The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.2.319
32(2), 319–330

Prediction of electricity consumption in A hotel using ensemble learning with temperature

Jaehwi Kim[a] · Jaehee Kim[a,1]

[a]Department of Statistics, Duksung Women's University

(Received January 18, 2019; Revised February 15, 2019; Accepted February 15, 2019)

Abstract

Forecasting the electricity consumption through analyzing the past electricity consumption a advantageous for energy planing and policy. Machine learning is widely used as a method to predict electricity consumption. Among them, ensemble learning is a method to avoid the overfitting of models and reduce variance to improve prediction accuracy. However, ensemble learning applied to daily data shows the disadvantages of predicting a center value without showing a peak due to the characteristics of ensemble learning. In this study, we overcome the shortcomings of ensemble learning by considering the temperature trend. We compare nine models and propose a model using random forest with the linear trend of temperature.

Keywords: ensemble learning, temperature, bagging, random forest, time series forecast

1. 서론

오늘날 전기에너지는 일상생활에서 손쉽게 접할 수 있는 에너지 중 하나이다. 최근에는 전기 자동차의 개발로 인해 운송 수단에서까지 전기에너지가 사용되고 있다. 이렇듯 전기에너지는 다른 에너지와 비교하여 다양한 분야에서 사용되고 있는 에너지라고 할 수 있다. 대체로 전기에너지는 공급과 수요가 동시에 일어나는 특징이 있다. 즉, 공급의 불안정은 전력수요의 불안정을 필연적으로 일으킬 수밖에 없다. 또한, 세계적으로 환경문제로 인한 온실가스 감축 압력이 증가하고, 원전에 대한 지역 주민들의 공포와 안전 요구가 증가하는 등 다양한 갈등이 존재한다. Lee (2015)에서는 이러한 문제로 인해 주요 선진국을 중심으로 에너지 정책을 공급 중심에서 수요관리 중심으로 빠르게 전환이 되는 추세라고 언급한다. 전력수요를 관리하는데 가장 기본이 되는 연구는 바로 전력 소비 패턴을 이해하고, 전력소모량을 예측하는 것이다. 미래의 전력소모량을 정확하게 예측할 수 있다면 그에 맞는 전력 생산을 통해 안정적인 공급을 할 수 있을 뿐만 아니라 불필요한 전력을 생산하여 낭비하는 경우를 줄일 수 있다. 또한, 소비자들의 관점에서 경제적으로 전기에너지를 이용하기 위해 소비 계획을 세울 수 있고, 전기 요금 절감으로 이윤을 창출할 수 있다.

This research was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP), Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20161210200610) and the Korea Electric Power Corporation (Grant number: R18XA01).
[1]Corresponding author: Department of Statistics, Duksung Women's University, 419 Samyang-ro 144 Gil 33, Dobong-Gu, Seoul 01369, Korea. E-mail: jaehee@duksung.ac.kr

국가간의 교류도 활발해지고, 전세계적으로 인구의 고령화 현상이 일어나면서 관광산업은 주목받고 있는 산업 중 하나이다. 우리나라 역시 2018년 2월 평창동계올림픽을 개최하는 등 관광산업 육성에 관심을 가지고 있으며, Yun (2018)에 의하면 문화체육관광부도 새 정부의 문화정책의 프레임을 '사람이 있는 문화'로 정하고 그에 맞는 정책과제들을 준비하고 있다. 또한 Choi (2018)에서는 시대적 변화 상황에 맞춰 지방자치단체들의 관광 정책 개발와 효과적인 관광전략을 제시하며 관광산업이 지역 경제의 성장 동력이 될 것이라고 주장한다. 이에 따라 호텔과 같은 숙박업에 대한 관심 또한 높아지며, 숙박시설과 서비스에 대한 개선을 중요시하게 되었다. 숙박시설에 있어 전기, 물 등의 에너지 사용에 대한 비용 절감은 경영에 있어 큰 이득을 가져오고, 다른 부분에 투자할 수 있는 기회를 제공한다. 그러므로 에너지 사용량에 대한 정확한 예측은 숙박시설을 운영함에 있어 중요한 요소라고 할 수 있다.

미래의 전력소모량을 예측하는 연구는 이전부터 다양한 방법으로 시행됐으며, 요즘 자주 사용되는 예측 방법은 기계학습(machine learning)이다. 기계학습은 특정 자료들을 알고리즘을 통해 컴퓨터가 학습할 수 있도록 하는 기법으로 학습 결과를 토대로 분류 및 예측하는 분석 방법이다. 요즘 알파고로 인해 대중들에게 잘 알려진 딥 러닝(deep learning) 뿐만 아니라 서포트 벡터 머신(support vector machine; SVM), 앙상블 학습(ensemble learning) 등을 기계학습의 예로 들 수 있다. 최근 관련 해외 논문으로 Massidda와 Marrocu (2018)는 랜덤 포레스트(random forest)와 선형 회귀에 기반한 하이브리드 기계학습 방법을 제시하여 전력소모량을 예측했다. 또한 Shi 등 (2018)에서는 a novel pooling-based deep RNN (PDRNN)을 제안하며 가정용 전력소모량을 예측했다. 국내에서도 기계학습을 이용한 전력소모량 예측이 활발히 이루어지고 있다. Park 등 (2017)은 호주의 National Electricity Market (NEM) 데이터를 SLPN, MLPN, CNN, RNN 모형을 사용하여 예측 정확도를 비교하였으며, 전처리 과정을 거치지 않은 데이터와 전처리 과정을 거친 데이터를 비교 실험하여 비교 측정하였다. Shin과 Kim (2016)은 R과 텐서플로우(Tensorflow)를 이용하여 딥 러닝 기반 전력수요예측 방법에 관해 연구하였으며 평균온도, 최저온도, 최고온도, 불쾌지수, 체감온도, 냉방도시 등 다양한 전력수요 예측요소들을 딥 러닝의 입력 요소로 사용했다. Ahn 등 (2017)은 DNN 모형을 이용하여 동계 전력수요예측을 하였으며 이 또한 여러 가지 기상요소를 사용하여 전력수요를 예측하였다. Tak 등 (2016)은 SVM에 회귀 분석 함수를 적용한 support vector regression (SVR) 모형으로 기상정보를 취합하여 단기 전력수요예측 모형을 만들었다.

