머신러닝 및 딥러닝 모델의 스태킹 앙상블을 이용한 단기 전력수요 예측에 관한 연구

이정일*, 김동일*
*충남대학교 컴퓨터공학과
jungil.lee@o.cnu.ac.kr dkim@cnu.ac.kr

A Study on Short-Term Electricity Demand Prediction Using Stacking Ensemble of Machine Learning and Deep Learning Ensemble Models

Jung-Il Lee*, Dong-il Kim*
*Dept. of Computer Engineering, Chung-Namg University

요 약

전력수요는 월, 요일 및 시간의 계절성(Seasonality)을 보이는 데이터이다. 각 계절성에 따라 특성이 다르기 때문에, 전력수요를 예측하기 위해서는 계절성의 특성을 고려한 다양한 모델을 선정하고, 병합하는 방법이 필요하다. 본 연구에서는 전력수요의 계절성을 고려한 다양한 예측모델을 병합하여 이용할 수 있도록 스태킹 앙상블 적용하고 실험결과를 기술한다. 또한, 162개 도시의 기상 데이터와 인구 데이터를 예측에 이용하는 방법, Regression 모델과 Time-series모델에 입력하는 특징(Feature)의 전처리 방법, 베이지안 최적화를 이용한 머신러닝 및 딥러닝 모델의 하이퍼파라메터 최적화 방법을 제시한다.

1. 서론

단기 전력수요 예측은 전력회사의 발전기 운영계획을 수립하는데 이용되는 중요한 요소이다. 실제 수요보다 더 높게 수요를 예측하면, 생산된 전기가 남아서 손실이 발생한다. 반면에, 실제 수요보다 더 낮게 수요를 예측하면 공급할 수 있는 전기가 부족하여 정전이 발생하며, 고객이 정전으로 인하여 입은 손해를 배상하여야 한다. 이와 같이, 전력수요 예측은 전력회사의 경영에 영향을 미치는 중요한 요소이다.

본 논문에서는 전력수요 예측의 정확도를 개선하기 위한 머신러닝과 딥러닝의 스태킹 앙상블 모델을 적용한 방법에 대하여 소개하고자 한다. 또한, 수요예측을 위한 데이터의 전처리, 결측치 및 이상치 제거, 상관분석, 특징 전처리 및 하이퍼파라메터 최적화 방법에 대해서도 살펴보고자 한다.

2. 데이터 수집 및 전처리

2.1 전력수요 데이터 특징분석

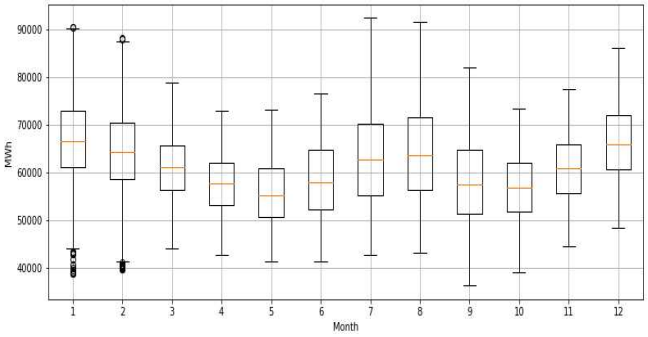
전력수요는 1시간 단위 데이터로, 단위는 MWh이다. 2013.01.01.부터 2021.03.31.까지 기간에 대한 통계정보는 표 1과 같다.

표 1 전력수요 데이터의 통계

count	mean	std	min	max
72288	61217.9	8556.1	36233	92478

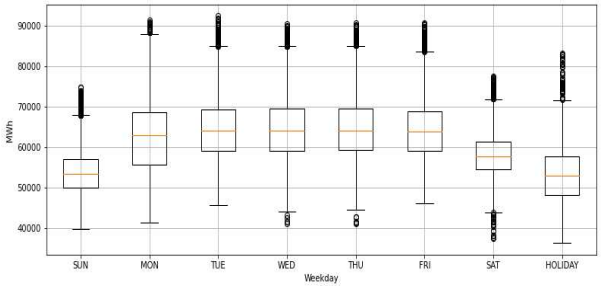
전력수요 데이터는 월, 요일, 시간별로 다른 특성을 보인다. 그림 1과 같이 여름철과 겨울철에는 냉난방 부하가 증가하여 전력수요가 높다. 반면, 봄과 가을은 전력수요가 낮다.

