시계열 데이터의 비정상성 문제해결을 통한 머신러닝 예측의 성능개선 방법

How to improve prediction of Machine Learning performance by solving Non-Stationary problems of time series data

저자 (Authors)
김동우, 신기범, 최승윤, 정준홍
Dongwoo Kim, Kibeom Sin, Sengyoon Choi, Joonhong Jung

출처 (Source)
한국정보과학회 학술발표논문집 , 2021.6, 873-875 (3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10583105

APA Style
김동우, 신기범, 최승윤, 정준홍 (2021). 시계열 데이터의 비정상성 문제해결을 통한 머신러닝 예측의 성능개선 방법. 한국정보과학회 학술발표논문집, 873-875.

시계열 데이터의 비정상성 문제해결을 통한 머신러닝 예측의 성능개선 방법

김동우○ 신기범 최승윤 정준홍
한전KDN
dwooo.kim@gmail.com, sin8616@naver.com, sychoi.84@kdn.com, nelcast@kdn.com

How to improve prediction of Machine Learning performance by solving Non-Stationary problems of time series data

Dongwoo Kim○, Kibeom Sin, Sengyoon Choi, Joonhong Jung
Kepco KDN

요 약

시계열 데이터의를 사용한 머신러닝 예측은 여러 도메인에서 시도되는 주제이다. 그러나 경제 분야에서는 데이터의 임의보행(Random Walk) 문제로 인해 예측이 불가능한 것으로 받아들여져 왔다. 본 논문에서는 경제 분야 시계열 데이터의 비정상성(Non Stationary) 상태를 정상성(Stationary) 상태로 변환하여 임의보행 문제를 해소하는 전처리 방법을 설명한다. 또한, 정상성 변환 결과를 활용하여 머신러닝 예측의 성능개선과 동적인 스태킹 방법에 대해 유연탄가격 예측 사례를 들어 설명 한다.

1. 서 론

시계열 데이터를 이용한 머신러닝 예측은 AI기술의 발전에 따라 많이 시도되는 주제이다.

Surachai Chancharat와 Abbas Valadkhani는 경제데이터가 통계적으로 비정상성(Non Stationary)임을 연구하였으며, 통계적인 분포를 갖지 않는 임의보행(Random Walk) 상태의 데이터는 가격예측이 불가능함을 연구하였다[1]. Rob J Hyndman,George Athanasopoulos는 통계적 비정상 상태의 시계열 데이터를 정상성 상태로 변환하기 위한 차분 및 로그변환 방법을 제안하였다[2].

또한 Luckyson Khaidem, Snehanshu Saha, Sudeepa Roy Dey은 가격예측을 위한 머신러닝 기법을 연구하였다[3]. Subba Rao Polamuri, K. Srinivas, A. Krishna Mohan는 Tree 계열의 예측력을 비교하였다[4]. 예측을 위한 앙상블 학습방법에 대해서는 A. Galicia a, R. Talavera-Llames a, A. Troncoso a, I. Koprinska b, F. Martínez-Álvarez가 연구하였다 [5].

경제분야에서 비정상성(Non Stationary) 상태의 시계열 데이터는 임의보행(Random Walk)으로 분류되고 있으며, 주가 및 가격의 변화를 예측할 수 없는 근거로 사용된다[1]. 본 논문에서는 시계열 데이터의 통계적 비정상성 상태를 확인하고[1], 이를 해소하기 위한 정상성 변환(Stationary Process)[2] 에 주안점을 두어, 예측이 가능한 데이터 상태를 확보하는 방법을 설명한다. 또한 머신러닝 예측방법과 동적인 앙상블 운영방법[3],[4],[5]을 기반으로, 동적인 스태킹 방법을 설계하여 국제 유연탄 가격인 GCI(Global Coal Index)를 예측하는 과정을 설명한다.

2. 본 론

2.1 Random Walk 해소를 위한 Stationary Process

본 논문에서는 국제 유연탄 가격인 GCI를 예측하기 위해 2010.01~2021.02 기간의 경제지표를 활용하였다. 예측에 사용된 모든 데이터에 대해 정상성 여부를 확인하는 방법으로 ADF테스트(Augmented Dicky Fuller Test)를 실행하였다. ADF테스트의 p-value > 0.05 이므로 Non-Stationary 상태로 확인 된다.(<그림 1> 참조).

GCI를 예측하기 위해 시계열 데이터를 정상성이 확보된 상태의 데이터로 변환한 후, 머신러닝예측을 실행한다. 정상성을 확보하는 방법으로는 예측에 사용되는 모든 입력 데이터를 1차 차분, 2차 차분, 로그변환, 멱급수처리 등의 방법으로 변환하는 Stationary Process를 수행한다[2].