국내의 선행 연구들을 살펴보면 대체로 기상정보를 활용하여 전력소모량을 예측하는 모형을 제시한다. 이와 다르게 Grmanova 등 (2016)은 기상정보 없이 전력소모량 데이터만을 가지고 앙상블 학습 기법을 통하여 미래의 전력소모량을 예측하는 방법을 보였다. 본 논문에서는 이 논문과 Laurinec (2017)의 방법을 도입하여 앙상블 학습을 통해 A 호텔의 일별 전력소모량을 예측했다. 하지만 일별 전력소모량 예측에 있어 앙상블 학습으로는 피크(peak) 예측에서는 불리한 면을 보였다. 그래서 평균기온 변수를 도입하여 선형회귀모형(linear regression model; LM)을 적합하고, 회귀모형의 잔차부분을 앙상블 학습을 통해 적합함으로써 예측의 정확도를 높였다.

본 논문의 구성은 다음과 같다. 2장에서는 전력소모량과 평균기온 데이터에 관해 설명과 패턴에 대한 특징을 언급한다. 3장에서는 본 논문에서 사용한 앙상블 학습과 그 방법에 관해 설명하고, 4장에서는 기온변수를 이용한 선형회귀모형에 관해 설명한다. 5장에서는 모형 설정과 그에 따른 분석결과를 언급 및 비교한다. 6장에서는 분석결과에 따른 결론을 내린다.

2. 전력소모량 및 온도 데이터

2.1. A 호텔 전력소모량

본 논문에서 사용한 데이터는 한국전력공사(KEPCO)의 전력 포털사이트 'i-smart'에서 제공받았으며,

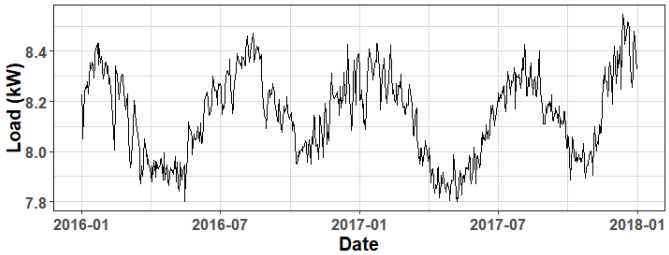


Figure 2.1. Daily electricity consumption data of A hotel from 2016.1.1 to 2017.12.31.

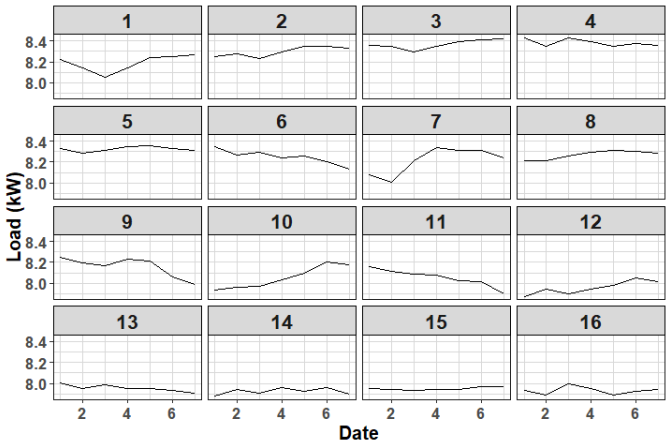


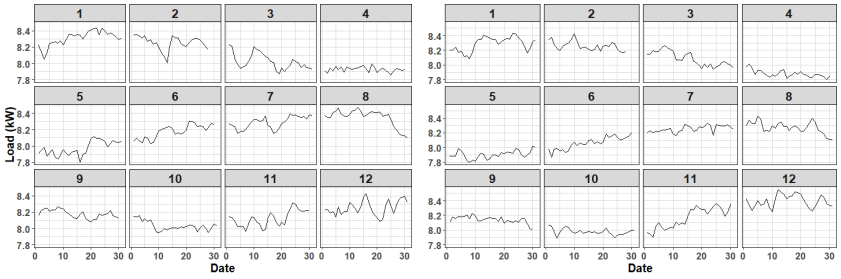
Figure 2.2. Comparison of weekly consumptions of A hotel during the first 16 weeks (2016.1–2016.4).