그림 1 월별 전력수요 박스플롯



요일별로는 그림 2와 같이 회사 및 공장 등 업무가 시작되는 월요일부터 금요일까지는 전력수요가 높고, 토요일과 휴일에는 전력수요가 낮다. 시간별로는 주간에는 전력수요가 높고, 야간에는 전력수요가 낮다.

그림 2 요일별 전력수요 박스플롯



2.2 기상 데이터 전처리

기상데이터는 인구가 많은 8개 광역시의 데이터를 이용하여 전국 단위의 기상데이터로 가중 평균하는 것이 일반적이다. 하지만, 본 연구에서는 전국 162개 기상관측 지점의 기상데이터를 이용하였으며, 각 관측지점의 가중치를 계산하기 위하여 162개 지역의 월별 인구데이터를 활용하였다. 각 기상관측 지점별로 평균 1.38%의 데이터가 결측된 것으로 나타났다. 그래서, 결측률이 1.38%를 초과하는 6개 지점은 제외하고, 인근지역의 기상데이터로 대체하였다. 나머지 156개 지점의 기상데이터 10,626,336개 중에서 기온은 17,888개(0.17%), 풍속은 28,266개(0.27%), 습도는 28,465(0.27%)개가 결측되었다. 결측치는 선형보간법으로 보정하였다.

기상 데이터를 예측모델에 적용하기 위해서는 전국 단위의 평균값으로 환산하여야 한다. 각 도시의 인구수를 반영하여 수식 1과 같이 가중 평균하였다.

수식 1 전국단위 기온 산출

$$Temperature_t = \sum_{t}^{n}\sum_{c}^{n} Temperature_{t,c} \times \frac{Population_{t,c}}{Population_t}$$

2.3 코로나19 데이터 전처리

코로나19로 인하여 사업장의 영업시간이 줄어들고, 손님 수가 감소하여 전력사용량에 변화가 발생할 것으로 예상된다. 코로나 데이터의 통계정보는 표 2와 같다.

표 2 코로나 데이터 통계정보

컬럼명	count	mean	std	min	max
확진자 수	421	244.8	266.3	0	1240
사망자 수	421	4.1	5.7	0	40
격리해제 수	421	225.7	274.8	0	2143
결과 음성 수	394	18938.7	15472.7	2472	66337
검사 완료 수	394	19189.6	15669.5	2550	67175
검사진행 수	421	47878	47253.8	40	193751
치료환자 수	394	4775.1	4448.9	623	18054

2.4. 이상치 제거

IQR을 이용하여 표 4와 같이 데이터의 이상치를 탐지하고 제거하였다. 제거된 이상치는 선형보간법으로 보정하였다.

표 3 데이터 이상치 탐지결과

데이터명		Lower Bound	Upper Bound	이상치 개수	이상치 비율
전력수요		36859.1	84736.1	347	0.48%
기상	기온	-22.0	47.6	308	0.003%
	풍속	-2.25	5.35	288383	2.72%
	습도	-1.0	143.0	3	0.00003%
코로나	확진자수	-496	944	17	4%
	사망자수	-5.0	11	39	9.3%
	격리해제수	-413	811	17	4%
	결과음성수	-12926	42018	49	11.6%
	검사완료수	-13038	42402	50	11.9%
	검사진행수	-56760.5	143931.5	35	8.3%
	치료환자수	-8207.5	16388.5	16	3.8%

3. 데이터 상관분석

3.1 상관분석 개요

표 4는 변수간의 상관계수를 나타낸다. 전력수요에 대하여 풍속과 습도의 상관계수가 비교적 높으며, 기온과 코로나 변수는 상관계수가 낮다. 코로나와 관련한 변수는 상관계수가 0.11~0.12로 비슷하고, 코로나 변수간의 상관계수가 0.84~0.92로 높으므로 대표적인 변수인 확진자수만 남기고 불필요한 변수는 제외한다.

