다변량 시계열 데이터의 2차원 재배치를 통한 성능 향상

김 민[1], 강승우[2], 조오현[1]
충북대학교[1], 레플러스[2]
{kmin010287, ohyunjo}@chungbuk.ac.kr[1], swkang@repul.net[2]

Performance Improvement through 2D Restacking of Multivariate Time Series Data

Min Kim[1], Seungwoo Kang[2], Ohyun Jo[1]
Chungbuk National University[1], Replus[2]

요 약

본 연구에서는 기존 시계열 데이터 분석에 특화된 파운데이션 모델(Foundation Model)인 TimesNet의 다변량 분석의 한계를 극복하기 위해, 전처리 단계에서 시계열 재배치(Restacking) 기법 활용을 제안한다. 제안하는 기법을 통해 예측 대상 시계열 변수와 상관도가 높은 연관 시계열 변수를 인접하게 재구성해 모델이 관계 정보를 효과적으로 학습하도록 한다. 또한 CNN 기반 시계열 예측 모델인 TimesNet을 기반으로 제안한 전처리 기법을 통한 다변량 시계열 예측 성능 향상을 검증한다. 실험에는 1시간 주기의 교통량 데이터 중 상위 105개 도로 구간을 선정하여, 4일 간 각 구간의 평균 속도 예측하였으며, 실험 결과 다변량 시계열 학습 환경에서 시계열 재배치를 적용한 TimesNet 모델이 적용하지 않은 경우에 비해 Mean Square Error가 8.1% 감소한 성능을 보인다.

Ⅰ. 서 론

다변량 시계열 데이터는 시간적 특성과 각 시계열 변수 간 상관성을 동시에 내포하고 있어 금융, 교통, 에너지 등 다양한 분야의 예측 과제에서 높은 정확도를 위해 반드시 고려되어야 한다. 전통적인 단변량 시계열 모델은 시간적 패턴만을 학습하도록 설계되어 시계열 변수 간 상관관계 정보 활용에 한계가 존재한다. 이러한 문제를 해결하기 위해 최근 연구에서는 다변량 시계열 데이터 예측에 CNN (Convolutional Neural Network) 기반 모델을 활용하고 있다 [1]. CNN 기반 모델의 Convolution 연산은 시간적 특성 및 상관성을 효과적으로 추출하면서도 비교적 적은 계산 자원을 활용하기에 모델의 복잡성 대비 우수한 성능을 보이고 있다 [2].

그 중 TimesNet 모델은 시계열 데이터에 내재된 주기들을 활용해 데이터를 2차원 형태로 재구성하여 학습하도록 설계되어있다 [3]. 그러나 해당 모델의 경우, 다변량 시계열 데이터 입력 시 단변량 시계열 데이터로의 변형이 필요하며, 이 과정에서 시계열 변수 간 상관관계를 소실하고 결과적으로 예측 성능이 저하되는 문제가 발생한다.

본 연구는 이러한 문제를 해결하기 위해 전처리 기법인 상관성 기반 시계열 재배치(Restacking)기법 활용을 제안한다. 제안한 전처리 기법을 통해 모델 변형 없이 TimesNet이 다변량 시계열 예측에서 향상된 성능이 제공되는지 분석한다. 또한 공공 데이터인 서울시 교통량 시계열 데이터를 활용해 전처리 기법을 통한 모델 성능 향상을 검증한다.

Ⅱ. 본론

1. 데이터 전처리

본 연구에서는 서울특별시의 공공데이터 '고속도로 구간별 1시간 단위 평균 속도 누적치'를 활용한다. 총 270개의 구간 시계열 변수를 가지며, 전체 측정 기간은 2017년 1월 1일부터 2018년 7월 9일까지 총 555일이다. 구간 시계열 변수 중 예측 대상 시계열 변수는 강변북로 한남대교 북단에서 반포대교 북단까지 구간의 속도이다.