A 호텔의 2016년 1월 1일부터 2017년 12월 31일까지 일별 전력소모량 데이터이다. A 호텔은 대한민국 경기도 안산시 상록구에 위치하고 있으며 객실 수가 70개인 3성급 호텔이다. 연간 데이터를 365일로 동일하게 맞춰주기 위해서 2016년 2월 29일 데이터는 제거했다. 또한 2016년 2월 22일부터 2월 26일, 7월 8일부터 7월 10일은 결측값 또는 관측의 오류로 판단되는 값이기 때문에 Moritz와 Bartz-Beielstein (2017)에서 제안한 스플라인 보간법(spline interpolation)을 이용하여 그 값들을 추정했다. 또한 모형이 잘 추정하는지 비교 대상이 될 시험 데이터(test data)로써 2018년 1월 1일부터 30일까지 동일 장소에서의 전력소모량을 사용했다. 그리고 전력소모량을 로그변환하여 값의 단위를 낮추고 분산을 안정화했다.

Figure 2.1을 보면 A 호텔의 전력소모량 데이터는 여름과 겨울에 소모량이 많고 봄과 가을에 소모량이 적은 패턴을 가지고 있다. Figure 2.2와 Figure 2.3에서는 특별히 일주일이나 한달을 주기로 반복되는 패턴은 보이지 않는 것을 확인할 수 있다.

2.2. 수원시 평균기온

A 호텔은 안산시에 위치하고 있지만 기상청에서 제공하는 평균기온 데이터의 경우 안산시는 누락되어 있기 때문에 안산시와 가장 가까운 수원시의 일별 평균기온 데이터를 사용했다. 수원시를 비롯한 대한민국의 기후는 4계절이 뚜렷한 기후이다. 여름에는 기온이 30°C가 넘는 더운 날씨이며, 겨울에는 영하로 내려가는 매우 추운 날씨이다. 이러한 패턴은 A 호텔의 전력소모량의 패턴과 비교했을 때 약 2배의



(a) Cosumptions in 2016 (b) Consumptions in 2017

Figure 2.3. Comparing with monthly consumptions of A hotel in 2016 and 2017.

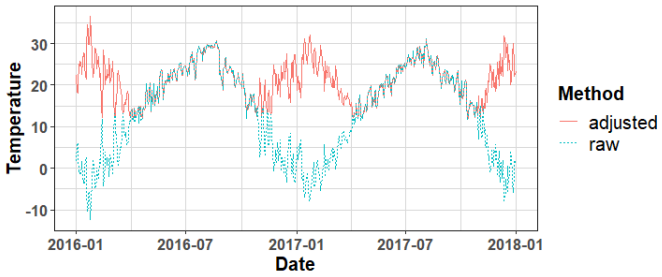


Figure 2.4. Daily average temperature and adjusted temperature in Suwon city.

주기 차가 발생한다. 그러므로 평균기온 데이터를 식 (2.1)을 통해 조정하여 A 호텔의 전력소모량과 주기가 같게 맞춰주었다.

$$\text{adj\_temp}_t = \sqrt{(\text{temp}_t - \text{temp}_{\text{ave}})^2} + \text{temp}_{\text{ave}}, \tag{2.1}$$

여기서 temp_t은 일평균 기온이고, t는 시계열의 시점이며 temp_{ave}은 1981년부터 2010년까지의 연평년이다. Figure 2.4를 보면 주기가 대략적으로 절반으로 감소한 것을 볼 수 있다.

3. 앙상블 학습을 이용한 전력소모량 데이터 분석

3.1. 의사결정나무

의사결정나무(decision tree)는 데이터를 일정 규칙에 의해 분류하여 예측하는 기법으로 그 과정이 나무와 유사하여 의사결정나무라 이름 지어졌다. 의사결정나무는 반응변수의 전체 집합인 뿌리마디(root node)로부터 설명변수들의 일정한 규칙에 의해 중간마디(intermediate node)와 끝마디(terminal node)로 나누어진다. 끝마디는 잎(leaves)이라고도 불리며, 마디 사이의 연결을 가지(branches)라고 부른다. 의사결정나무의 특징은 각각의 잎들이 서로 교집합을 가지지 않는다는 것이다. 즉 각 잎들에 속하는 데이터의 갯수를 더하면 뿌리마디의 데이터 수와 일치한다. 의사결정나무는 반응변수가 어떤 형태의 데이터인지에 따라 크게 분류나무(classification tree)와 회귀나무(regression tree)로 분류할 수 있다. A 호텔의 전력소모량은 연속형 변수이기 때문에 본 논문에서는 회귀나무에 대해 다룬다. 회귀나무의 경우에는 잎으로 분류된 반응변수들의 평균을 예측값으로 반환한다. 즉 잎의 갯수에 따라 도출하는 값의 수가 결정된다. 본 논문에서는 Therneau와 Atkinson (1997)에서 제안한 recursive partitioning regression trees (RPART) 모형과 Hothorn 등 (2010)에서 제안한 conditional inference trees

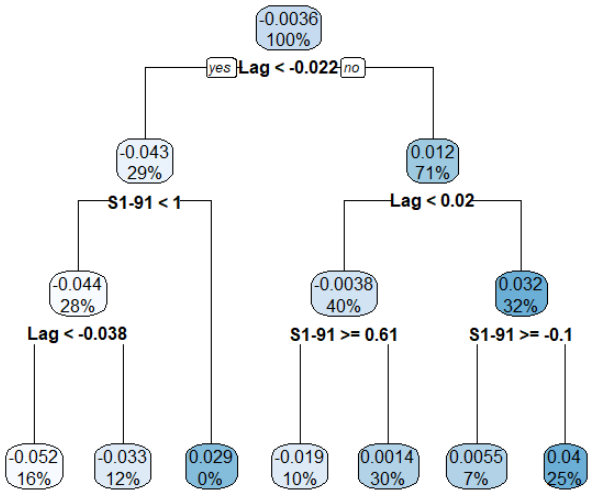


Figure 3.1. Regression tree example.

