluster 4	(-20,-10)	(5,15)	900

각 군집에서의 측정 데이터들을 월별·요일별로 분류한 결과는 표 2와 표 3과 같다. 표 2에서 군집 1, 3, 4에는 6월에서부터 8월에 해당하는 여름 계절 데이터가 다른 월에 비해 많이 나타났다. 반면에 군집 2에서는 11월부터 3월에 해당하는 겨울 데이터가 다른 월에 비해 많이 나타났다. 하지만, 월별 데이터 분포에서 확인할 수 있듯이 이러한 계절적 특성만이 두드러지게 나타나는 것은 아니고 계절성 이외의 요인에 의한 전력 수요 패턴도 함께 섞여 있다는 것을 확인할 수 있다.

표 3에서는 각 군집에 포함된 데이터의 요일별 특성을 살펴볼 수 있는데 군집 1과 3에서는 요일별 특성이 잘 보이지 않는다. 반면에 군집 2의 경우에는 토요일과 일요일에 해당하는 휴일의 데이터가 다른 요일에 비해 많이 나타났다. 그리고 군집 4는 군집 2와는 반대로, 휴일 데이터보다 평일의 데이터가 훨씬 많이 나타났다.

Table 2. Data distribution with respect to months

Month	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Jan.	115	309	86	31
Feb.	139	295	36	35
Mar.	143	301	55	52
Apr.	123	180	132	42
May.	181	55	108	90
Jun.	218	1	108	148
July.	222	0	183	112
Aug.	232	0	194	88
Sep.	194	22	145	131
Oct.	179	70	167	109
Nov.	80	209	182	46
Dec.	48	303	147	16

Table 3. Data distribution with respect to days of week

Day	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Sun.	381	550	6	0
Mon.	74	135	86	442
Tues.	315	163	252	138
Wed.	320	174	248	104
Thu.	308	156	269	137
Fri.	354	162	283	79
Sat.	122	405	409	0

표 3에서 군집 3는 특별한 요일적 특성이 보이지 않았지만 이를 계절성과 함께 분석할 경우, 특별한 패턴을 보임을 확인할 수 있었다. 표 4는 계절성과 요일적 특성을 함께 분석한 결과로 군집 3에 속한 데이터 중 여름에 해당하는 7월, 8월과 겨울에 해당하는 12월, 1월의 요일별 데이터 분포이다. 표 4에서 확인할 수 있듯이 겨울에 해당하는 12월, 1월에는 토요일에 해당하는 데이터의 수가 평일에 해당하는 데이터의 수보다 훨씬 적은 비율을 차지하고 있다. 반면에 여름에는 토요일에 해당하는 데이터의 비율이 크게 상승했다.

표 5는 표 3의 군집 1에 속한 데이터 중 여름에 해당하는 7월, 8월과 겨울에 해당하는 12월, 1월의 요일별 데이터 분포이다. 군집 1의 경우 일요일 데이터가 여름에 집중적으로 분포되어 있고 나머지 계절에서는 일요일 데이터보다는 평일 데이터가 더 많이 분포하는 것을 확인할 수 있었다. 따라서 각 군집은 전력 수요를 결정하는 여러 요인들이 복합적으로 작용하고 있으며 그 작용 메커니즘도 군집별로 다양하게 나타난다는 것을 알 수 있다.

Table 4. Comparison of Data distributions for July., Aug., Dec., and Jan. for cluster 3

Day	July	Aug	Dec	Jan
Sun.	1	2	0	0
Mon.	0	2	9	13
Tues.	21	28	32	23
Wed.	24	34	31	18
Thu.	24	30	37	22
Fri.	40	35	38	20
Sat.	75	63	0	0

Table 5. Comparison of Data distributions for July., Aug., Dec., and Jan. for cluster 1

Day	July	Aug	Dec	Jan
Sun.	80	77	0	0
Mon.	5	10	6	7
Tues.	37	34	10	22
Wed.	30	32	12	26
Thu.	34	28	8	24
Fri.	29	35	7	28
Sat.	7	16	5	8

지금까지의 분석을 통해서 t-SNE를 적용한 결과 나타난 군집은 단일 특성으로는 설명이 불가능한 패턴이 존재했으며, 요일별·계절별 특성 이외의 요인에 의해서도 패턴이 결정됨을 확인할 수 있었다. 따라서, t-SNE로 얻은 군집들에 다시 t-SNE를 적용하여 군집 분석을 수행하였다.

그림 5는 표 1에서 소개한 4개의 군집에 대해 각각 t-SNE를 적용한 결과이다. 더 세밀한 군집을 찾기 위해서 이전 t-SNE에서 사용한 혼란도(200)보다 더 적은 값(5)를 사용하였다. 그림 5에서 확인할 수 있듯이 각각의 군집은 다시 더 작은 단위의 군집으로 나뉘는 것을 알 수 있다. 따라서 계층적으로 계속 t-SNE를 적용할 경우 다양한 수요 패턴이 나오는 것을 기대해 볼 수 있다. 하지만 그림 5와 같은 2단계 계층에서의 군집 분석 결과는 계절성과 요일적 특성만으로는 수요 패턴의 설명을 할 수가 없다. 따라서, 하위 계층의 군집 분석 결과로 찾은 군집의 패턴을 이해하기 위해서는 계절과 요일 이외의 다른 정보(예를 들어, 기상 정보, 연도)을 활용하여 수요 패턴을 이해할 필요가 있다.