시계열 변수 중 결측치가 1% 이하인 시계열 변수 150개를 선별하며, 선별된 시계열 변수에 대해 수식 (1)을 통해 제안된 전처리 기법을 수행한다. 수식 (1)은 시간 임베딩 기법(Time-Embedding) 중 피어슨 상관계수 계산을 통한 재배치 기법을 표현한다 [4].

$$C_n = pcc(V_t, V_n) \qquad (1)$$

수식에서 V_t, V_n 은 각각 예측 대상 시계열 변수와 나머지 시계열 변수를 의미한다. 해당 과정을 통해 시계열 변수 간 상관성 파악을 위해 피어슨 상관계수를 계산하고, 예측 대상 시계열 변수를 중심으로 시계열 변수를 차례대로 인접하게 배치한다.

2. 학습 모델

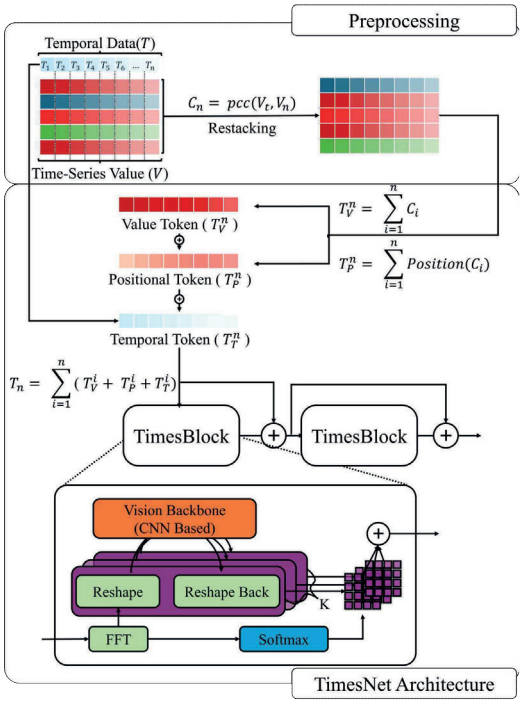

그림 1. 재구성된 TimesNet의 학습 모델 구성도

그림 1은 제안한 전처리를 포함한 TimesNet 모델의 다변량 시계열 데이터 학습 과정을 나타낸다. 먼저, 입력된 다변량 시계열 데이터를 시간 축과 시계열 데이터로 분리한다. 분리된 시계열 데이터는 제안된 전처리 기법을 활용하여 예측 대상 시계열 변수를 중심으로 나머지 시계열 변수들을 상관성에 따라 재배치한다. 이후, 분리한 시간 축은 시간 토큰(Temporal Token)으로, 재배치된 시계열 데이터는 위치 토큰(Position Token)과 값 토큰(Value Token)으로 변환하여 토큰들을 1차원 형태의 입력 토큰으로 합성한다. 모든 전처리 과정이 끝난 입력 토큰을 TimesNet 모델의 입력으로 사용하여 학습을 진행한다.

TimesNet 모델은 2개의 TimesBlock으로 구성되며, TimesBlock에 입력된 입력 토큰은 고속 푸리에 변환(Fast Fourier Transform)을 통해 주기 성분 상위 k개를 추출한 뒤 주기별로 분리된다. 이후 분리된 토큰은 k개의 2차원 형태 토큰으로 재구성하며, CNN 기반 레이어인 인셉션 레이어(Inception Layer)에서 합성곱 기반 특징 학습을 수행해 시계열의 시·공간적 패턴을 포착한다. 특징 학습이 끝난 토큰은 다시 1차원 형태의 토큰으로 재변환되며, 이전 단계의 출력 토큰과 합성되어 다음 TimesBlock의 입력으로 활용된다. 모든 TimesBlock을 통과한 최종 출력 토큰은 역정규화를 통해 다시 원래 형태로 변형되며, 변형된 출력 중 예측 대상 부분만을 추출하여 최종 결과로 출력한다.

3. 성능 평가

3-1. 실험환경

본 논문에서는 회귀 예측 성능을 비교하기 위한 지표로 평균제곱오차(Mean Squared Error), 평균절대오차(Mean Absolute Error)를 사용한다. 해당 지표들은 회귀 예측에 적합하며, 값이 작을수록 좋은 예측 결과를 낼 수 있는 모델로 판단한다. 또한, 모델의 설명력을 확인하기 위해 R^2 점수(R-Squared Score)를 사용한다. 해당 지표는 모델의 독립 변수와 종속 변수 간 관계성을 0과 1 사이의 값으로 표현하며, 값이 클수록 해당 모델이 독립 변수와 종속 변수 간 관계성을 잘 설명하는 것으로 판단한다.