표 4 상관계수

	전력수요	기온	풍속	습도	확진자수	사망자수	격리해제수	결과음성수	검사완료수	검사진행수	치료환자수
전력수요	1	-0.13	0.24	-0.28	0.12	0.11	0.11	0.12	0.12	0.12	0.11
기온	-0.13	1	0.07	0.22	-0.18	-0.19	-0.17	-0.14	-0.14	-0.17	-0.19
풍속	0.24	0.07	1	-0.49	0.01	0.01	0.01	0.02	0.02	0.01	0.02
습도	-0.28	0.22	-0.49	1	-0.05	-0.06	-0.04	-0.04	-0.04	-0.03	-0.06
확진자수	0.12	-0.18	0.01	-0.05	1	0.81	0.78	0.84	0.84	0.9	0.91
사망자수	0.11	-0.19	0.01	-0.06	0.81	1	0.84	0.81	0.81	0.88	0.91
격리해제수	0.11	-0.17	0.01	-0.04	0.78	0.84	1	0.87	0.87	0.89	0.89
결과음성수	0.12	-0.14	0.02	-0.04	0.84	0.81	0.87	1	1	0.92	0.89
검사완료수	0.12	-0.14	0.02	-0.04	0.84	0.81	0.87	1	1	0.92	0.9
검사진행수	0.12	-0.17	0.01	-0.03	0.9	0.88	0.89	0.92	0.92	1	0.96
치료환자수	0.11	-0.19	0.02	-0.06	0.91	0.91	0.89	0.89	0.9	0.96	1

3.2 기상변수의 월별 상관분석

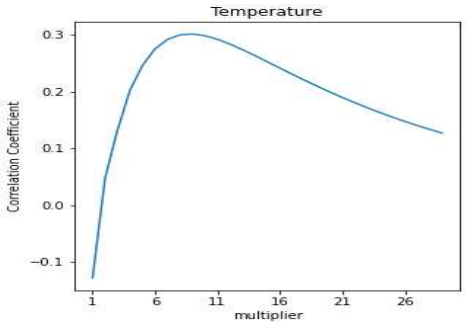
기상변수는 표 5와 같이 상관계수가 월별로 다르기 때문에, 비선형성을 높여줘야 한다. 기온, 풍속 및 습

도에 대하여 2승부터 30승까지 계산한 결과, 그림 3과 같이 기온만 8승에서 상관계수가 0.3으로 가장 높게 나타났다.

표 5 기상변수와 전력수요의 월별 상관계수

월	기온	풍속	습도	월	기온	풍속	습도
1	-0.15	0.17	-0.28	7	0.68	0.33	-0.5
2	-0.16	0.2	-0.24	8	0.67	0.38	-0.53
3	-0.03	0.23	-0.18	9	0.48	0.36	-0.26
4	0.18	0.25	-0.2	10	0.17	0.16	-0.28
5	0.48	0.34	-0.34	11	-0.14	0.18	-0.38
6	0.62	0.49	-0.43	12	-0.08	0.28	-0.34

그림 3 기온의 상관계수 변화



3.2 자기 상관분석

표 6은 전력수요의 자기상관계수이다. t-24와 t-168의 전력수요가 상관계수가 높고, 과거로 갈수록 상관계수가 낮아지는 것을 알 수 있다.