<그림 1>는 예측에 사용되는 모든 경제데이터 데이터의 정상성을 확보하기 위해 Stationary Process를 적용 후 ADF 테스트로 정상성을 검사한 결과이다. 원본 데이터인 269종의 데이터는 대부분 Non Stationary 상태이다. 1차 차분, 2차 차분을 실행한 결과는 모두 정상성이 확보 되었다. 정상성이 확보된 시계열 데이터는 통계적 특성이 일정하게 유지되므로, 예측을 위한 데이터

로 사용할 수 있다.

	raw	1st	2nd	log	power	power_1st	power_2nd
BCI14 Index	True	True	True	True	True	True	True
CCOKFOUN Index	True	True	True	True	False	True	True
CHJPM322 Index	True	True	True	True	True	True	True
USDZAR Curncy	False	True	True	True	False	True	True
CNEVCOAL Index	False	True	True	True	False	True	True
...	...	...	...	...	...	...	...
GCN_6000NAR_FOB_FUTURES	False	True	True	True	False	True	True
DES_ARA_6000NAR_CIF	False	True	True	True	False	True	True
RB_6000NAR_FOB	False	True	True	True	False	True	True
AUITCOKV Index	True	True	True	True	True	True	True
LOCADY Comdty	True	True	True	True	False	True	True

269 rows × 7 columns

```
print('[raw]count of True ',df_adfuller['raw'].sum()) # 269개의 True 합 = 0
print('[1st]count of True ',df_adfuller['1st'].sum()) # 269개의 True 합 = 269
print('[2nd]count of True ',df_adfuller['2nd'].sum()) # 269개의 True 합 = 269
print('[log]count of True ',df_adfuller['log'].sum()) # 269개의 True 합 = 0
print('[power]count of True ',df_adfuller['power'].sum()) # 269개의 True 합 = 269
print('[power1st]count of True ',df_adfuller['power_1st'].sum()) # 269개의 True 합 = 269
print('[power2nd]count of True ',df_adfuller['power_2nd'].sum()) # 269개의 True 합 = 269
```

```
[raw]count of True  67
[1st]count of True  269
[2nd]count of True  269
[log]count of True  253
[power]count of True  64
[power1st]count of True  246
[power2nd]count of True  246
```

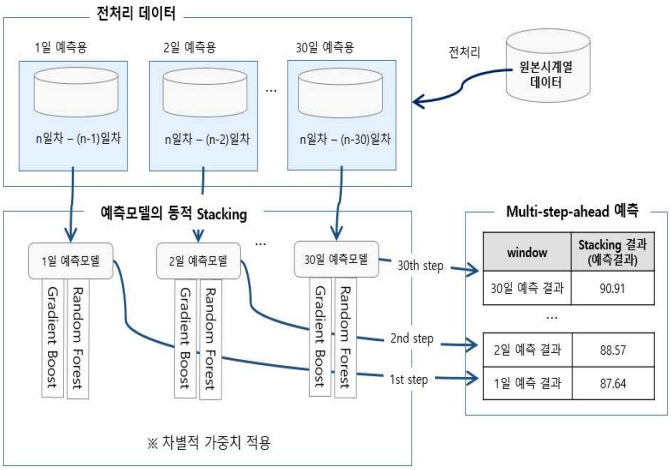
<그림1> 정상성 변환 후 ADF테스트 결과

2.2 동적 스태킹을 통한 예측 방법 설계

예측 결과를 조합하는 스태킹 방법은 예측에 참여한 모델의 예측결과를 조합하는 앙상블 방법이다. 본 논문에서 사용한 예측 방법에서는 두가지의 동적인 스태킹 방법을 사용한다.

첫째, 동적인 Multi-step-ahead 예측이다. Multi-step-ahead 예측은 일반적으로 Dynamic forcasting과 같이 예측결과가 입력값으로 재사용되는 방식이다. 그러나 이 경우 오류누적에 의한 정확도 저하의 문제가 발생한다[5]. 따라서 각 step을 One-step-ahead로 예측하여 오류누적 문제를 해소 한다. 각 step별 예측결과를 연결하여 Multi-step-ahead의 결과를 생성한다. GCI 예측은 1일~30일 이후의 값을 예측하는 것으로 하며, 1-step ~ 30-step의 예측이 각각 실행된다.

둘째, Stacking 방법으로 각 예측결과의 평균을 구하는 방법이 아닌 이전 예측결과의 정확도 별로 차등적인 가중치를 적용하여 최종결과를 계산하는 조합방법을 사용한다[5]. GCI 예측을 위해 Tree계열의 대표적인 2가지의 알고리즘(Gradient Boost, Random Forest)으로 학습한 예측모델을 사용하여 학습한다.



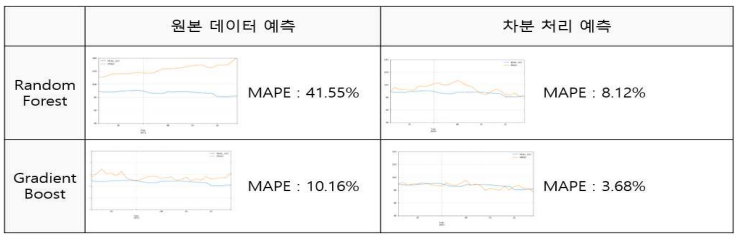
<그림2> GCI의 예측 구성도

예측결과를 조합하는 스태킹 단계에서는 매번 예측 step마다 다른 가중치를 적용하여 조합한다.