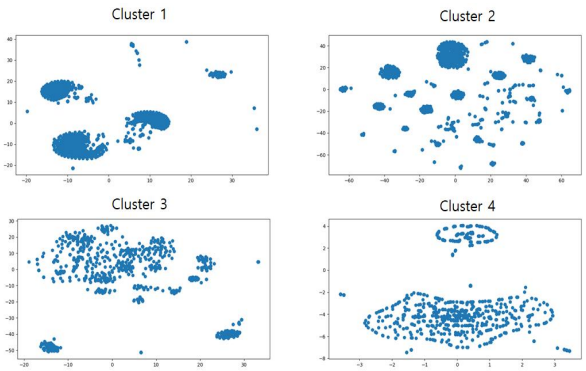

Fig. 5. Result of 2nd level t-SNE

특히 그림 5의 군집 2의 경우에는 규모가 큰 군집 이외에 적은 수의 데이터로 구성된 군집들도 나타났다. 따라서, 계층적인 군집 분석을 시도할 경우, 혼란도를 적절한 수준으로 감소시키는 방법의 개발 또한 필요해 보인다.

IV. Conclusions

본 논문은 전력 운영에서 중요한 역할을 담당하는 전력 수요 예측의 정확도를 높이는 하나의 방안인 전력 수요패턴 분석을 다룬다. 이를 위해 전력거래소에서 수집된 2000년 1월 1일부터 2017년 12월 13일까지 시간대별 전력 수요 데이터를 대상으로 군집 분석을 수행하였다. 군집 분석은 t-SNE를 이용하였고 시계열 데이터에 이 방법을 적용하기 위하여 동적시간 와핑을 적용하였다.

군집분석 결과로 얻는 각각의 군집은 계절성과 같은 특정 성질이 단독으로 포함되기 보다는 여러 성질이 복합적으로 있는 것으로 분석되었다. 하지만, 한 번의 군집분석으로 얻은 군집들을 얻은 이후, 개별 군집에 대해 다시 군집분석을 수행할 경우, 더 세밀한 군집으로 나누는 것이 가능하였다. 이렇게 계층적인 군집분석을 수행할 경우, 전력수요의 패턴을 더 세밀하게 분석하는 것이 가능할 것으로 기대된다.

이 논문의 후속연구로 계층적 군집분석 방법을 적용하여 최대 몇 개까지의 군집으로 전력수요 데이터를 나눌 수 있는지를 확인하고 이렇게 얻은 군집에 대한 의미를 밝히는 것이 필요하다. 뿐만 아니라, 하루 단위의 전력수요를 나누는 기준을 다르게 적용할 필요도 있을 것으로 판단된다. 예를 들어, 하루의 정의를 해당일의 0시부터 24시로 정하는 것 대신에 오전 6시부터 다음날의 오전 6시까지를 하루의 정의로 사용할 경우의 분석 등이 가능할 것으로 생각한다. 그리고 시간이 지날수록 수요의 패턴이 달라질 수 있기 때문에 시간의 흐름에 따라 패턴이 달라지고 있는지를 확인하는 것도 중요하다. 전력 수요에 대한 군집 분석의 목적은 전력 수요의 패턴을 이해하고 이를 이용하여 더 정확한 수요예측 모델을 만드는 것이다. 따라서 각 군집에 대한 예측 모델을 구축하고 그 정확도를 측정하는 것 또한 필요하다.

REFERENCES

[1] Korea Power Exchange, http://www.kpx.or.kr
[2] P.J. Brockwell and R.A. Davis, "Introduction to Time Series and Forecasting" Springer, 2016.
[3] D. Park, S.H. Yoon, "Clustering and classification to characterize daily electricity demand," Journal of the Korean Data & Information Science Society, Vol. 28, No. 2, pp. 395-406, March 2017.
[4] J.H. Lim, S.Y. Kim, J.D. Park, K.B. Song, "Representative temperature assessment for improvement of short-term load forecasting accuracy," Journal of the Korean Institute of Illuminating and Electrical Installation Engineers, Vol. 27, No. 6, pp. 39-43, June 2013.
[5] S.H. Yoon, Y.J. Choi, "Functional clustering for electricity demand data: A case study," Journal of the Korean Data & Information Science Society, Vol. 26, No. 4, pp. 885-894, July 2015.
[6] L.J.P. van der Maaten and G.E. Hinton, "Visualizing high-dimensional data using t-SNE," Journal of Machine Learning Research, Vol. 9, pp. 2579-2695, Nov 2008.
[7] G.E. Hinton and S.T. Roweis, "Stochastic neighbor embedding," Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 833-840, 2002.
[8] J.Kruskal and M. Liberman, "The symmetric time warping problem: From continuous to discrete," Proceedings of Time Waprs, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, pp. 125-161, 1983.
[9] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," Intelligent Data Analysis, Vol. 11, No. 5, pp. 561-580, Oct. 2007.
[10] N.V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using Bayesian framework," Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 156-161, Dec. 2013.

Authors

Yunhong Min received the B.S. degree in Industrial Engineering & Management Science from POSTECH, and Ph.D. in Industrial Engineering from Seoul National University in 2006 and 2012, respectively. Dr. Min joined the faculty of Graduate School of Logistics at Incheon National University in 2017. Before joining Incheon National University, he was a research staff member of Samsung Advanced Institute of Technology. He is interested in mathematical optimization and machine learning and their applications to logistics and supply chain management.