표 6 과거 전력수요와의 요일별 상관계수

시점	전체	일	월	화	수	목	금	토
t	1	1	1	1	1	1	1	1
t-24	0.78	0.88	0.68	0.95	0.97	0.96	0.97	0.88
t-48	0.58	0.73	0.75	0.73	0.93	0.94	0.93	0.84
t-72	0.54	0.72	0.88	0.79	0.71	0.91	0.91	0.81
t-96	0.53	0.69	0.89	0.87	0.77	0.7	0.89	0.77
t-120	0.53	0.67	0.88	0.88	0.85	0.75	0.69	0.7
t-144	0.71	0.62	0.88	0.88	0.87	0.84	0.76	0.78
t-168	0.9	0.87	0.91	0.88	0.88	0.87	0.84	0.8

4. 특징 전처리

본 연구의 수요예측 모델에서 활용하는 특징은 표 7과 같이 전력수요의 계절적 특성을 나타내는 범주형 변수와 전력수요와 상관관계를 갖는 수치형 변수로 구성된다. 수치형 변수는 MinMax 정규화를 적용한다. Regression 모델에는 자기상관도를 반영하기 위하여 전력수요$_{t-24}$와 전력수요$_{t-168}$ 특징을 추가하며, 특징의 형태는 (46656, 52)이다.

Time-series 모델에서는 특징을 순차적으로 생성한다. 전력수요 특징은 자기상관도가 높은 (t-24, t-48, t-72)으로 하고, 전력수요를 제외한 다른 특징은 (t, t-24, t-48)로 하여 시퀀스를 생성한다. 특징의 형태는 (46656, 3, 50)이고, Many-to-One 모델로 t시점의 전력수요를 예측한다.

표 7 특징 설명

명칭	유형	특징 변환
월	범주	12개의 One-hot 인코딩
시간	범주	24개의 One-hot 인코딩
요일	범주	7개의 One-hot 인코딩
공휴일	범주	
기온	수치	
Log(기온^8)	수치	기온^{8}을 Log1p 계산
Log(풍속)	수치	풍속을 Log1p 계산
습도	수치	
확진자수	수치	

5. 모델 학습 및 예측결과

5.1 하이퍼 파라메터 최적화

각 모델의 하이퍼파라메터는 검증 데이터의 MSE를 최소화하는 목적식과 표 8의 탐색공간을 바탕으로 베이지안 최적화[1]를 이용하여 최적화하였다

표 8 하이퍼 파라메터 탐색공간

모델명	하이퍼 파라메터	최소값	최대값
XGBoost [2]	max_depth	6	80
	subsample	0.9	1
	n_estimators	1000	2000
	learning_rate	0.01	0.2
MLPs[3]	n_hidden	1	100
	n_neurons	50	1000
	dropout_rate	0	0.05
	batch_size_rate	0.001	0.01
LSTM[4]	n_hidden	1	40
	n_neurons	100	300
	dropout_rate	0	0.001
	batch_size_rate	0.001	0.01
TCN[5]	nb_filters	32	64
	kernel_size	2	4
	dilations	2	8
	dropout_rate	0	0.001
	batch_size_rate	0.001	0.01

튜닝을 하여도 성능에 큰 영향을 주지 않는 하이퍼파라메터는 다음과 같이 적용한다.

- 활성화 함수 : RELU, 옵티마이저 : ADAM
- 학습률 : 0.002, Epoch : 500, 조기종료 : 50
- 각 반복마다 MSE가 낮은 모델을 저장

5.2 학습, 검증 및 테스트 범위

12개월에 대한 모델의 정확도를 검증하기 위하여

표 9와 같이 실험계획을 수립하였다. 학습과 검증 데이터셋의 비율은 80:20이며, 시계열 데이터이므로 혼합하지 않는다.

표 9 학습, 검증 및 테스트 범위

구분	학습 및 검증		테스트(예측)	
	시작일	종료일	시작일	종료일
0	2013.03.01	2020.02.31	2020.03.01	2020.03.31
…	…	…	…	…
11	2014.02.01	2021.01.31	2021.02.01	2021.02.31

5.3 모델 앙상블 및 예측결과

전력수요는 월, 요일 및 시간별로 다른 특성을 보인다. 그러므로, 특성에 맞는 다양한 모델을 적용하기 위한 앙상블 기법이 필요하다. 본 논문에서는 그림 4와 같이 스태킹(Stacking)[6] 앙상블을 이용하였으며, 실험결과는 표 10과 같다.