(CTREE) 모형을 사용했다. RPART 모형은 간단한 비모수적 방법인 반복 분할(recursive partitioning)을 사용한 의사결정나무로 기존의 의사결정나무보다 더 직관적인 모형을 생성하고, 더 정확할 수 있다. CTREE는 반복 분할을 조건부 추론의 틀에 맞춘 의사결정나무로 기존의 의사결정나무보다 과적합(overfitting)의 문제를 해결한다. Figure 3.1은 전력소모량 데이터로부터 얻어낸 설명변수들(Lag, S1–91 등)로 RPART 모형을 적합하여 로그변환된 A 호텔의 전력소모량을 추정한 것을 보여준다. 각 마디에서 보이는 % 값은 각 마디로 분류된 데이터가 전체의 몇 %인지 보여주는 값이다. 잎들의 값을 전부 합하면 100%가 됨을 알 수 있는데 이는 의사결정나무의 잎들이 서로 교집합을 가지지 않는다는 특징을 잘 나타낸다. % 위에 적힌 값은 각 마디에서 로그 변환된 전력소모량의 추정값을 나타낸다. 잎에서 나타나는 추정값은 새로운 데이터로 예측을 할 때 각 값에 따른 예측값이 된다. 각 마디의 하단에 나타나는 규칙은 각 마디에서 하위 마디로 분류할 때 사용된 규칙을 나타낸다. Figure 3.1에서는 간단하고 가시적인 예시를 위해 깊이가 3인 의사결정나무를 만들었지만, 실제 분석에서는 깊이가 26–30인 의사결정나무를 생성하여 값을 더욱 정교하게 분류했다.

3.2. 피쳐 엔지니어링

피쳐 엔지니어링(feature engineering)은 머신러닝 모형의 성능을 높이거나 머신러닝 모형에 입력할 데이터를 생성하기 위해 데이터에 대한 지식을 활용하여 초기 데이터를 가공하고 특징(feature)을 만들어내는 과정이다. 피쳐 엔지니어링은 모형 성능에 미치는 영향이 크기 때문에 머신러닝 응용에 있어서 굉장히 중요한 기법이다. 피쳐 엔지니어링에는 특징 선택(feature selection), 특징 추출(feature extraction), 특징 생성 또는 구축(feature generation or construction) 등이 있다. 여기에서는 R 프로그램 *forecast* 패키지의 *fourier* 함수를 이용하여 전력소모량의 주기를 사인과 코사인 곡선으로 나타냈고, 그 값을 의사결정나무의 설명변수로써 사용했다. 또한 전력소모량의 계절 특성을 설명변수로 사용하여 의사결정나무가 잘 예측할 수 있도록 했다. 하지만 이러한 값들을 설명변수로 사용하게 되면 전력소모량의 추세를 반영하지 못하게 된다. 그러므로 Cleveland 등 (1990)에서 제안한 seasonal and trend decomposition using Loess (STL)을 이용하여 시계열 데이터를 각 성분으로 분리한 후 추세 부분은 다른 모형을 이용하여 예측했다. 이 부분에 대한 설명은 3.4절에서 다시 다루겠다.

3.3. 앙상블 학습

앙상블 학습은 베이스 모형(base model)들의 집합을 사용하는 접근법이다. 하나의 베이스 모형을 사용하는 경우 과적합될 가능성이 있으며, 분산이 높아 훈련 데이터(train data)에는 잘 적합하지만 예측력은 매우 떨어질 수 있다. 이를 보완하는 방법으로 Grmanova 등 (2016)에서는 앙상블 학습을 제안한다. 앙상블 학습은 각각의 베이스 모형들을 훈련 데이터를 이용하여 적합하고 그 모형들의 대푯값을 취하는 기법이다. 그러므로 과적합의 가능성을 줄이고 분산을 줄여 예측력을 높일 수 있다. 앙상블 학습은 크게 두 가지로 나눌 수 있는데, 베이스 모형이 같은 형태인 경우 동질적 학습(homogeneous learning), 다른 형태인 경우 이질적 학습(heterogeneous learning)이라고 한다. 본 논문에서 사용한 방법은 의사결정나무를 베이스 모형으로 사용하는 동질적 학습으로, 배깅(Bagging: bootstrap aggregation)과 랜덤 포레스트를 사용한다.