<그림2>는 원본 시계열 데이터의 전처리와 예측과정에 대한 구성도이다. 전처리 단계에서는 차분을 통해 정상성을 확보하고, 정상성이 확보된 데이터는 동적 Stacking을 통해 예측결과를 생성한다.

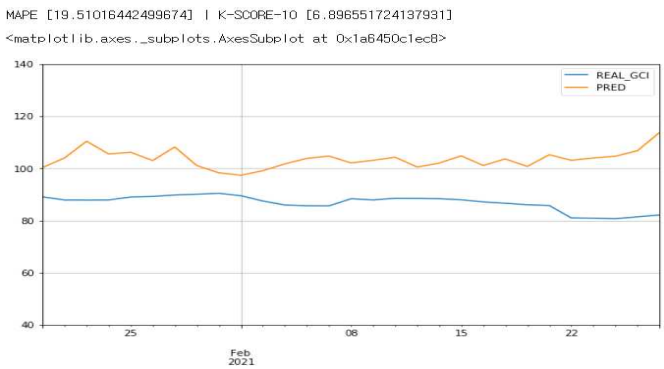
2.3 GCI 예측결과 비교

정상성 확보 전후의 알고리즘 별로 비교한 결과는 <그림3>과 같다.

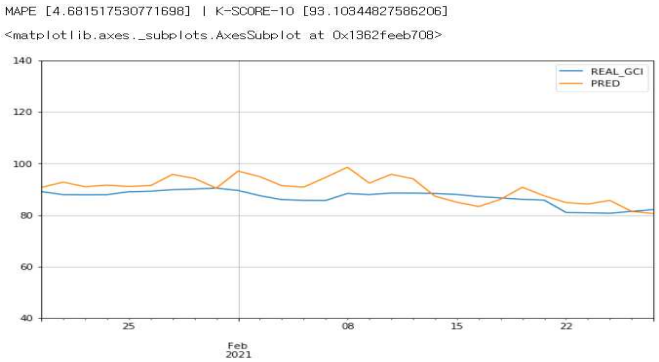


<그림3> 정상성 처리 전후 성능비교

정상성 처리 후 각 알고리즘의 월등한 성능개선을 확인할 수 있다. Random Forest 와 Gredient Boost 알고리즘을 기반으로 스태킹한 예측결과는 <그림4,5> 와 같다.



<그림4> 비정상성 데이터의 GCI예측, MAPE : 19.51%



<그림5> 정상성 데이터의 GCI예측, MAPE : 4.68%

스태킹 실험에 사용한 두 개의 알고리즘에 대하여, 예측 step마다 가중치를 각각 70%와 30%로 차등부여 후 결합하였다. 가중치 차등부여의 근거는 이전 예측 결과에서의 알고리즘별 정확도를 기준으로 우수한 정확도를 보인 알고리즘에 높은 가중치를 부여하는 방식이다. 두 가지의 예측 결과를 비교한 결과, 정상성 처리 후 실행한 예측이 비정상성 상태에서 예측보다 우수한 결과를 얻었다.

3. 결론 및 향후 연구

비정상성(Non Stationary) 데이터는 정상성 처리(Stationary Process)를 통해 정상성(Stationary) 상태로 변환이 가능하다. 특히 본 논문에서 실험한 국제 경제데이터에 대하여 차분 처리로 정상성이 확보 되었다.

트리계열의 앙상블 알고리즘은 정상성 변환 후 예측할 때 우수한 결과를 보이는 것을 발견하였다.

경제 분야의 가격예측이 데이터의 임의보행(Random Walk) 문제으로 인해 불가능 하다는 근거[1]는 본 논문에서 사용한 정상성 변환 방법으로 반박 가능하다. 즉, 데이터 특성에 맞는 전처리 방법과 머신러닝 예측을 함께 적용할 경우 경제 분야의 가격예측도 추세적 변화를 예측할 수 있는 것으로 확인되었다.

본 논문의 연구결과는 발전사의 발전연료 구매를 위한 의사결정 시스템의 기능으로 사용하고 있다. 앞으로 시계열 데이터 특성에 적합한 알고리즘별 전처리 방법을 연구하고 결합방법을 최적화 하는것이 앞으로 연구할 과제이다.

참 고 문 헌

[1] Surachai Chancharat, Abbas Valadkhani, “Structural Breaks and Testing for the Random Walk Hypothesis in International Stock Prices”, THE JOURNAL OF THE KOREAN ECONOMY, vol. 8, 2007

[2] Rob J Hyndman, George Athanasopoulos, “forecasting: Principles and Practice“, 2013

[3] Luckyson Khaidem, Snehanshu Saha, Sudeepa Roy Dey, “Predicting the direction of stock market prices using random forest”, Applied Mathematical Finance, 2016

[4] Subba Rao Polamuri, K. Srinivas, A. Krishna Mohan, “Stock Market Prices Prediction using Random Forest and Extra Tree Regression”, International Journal of Recent Technology and Engineering, 2019

[5] A. Galicia a, R. Talavera-Llames a, A. Troncoso a, I. Koprinska b, F. Martinez-Alvarez, “Multi-step forecasting for big data time series based on ensemble learning”, Knowledge-Based Systems, 2018