그림 4 스태킹 앙상블 구성도

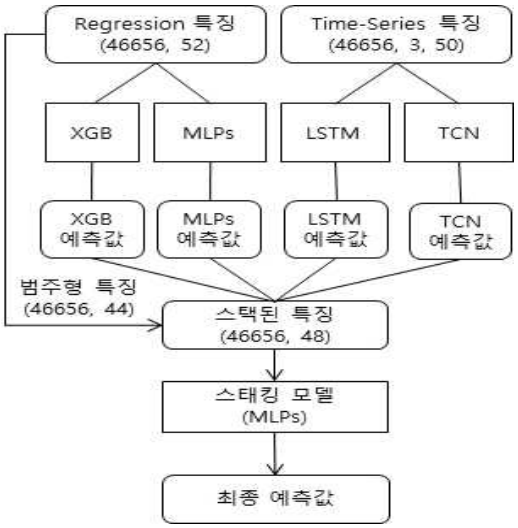


표 10 예측 MAPE

구분	XGB	MLP	LSTM	TCN	Stacking
0	1.95	1.89	1.77	1.78	1.7
1	2.07	2.31	1.92	1.91	1.95
2	2.53	2.53	2.81	2.78	2.53
3	1.88	1.98	1.9	2	1.85
4	1.85	1.94	1.83	1.8	1.79
5	2.44	2.39	1.99	1.87	1.92
6	2.05	1.86	1.77	1.77	1.7
7	1.96	1.82	1.74	1.74	1.63
8	1.56	1.69	1.7	1.6	1.45
9	1.83	2.06	2.08	1.93	1.74
10	2.84	2.07	2.58	1.85	1.75
11	2.55	2.61	2.23	2.21	2.02
평균	2.12	2.09	2.02	1.93	1.83

6. 결론

본 논문에서는 다음 날의 24시간 전력수요를 예측하기 위하여 전력수요, 기상 데이터 및 코로나 데이터의 결측치 및 이상치 보정 등 전처리를 수행하였으며, 데이터 간의 상관관계를 분석하여 Regression 모델과 Time-seriesm모델의 입력으로 적합한 특징을 도출하였다.

전력수요는 월, 요일, 시간별 특성이 다르기 때문에, 2개의 Regression 모델과 2개의 Time-series 모델을 적용하였으며, 모델별 탐색공간을 정의하고, 손실이 최소화되는 목적식을 바탕으로 베이지안 최적화를 이용하여 하이퍼파라메터를 최적화하였다. 그리고, 4개 모델의 예측결과를 병합하기 위하여, 스태킹 앙상블을 이용하였다.

실험을 수행한 결과, 개별적인 머신러닝 및 딥러닝 모델의 예측 정확도보다 스태킹 앙상블의 예측 MAPE가 0.1~0.29% 낮은 것을 확인하였다. 결론적으로, 계절적 특성(월, 요일, 시간)을 가지는 전력수요의 예측에 다양한 머신러닝 및 딥러닝 모델을 스태킹 앙상블로 병합하면 보다 나은 성능을 얻을 수 있다고 할 수 있다. 향후에는 더 많은 딥러닝 모델을 추가로 이용하여 정확도를 향상시킬 예정이다.

참고문헌

[1] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas, “The Application of Bayesian methods for seeking the extrenum, Towards Global Optimization, 2:117-129, 1978

[2] Tianqi Chen and Carlos Guestrin, “XGBoost:A Scalable Tree Boosting System”, In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016

[3] David Rumelhart et al, “Learning International Representations by Error Propagation”, Defense Technical Information Center technical report, 1985

[4] Sepp Hochreiter and Jurgen Schmidhuber, “Long Short-Term Memory”, Neural Computation 9(8):1735-1780, 1997

[5] Karen Simonyan, “WAVENET : A Generative model for raw audio”, arXiv:1609.03499v2, 2016

[6] David H. Wolpert, “Stacked Generalization” Neural Networks 5, no. 2, (1992):241-259