3.3.1. 배깅 배깅은 붓스트랩 샘플링(bootstrap sampling)을 이용한 앙상블 기법이다. 의사결정나무는 대체로 높은 분산을 가진다는 문제점이 있다. 심지어 훈련 데이터를 임의로 두 집단으로 나눠 각각 의사결정나무 모형을 적용했을때 두 모형의 결과는 매우 상이할 수도 있다. 이를 보완하기 위해 붓스트랩 기법을 적용한 방법이 바로 배깅이다. 배깅은 하나의 훈련 데이터로부터 여러 개의 붓스트랩 샘플을 생성하고 각 샘플을 다른 의사결정나무에 적합한다. 이렇게 나온 각 의사결정나무의 결과들을 모아 평균을 구함으로써 분산을 줄일 수 있다. 이 원리는 마치 분산이 σ^2인 모집단으로부터 n개의 표본을 뽑았을 때 표본평균의 분산은 σ^2/n의 되는 것과 같은 이치이다. 다만 일반적으로 한 번에 여러 개의 훈련 데이터를 가지고 분석을 할 수 없기 때문에 붓스트랩 기법을 사용하는 것이다. James 등 (2013)에서는 배깅으로 얻어진 결과 값을 $\hat{f}_{\text{bag}}(x)$라 할 때, 배깅을 식 (3.1)과 같이 표현한다.

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x), \tag{3.1}$$

여기서 B는 붓스트랩 샘플의 갯수이며, $\hat{f}^{*b}(x)$는 b번째 붓스트랩 샘플을 의사결정나무에 적용하여 얻은 값이다.

3.3.2. 랜덤 포레스트 배깅은 분산을 줄일 수 있는 뛰어난 방법이지만 한 가지 문제점이 발생할 수 있다. 의사결정나무의 설명변수 중에서 다른 변수들보다 영향력이 매우 큰 변수가 존재한다면, 의사결정나무를 여러 번 반복하더라도 영향력이 큰 변수로 먼저 분리하는 나무만 생성 될 수 있다. 이 경우에는 아무리 많은 의사결정나무 모형이 있다 하더라도 그 모형들로부터 얻은 결과값은 유사할 수밖에 없다. 왜냐하면 의사결정나무의 특성상 초반 분류 방법에 따라 나무의 형태가 어느정도 윤곽이 잡히기 때문이다. 그러므로 설명변수 전부를 사용하지 않고 몇 개만 사용하여 의사결정나무 모형을 만듦으로 모든 나무 모형이 비슷한 경우를 방지할 수 있다. 이러한 알고리즘을 랜덤 포레스트라고 한다. 랜덤 포레스트는 설명변수가 p개 있다고 가정할 때 통상 분류나무에서는 $\sqrt{p}$개, 회귀나무에서는 $p/3$개의 변수만을 사용하여 의사결정나무를 적용한다. 그러므로 영향력이 매우 큰 변수가 존재하더라도 의사결정나무들이 전부 유사한 결과를 가지는 경우를 어느정도 배제할 수 있다.

3.4. 추세 추정

의사결정나무의 설명변수만으로는 전력소모량의 추세를 예측하기에는 어려움이 있다. 그러므로 전력소모량의 추세 부분을 따로 분류하여 autoregressive integrated moving average (ARIMA) 모형과 홀트-윈터스(Holt-Winters) 모형을 이용해 예측했다.

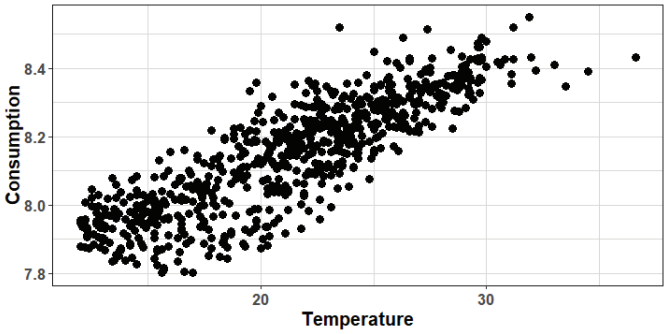


Figure 4.1. Scatter plot between log-transformed electricity consumption and adjusted temperature.

3.4.1. ARIMA 모형 만약 시계열 데이터 $\{y_t\}$가 자기회귀과정(autoregressive; AR)과 이동평균과정(moving average; MA)을 혼합한 ARMA(p,d) 모형을 따른다고 할 때 Cowpertwait와 Metcalfe (2009)에서는 y_t을 식 (3.2)와 같이 나타낸다.

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + w_t + \sum_{j=1}^{q} \beta_j w_{t-j}, \tag{3.2}$$

여기서 p와 q는 차수, α와 β는 모형의 모수(parameter)이고 $\{w_t\}$는 백색잡음(white noise)이다. y_t를 d번 차분한 $\bigtriangledown^d y_t$가 ARMA(p,q) 모형을 따르면 y_t는 ARIMA 모형을 따른다고 하며 ARIMA(p,d,q)로 표기한다.

3.4.2. 홀트-윈터스 모형 홀트-윈터스 모형은 지수평활법(exponential smoothing)의 일종으로 가법과 승법으로 나뉜다. Cowpertwait와 Metcalfe (2009)에서는 y_t를 시계열 데이터라고 할 때, 승법 홀트-윈터스 평활법을 식 (3.3)과 같이 나타낸다.

$$\begin{aligned} a_t &= \alpha \left(\frac{y_t}{S_{t-p}} \right) + (1-\alpha)\left(a_{t-1} + b_{t-1}\right), \\ b_t &= \beta \left(a_t - a_{t-1}\right) + (1-\beta) b_{t-1}, \\ S_t &= \gamma \left(\frac{y_t}{a_t} \right) + (1-\gamma) S_{t-p}, \\ \hat{y}_{t+k|t} &= (a_t + k b_t) S_{t+k-p}, \quad k \le p, \end{aligned} \tag{3.3}$$

여기서 p는 주기(period), a_t는 수준(level), b_t는 기울기(slope), S_t는 계절효과(seasonal effect)이고, α, β, γ는 각각의 평활 모수(smoothing parameter)이다.

