시계열 데이터 특징에 따른 인공지능 예측 모델 선정에 대한 연구

A Study on the Selection of AI Prediction Models According to the Characteristics of Time Series

신의현[1] · 김동건[2] · 김광수[2]

Ui-Hyeon Shin, Dong-Keon Kim and Kwangsu Kim

[1]성균관대학교 시스템경영공학과

E-mail: shineh96@naver.com

[2]성균관대학교 소프트웨어학과

E-mail: kdk1996@skku.edu, kim.kwangsu@skku.edu

요 약

과거에는 시계열 예측 모델로 회귀분석 모델 또는 RNN 모델만을 활용하였으나, 최근 연구 결과로 다양한 모델을 활용한 시계열 예측 방법이 등장하고 있다. 하지만 시계열 데이터 특징에 따라 어떤 예측 모델이 적합한지에 대한 기준이 제시되어 있지 않다. 따라서 우리는 시계열 분해를 통해 추출한 시계열적 특징에 따른 예측 모델 선정 기준을 제시한다. 본 논문에서는 시계열 데이터의 특징을 추출하여 무작위 정도를 구하는 방법에 관해 설명하고, RNN 기반 예측 모델, CNN 기반 예측 모델을 활용한 예측 정확도 결과를 도출한다.

키워드 : 시계열 분해, 무작위 정도, 딥러닝, 시계열 예측

1. 서 론

시계열 예측은 시간 흐름에 따라 순차적으로 관찰된 과거 데이터를 통해 미래를 예측하는 것이다. 이러한 시계열 예측 방법론은 최근 기계학습 및 딥러닝 기술 발달로 인해 날씨, 에너지 소비, 주가지수, 교통량 등 다양한 예측 분야에 두드러진 연구 성과가 나타나고 있다 [1].

그러나 최근 연구에서 다양한 시계열 예측 방법들이 제시되고 있지만, 각 데이터 특징에 따라 적합한 모델이 서로 다르다. 더불어 각 특징에 따라 어떤 예측 모델이 적합한지에 대한 기준이 제시되어 있지 않아, 보유한 데이터에 따른 예측 모델을 선정하는 데 어려움이 있다.

본 논문에서는 시계열 데이터로부터 추세, 계절성을 추출하고, 규칙적인 패턴을 나타내는 추세와 계절성을 제거하여 잔차를 얻는다. 잔차 값을 바탕으로 데이터의 무작위 정도를 나타내는 기준을 제시하고, 그에 따라 데이터 특징에 적합한 딥러닝 모델을 선정하는 기준을 제시한다.

2. 본 론

2.1 시계열 분해를 통한 무작위 정도 산출

시계열 데이터는 추세와 계절성으로 데이터를 표현 할 수 있다. 시계열 분해는 데이터로부터 추세와 계절성을 분리하는 기법이다. 이를 통해 규칙적인 패턴(추세와 계절성)을 제거하면 시계열 데이터가 가진 고유한 무작위 값을 추출 할 수 있는데, 이것을 잔차라 한다. 본 연구에서는 시계열 분해를 통해 얻은 잔차를 데이터의 시계열적 특징으로 고려한다. 우리는 데이터의 특징을 나타내는 방법으로 잔차(x_i)에 분산을 적용한 방법을 식(1)과 같이 정의하고, 이를 무작위 정도(r)로 표현한다. 무작위 정도가 높을수록 데이터가 불규칙적인 것을 의미한다.

$$r = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (1)$$

2.2 데이터 선정 및 분석

본 논문에서는 다양한 데이터에서 나타나는 시계열적 특징을 고려하기 위해, 날씨, 에너지 소비, 주가지수, 교통량 분야의 10개의 시계열 데이터를 사용한다. 각 데이터는 일 단위로 측정된 총 1,000개의 샘플을 가지며 이 중 train 데이터 700개, valid 데이터 200개, test 데이터 100개로 구분한다. 일반적으로 시계열 데이터 정규화에는 min-max scale과 z-score scale을 사용하는데, 데이터에 이상치가 존재할 경우 z-score scale 방법이 더 효과적이기 때문에 본 연구에서는 각 데이터를 z-score를 통해 정규화한 후, 식(1)을 통해 무작위 정도 r을 계산한다(표 1).

표 1의 무작위 정도를 통해 주가지수 데이터는 규칙적 패턴을 보이고, 에너지 소비, 교통량 데이터는 불규칙적임을 알 수 있다.

2.3 사용모델 선정

본 논문에서는 여러 딥러닝 시계열 예측 모델 중 우수한 성능을 보이는 RNN 기반의 Long Short-Term Memory (LSTM) [2], Gated Recurrent Units (GRU) [3]와 CNN 기반의 1d-CNN [4], Temporal Convolutional Network (TCN) [5] 4가지 모델을 선정하여 각 데이터 특징별 모델 간의 성능을 비교한다.

감사의 글 : 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신산업진흥원의 지원을 받아 수행된 헬스케어 AI 융합 연구개발 사업임(No.S0316-21-1006)

각 모델은 Lara-Benitez [6]가 진행한 시계열 예측 실험을 통해 도출된 최적의 모델 구조 Hyper-parameter 값을 차용하여 구성한다.