4. 기온변수를 고려한 앙상블 학습

4.1. 기온변수를 이용한 선형회귀모형

A 호텔의 일별 전력소모량 데이터는 앞서 언급했듯이 여름과 겨울에 높고 봄과 가을에는 낮은 패턴을 가지고 있다. 전력소모량의 주기를 설명변수로 사용하여 의사결정나무를 모델링하고 배깅과 랜덤포레스트를 통해 예측력을 높였지만 약 반년정도의 길이를 가지는 주기로 정확한 전력소모량 값을 예측하기

에는 몇가지 어려움이 있었다. 가장 두드러지는 문제점은 예측된 값들이 중심편향적이라는 것이다. 즉 전체적으로 오차는 줄일 수 있지만 전력소모량의 그래프 형태가 최대 또는 최소를 향해 역동적으로 나타내지 못하고 평균값으로 부드럽게 그려졌다. 그러므로 본 논문에서는 비슷한 패턴을 보이는 수원시의 평균기온을 이용해 좀 더 정확하게 예측하는 새로운 방법을 제안한다. 데이터 부분에서 설명한 조정된 기온 변수 adj_temp_t를 이용하여 전력소모량과 선형회귀모형을 적합함으로써 중심편향적 문제를 해결했다. 로그변환된 A 호텔의 전력소모량과 조정된 기온 간의 상관계수는 0.86으로 매우 높은 상관관계를 가지고 있다. 또한 Figure 4.1에서 확인할 수 있듯이 선형모형을 적합하기에도 아주 좋은 데이터 형태를 취하고 있다. 이에 따라 식 (4.1)과 같이 선형회귀식을 세우고 모형을 적합했다.

$$\text{LM} : \text{ec}_t = 7.538 + 0.029 \times \text{adj\_temp}_t + \xi_t, \tag{4.1}$$

여기서 ec_t는 로그변환된 일별 전력소모량이고, ξ_t은 회귀모형의 잔차이다. ξ_t는 앙상블 학습 기법 + ARIMA 또는 앙상블 학습 기법 + 홀트-윈터스 모형으로 추정했다. 이 부분에 대한 자세한 내용은 5.1장 모형 설정에서 다루도록 한다.

5. 모형 설정과 분석 결과

5.1. 모형 설정

본 논문에서는 크게 두 가지 방법으로 모형을 설정하였다. 첫 번째는 앙상블 학습 기법 + 홀트-윈터스 모형이다. 앙상블 학습 기법으로 RPART와 CTREE를 이용한 배깅과 랜덤포레스트를 사용했다. 추세 추정 방법에서 ARIMA 모형의 성능이 좋지 않았기 때문에 여기에서는 홀트-윈터스 모형만 다루었다. 두 번째는 선형회귀모형(LM) + 앙상블 학습 기법 + ARIMA or 홀트-윈터스 모형이다. 전력소모량과 높은 상관관계에 있는 조정된 온도변수를 이용하여 모형의 단점을 극복한 방법이다. 총 9개의 모형을 설정했고 아래와 같이 정리했다.

Model 1: Bagging with RPART + Holt-Winters

Model 2: Bagging with CTREE + Holt-Winters

Model 3: Random forest + Holt-Winters

Model 4: LM + Bagging with RPART + ARIMA

Model 5: LM + Bagging with CTREE + ARIMA

Model 6: LM + Random forest + ARIMA

Model 7: LM + Bagging with RPART + Holt-Winters

Model 8: LM + Bagging with CTREE + Holt-Winters

Model 9: LM + Random forest + Holt-Winters

5.2. 모수 설정

9개 모형에 각각 사용된 모수들은 Table 5.1과 같다. Period는 추세 추정에 사용된 데이터의 주기를 나타내며 α, β, γ는 홀트-윈터스 모형의 모수이다. ARIMA 모형을 사용한 경우는 ARIMA(p,d,q)의 형태로 표현했다. 배깅을 사용한 모형(Model 1, 2, 4, 5, 7, 8)은 100개의 붓스트랩 샘플을 사용했고, 랜덤포레스트를 사용한 모형(Model 3, 6, 9)은 1000개의 붓스트랩 샘플을 사용했다.

Table 5.1. Parameters used in each model

Model	Trend parameters			
	Period	α	β	γ
1, 2, 3	2	0.9978106	0.8288478	0.1417098
4, 5, 6	91		ARIMA$(0, 2, 0)$	
7, 8, 9	2	0.2674424	0.1843937	0.3083806

Table 5.2. Accuracy based on train and test data

Model	t.MAPE (%)	RMSE	MAE	MAPE (%)
1	0.7569	0.1134	0.0932	1.1117
2	0.8002	0.1175	0.0942	1.1240
3	1.0246	0.1115	0.0895	1.0666
4	0.4478	0.1021	0.0853	1.0113
5	0.4903	0.1028	0.0859	1.0184
6	0.6094	**0.0998**	**0.0835**	**0.9898**
7	0.4461	0.1021	0.0853	1.0114
8	0.4888	0.1028	0.0859	1.0185
9	0.6077	**0.0998**	**0.0834**	**0.9897**

t.MAPE = train data MAPE; RMSE = root mean squared error; MAE = mean absolute error; MAPE = mean absolute percentage error.

5.3. 측정 도구

9개의 모형들을 비교하는 측도로 훈련 데이터에는 mean absolute percentage error (MAPE)를 사용하고, 예측에 활용한 시험 데이터에는 root mean squared error (RMSE), mean absolute error (MAE), MAPE를 사용한다. 동일한 계산 방법이지만 구분을 위해서 훈련 데이터에 사용한 MAPE는 train data MAPE (t.MAPE)라고 표기하겠다. 이러한 측도들은 식 (5.1)과 같이 표현한다.

$$\begin{aligned}
\text{t.MAPE} &= \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t-\hat{y}_t}{y_t}\right|\times 100\ \ (\%),\\
\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_t)^2},\\
\text{MAE} &= \frac{1}{n}\sum_{t=1}^{n}\left|y_t-\hat{y}_t\right|,\\
\text{MAPE} &= \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right|\times 100\ \ (\%),
\end{aligned} \tag{5.1}$$

여기서 t는 시점이고, N은 훈련 데이터 갯수이고 n은 시험 데이터 갯수이다. 각 측도에 대한 판단은 값이 작을수록 실제 값과 오차율이 적은 것으로 더 좋은 모형으로 볼 수 있다.

5.4. 분석 결과

모형 비교 측도로 Table 5.2를 보면 Model 1–3 보다 Model 4–9가 더 좋은 결과를 보였다. 특히 선형회귀모형에 랜덤 포레스트를 조합한 모형(Model 6, 9)이 가장 좋은 성능을 보였다. 선형회귀모형(LM)을 사용한 모형들 중에서 ARIMA 모형으로 추세를 추정하는 것과 홀트-윈터스 모형으로 추세를 추정하는

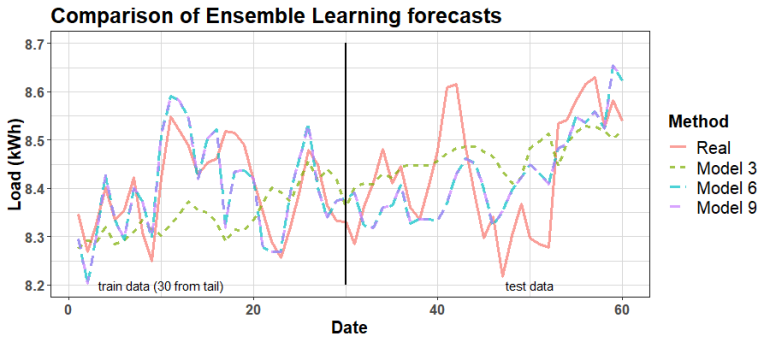


Figure 5.1. Comparison of random forest forecasts

것은 큰 차이가 없었다. 앙상블 모형 중에서는 배깅의 방법 보다는 랜덤포레스트의 방법이 더 좋은 결과를 도출했다. Figure 5.1은 앙상블 모형으로 랜덤 포레스트를 사용한 모형(Model 3, 6, 9)들을 비교한 그래프이다. 왼쪽은 훈련 데이터 중 끝에서 30일의 데이터를 적합된 각 모형과 비교한 것이고 오른쪽은 시험 데이터와 각 모형들의 예측값을 비교한 것이다. 그래프에서도 알 수 있듯이 선형회귀모형을 사용한 모형들이 더 잘 적합되고, 더 잘 예측하는 것을 보여준다. Figure 5.1를 보면 훈련 데이터에서 실제 데이터와 비교하는 그림은 어느정도 잘 맞아 보이지만 오른쪽 예측의 부분은 왼쪽에 비하면 그렇게 완전히 들어맞아 보이지는 않는다. 이는 조정된 기온 데이터 역시 미래의 값을 예측해야 하기 때문에 그 부분에서 발생하는 문제점이라고 볼 수 있다. 본 연구에서는 간단한 시계열 분석 방법으로 예측하였기에 많은 차이를 보이지만 더 좋은 모형을 통해 평균 기온을 더 정확하게 예측한다면 이 모형도 더욱 발전할 수 있는 기회가 될 것이라 예상한다.

6. 결론

본 연구에서는 A 호텔의 일일 전력소모량 데이터에 대한 분석 방법으로 앙상블 학습 기법을 도입했고, 모형의 부족한 부분을 조정된 기온 변수를 이용한 선형회귀모형을 접목했다. 그리고 이를 토대로 A 호텔의 전력소모량을 예측했다. 로그변환된 전력소모량 데이터는 조정된 기온 데이터와 상관계수가 0.86으로 높은 상관관계에 있었으며 이를 이용한 선형회귀모형은 앙상블 학습 기법을 잘 보완해 주었다. 그 근거로 선형회귀모형(LM) + 앙상블 학습 기법 + ARIMA 또는 홀트-윈터스 모형이 전반적으로 앙상블 학습 기법 + 홀트-윈터스 모형보다 더 예측을 잘하는 것으로 나타났다. 의사결정나무를 이용하여 A호텔의 일일 전력소모량을 예측하는데에는 설명변수로 사용할 수 있는 특징의 한계가 있고 전력소모량과 선형관계에 있는 조정된 기온 변수의 도입이 이러한 결과를 이끌어 냈다고 볼 수 있다. 앙상블 학습 기법 중에서는 차이가 미미하지만 배깅보다는 랜덤포레스트가 더 좋은 결과를 보였으며, 이는 이론상 랜덤 포레스트가 배깅의 단점을 보완한 것임을 어느정도 뒷받침 해주고 있다.