표 1 모델링을 위한 초기 파라미터들
Table 1. Initial parameters for modeling.

#	분야	데이터	무작위정도
1	주가지수	아마존 주가 종가	0.003
2	주가지수	3M 주가 종가	0.009
3	주가지수	구글 주가 종가	0.010
4	주가지수	야후 주가 종가	0.018
5	날씨	서울의 평균기온	0.067
6	날씨	대구의 평균기온	0.074
7	날씨	제주의 평균기온	0.077
8	에너지소비	경기도 lng 연료 소비량	0.400
9	교통량	고속도로 통행량	0.643
10	교통량	대구 버스 이용자수	0.772

표 2. 모델 구조 Hyper-parameter 값.
Table 2. Model Architecture Hyper-parameter values.

	LSTM	GRU	1d-CNN	TCN
Layers	2	2	4	1
Units(Filters)	32	32	(16)	(64)
Return sequence	True	True	-	False
Kernel size	-	-	{7,5,3,3}	{3}
Pool size	-	-	0	-
Dilations	-	-	-	[1,2,4,8]

3. 실험 및 결과

3.1 실험 설계

각 모델의 학습 Hyper-parameter는 Lara- Benitez가 제시한 값으로 구성한다. Learning Rate는 0.001, Past History factor는 1.25, 배치의 크기는 LSTM은 32, GRU/1d-CNN/TCN은 64로 하여 5 Epoch 동안 학습한 뒤, 다음 30개 샘플에 대한 예측을 수행한다. 성능평가 지표로는 Mean Absolute Error(MAE)를 사용하고, 각 모델에 대해 실험을 10번씩 실행하여 기록한 성능의 평균과 각 데이터에서 모델별 성능 순위를 구한다.

3.2 실험 결과

표 3의 실험 결과는 데이터의 무작위 정도에 따라 적합한 예측 모델이 다른 것을 나타낸다. 무작위성이 작은 1-3 데이터는 GRU가, 무작위성이 중간인 4-7 데이터는 LSTM과 TCN이, 무작위성이 높은 8-10 데이터는 1d-CNN이 좋은 성능을 보인다.

GRU와 LSTM과 같은 RNN 기반 예측 모델은 연속적인 정보의 흐름을 학습하여, 샘플 간의 연관성이 높은 데이터, 즉 무작위 정도가 작은 데이터에서 좋은 성능을 보인다. 반면, 무작위성이 강한 시계열 데이터에서는 CNN 기반 예측 모델이 우수한 성능을 보인다.

표 3. 모델 별 시계열 예측 실험 결과.
Table 3. Time series prediction experiment result

#	무작위정도	LSTM	GRU	1d-CNN	TCN
1	0.003	0.447(2)	**0.313(1)**	0.614(3)	1.178(4)
2	0.009	0.497(4)	**0.259(1)**	0.417(3)	0.406(2)
3	0.010	0.582(4)	**0.263(1)**	0.309(2)	0.312(3)
4	0.018	0.594(3)	**0.509(1)**	0.669(4)	0.529(2)
5	0.067	0.333(2)	0.451(4)	0.426(3)	**0.332(1)**
6	0.074	**0.311(1)**	0.417(4)	0.382(3)	0.336(2)
7	0.077	0.390(3)	0.496(4)	0.359(2)	**0.291(1)**
8	0.400	0.751(3)	0.760(4)	**0.651(1)**	0.745(2)
9	0.643	0.708(4)	0.707(3)	**0.563(1)**	0.704(2)
10	0.772	0.735(4)	0.707(3)	**0.459(1)**	0.630(2)

4. 결론 및 향후연구

4.1 결론

본 논문에서는 시계열 분해로 추출된 잔차를 통해 데이터의 불규칙한 특징의 정도를 산출 할 수 있는 지표를 제시하고, 산출한 무작위정도에 따른 적합한 예측 모델을 선정하는 기준을 제시한다. 무작위성이 작은 데이터는 GRU, LSTM과 같은 RNN 기반 예측 모델을 선정하는 것이 유리하고, 무작위성이 높은 데이터는 1d-CNN, TCN과 같은 CNN 기반 예측 모델을 선정해야 한다.

4.2 향후 연구

향후 연구에서는 다른 시계열적 특징인 추세와 계절성과 예측 모델들의 연관성에 대해 비교하여 클러스터링을 통한 모델 선정 기준을 제시하고자 한다.

참 고 문 헌

[1] Lim, Bryan, and Stefan Zohren. "Time Series Forecasting With Deep Learning: A Survey." stat 1050 (2020): 27.

[2] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[3] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[4] Selvin, Sreelekshmy, et al. "Stock price prediction using LSTM, RNN and CNN-sliding window model." 2017 international conference on advances in computing, communications and informatics (icacci). IEEE, 2017.

[5] Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).

[6] Lara-Benitez, Pedro, Manuel Carranza-Garcia, and Jose C. Riquelme. "An Experimental Review on Deep Learning Architectures for Time Series Forecasting." International Journal of Neural Systems (2020).