통상 전력소모량은 전기를 사용하는 용도와 상황에 따라 다른 패턴을 보인다. 또한 데이터 수집 방법에 따라서도 새로운 패턴이 보일 수 있다. 그러므로 이러한 특징들을 잘 파악하고 탐색적 분석을 통해 적절한 분석 방법 및 모형을 고려하는 것이 중요하다. 전력소모량에 대한 분석은 우리가 전기를 사용하는 한 매우 중요한 요소 중 하나이며, 꾸준한 연구와 관심 그리고 지속적인 데이터 수집이 이루어져야 할 것이다. 또한 기온 뿐만 아니라 전력소모량에 영향을 주는 다른 외부 변수, 예를 들어 호텔의 경우 연휴, 투숙객 및 회의시설 이용 고객 수 등을 고려한 연구가 향후 이루어진다면 더 좋은 예측 모형이 개발될 것이라고 기대한다.

References

Ahn, J. Y., Park, S. M., and Kim, C. B. (2017). A study on neural network model for winter electric power demand prediction, *Journal of Korean Institute of Information Technology*, **15**, 1–9.

Choi, R. I. (2018). Tourism strategy and cultural policy research through development of local tourism resources, *Asia-Pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, **8**, 43–51.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess, *Journal of Official Statistics*, **6**, 3–73.

Cowpertwait, P. S. P. and Metcalfe, A. V. (2009). *Introductory Time Series with R*, Springer.

Grmanova, G., Laurinec, P., Rozinajova, V., *et al.* (2016). Incremental ensemble learning for electricity load forecasting, *Acta Polytechnica Hungarica*, **13** 97–117.

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. (2010). Party: A laboratory for recursive partytioning.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Laurinec, P. (2017). Ensemble learning for time series forecasting in R, (2018, August 10), Retrieved from https://petolau.github.io/Ensemble-of-trees-for-forecasting-time-series/

Lee, S. I. (2015). Transition of energy management paradigm and future demand management technology prospect, *Future Horizon*, **24**, 12–15.

Massidda, L. and Marrocu, M. (2018). Smart meter forecasting from one minute to one year horizons, *Energies*, **11** 3520.

Moritz, S. and Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R, *The R Journal*, **9** 207–218.

Park, J. H., Na, W. S., and Xu, Y. (2017). Evaluation of demand power prediction performance based on deep learning algorithm and data preprocessing. In *Proceedings of the Korea Information Science Society*, 1882–1884.

Shi, H., Xu, M., and Li, R. (2018). Deep learning for household load forecasting-A novel pooling deep RNN, *IEEE Transactions on Smart Grid*, **9**, 5271–5280.

Shin, D. H. and Kim, C. B. (2016). A study on deep learning input pattern for summer power demand prediction, *Journal of Korean Institute of Information Technology*, **14**, 127–134.

Tak, H., Kim, T., Cho, H. G., and Kim, H. (2016). A new prediction model for power consumption with local weather information, *Journal of the Korea Contents Association*, **16**, 488–498.

Therneau, T. M. and Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines.

Yun, S. C. (2018). Focus on fair growth of contents industry and fostering high valued tourism industry, *Nara Economy* (nara.kdi.re.kr) 2018 Feb, Special Theme, 16–17.

앙상블 학습과 온도 변수를 이용한 A 호텔의 전력소모량 예측

김재휘[a] · 김재희[a,1]

[a]덕성여자대학교 정보통계학과

(2019년 1월 18일 접수, 2019년 2월 15일 수정, 2019년 2월 15일 채택)

요 약

과거의 전력소모량을 분석하여 미래의 전력소모량을 예측하는 것은 에너지 계획과 정책 결정에 있어 많은 이점을 가져다준다. 기계학습은 최근 전력소모량을 예측하는 분석 방법으로 많이 사용하고 있다. 그중 앙상블 학습은 모형의 과적합 현상을 방지하고 분산을 줄여 예측의 정확성을 높이는 방법으로 알려져 있다. 하지만 일별 데이터에 앙상블 학습을 적용했을 때 분석 방법의 특성으로 인해 피크를 잘 나타내지 못하고 중심값으로 예측하는 단점을 보였다. 본 연구에서는 앙상블 학습 전에 온도 변수와의 상관성을 고려하여 선형모형으로 적합함으로써 앙상블 학습의 단점을 보완한다. 그리고 9개의 모형을 비교한 결과 온도 변수를 선형모형으로 적합하고 랜덤포레스트를 사용한 모형이 결과가 가장 좋음을 보여준다.

주요용어: 앙상블 학습, 온도, 배깅, 랜덤 포레스트, 시계열 자료 예측

이 논문은 한국에너지기술평가원(KETEP)과 산업통상자원부(MOTIE)의 지원을 받아 수행된 연구과제입니다 (No. 20161210200610). 또한 한국전력공사의 지원을 받아 수행되었습니다 (Grant number:R18XA01).

[1]교신저자: (01369) 서울시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과.
E-mail: jaehee@duksung.ac.kr