해당 실험의 입력 시계열 길이는 예측 시점 이전으로 24시간으로 설정하였으며, 예측 길이는 예측 시점 이후 96시간으로 설정하였다.

최종 출력 레이어에서는 Softmax 활성화 함수를 사용하며, 최적화 함수는 Adam을 사용했다. 학습 과적합 방지를 위해 검증 손실이 3 Epoch 연속으로 개선되지 않을 경우 학습을 조기 종료하도록 설정하였다.

3-2. 실험 결과

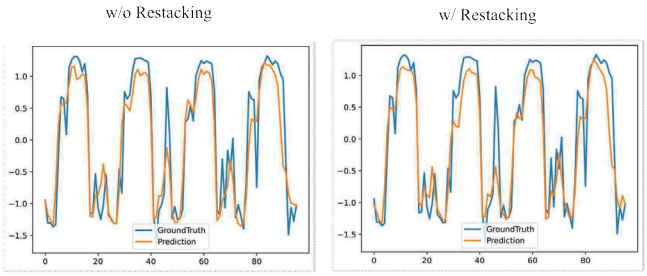

그림 2. 각 시계열 데이터 처리별 예측-실제값 그래프

그림 2는 입력 데이터의 전처리 과정에 따른 TimesNet의 예측 결과와 실제값을 그래프로 나타낸 것이다. 해당 그래프에서 가로축은 시간축을 나타내며, 세로축은 예측된 속도를 범위 조정하여 나타낸 값들이다. 예측 결과, 제안한 전처리를 적용할 경우, 상승 후 하락 구간에서 정답값에 더 근사하게 예측하는 것을 확인할 수 있다.

표 1은 제안된 전처리 과정 적용을 통한 다변량 시계열 학습 결과를 정리한 표이다. 위 결과를 통해 제안된 전처리 과정 진행 시 모델의 예측 정확도가 의미 있게 향상되는 것을 확인할 수 있다.

MSE가 0.3153에서 0.2899로 8.1 % 감소하고, MAE 역시 0.4016에서 0.3857로 4% 감소하였다. 그림 2와 위 결과를 비교 분석하였을 때, 제안한 전처리를 진행한 모델이 회귀할 때의 예측값과 실제값 간의 오차가 더 낮아져 정답값에 근사한 값 범위를 예측할 수 있다는 것을 알 수 있다.

Metric	MSE	MAE	R^2 Score
w/ Restacking	0.3153	0.4016	0.6887
w/o Restacking	0.2899	0.3857	0.7138

표 1. 각 시계열 데이터 처리별 예측 성능 지표

R^2 Score는 0.6887에서 0.7138로 3.6 %p 상승하였다. 결론적으로, 해당 전처리 과정은 예측 오차를 줄이고 전체 데이터의 패턴을 더욱 정확히 반영하여 모델의 예측 성능을 실질적으로 향상시킬 수 있음을 나타낸다. 또한 동일한 모델 구조와 하이퍼파라미터를 설정하여도 관측값의 분산을 더 잘 포착할 수 있음을 보인다.

Ⅲ. 결론

본 논문에서는 TimesNet 모델의 다변량 시계열 예측 성능 향상을 위한 전처리 기법인 재배치의 활용을 제안한다. 제안하는 전처리 기법을 활용한 결과 전반적인 오차가 줄었으며 예측 성능이 유의미하게 개선되었음을 확인하였다. 차후 연구에서는 제안된 전처리 기법에 특성을 활용하여 Convolution 연산을 줄여 낮은 복잡도와 높은 정확도를 가지는 경량화 모델을 설계할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00165, 5G+ 지능형 기지국 소프트웨어 모뎀 개발). 또한, 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1A2C2095289)

참 고 문 헌

[1] JungHun Byun, Ohyun Jo "LowComplexityCNN Model byUsing2D Imaging od Time Series Traffic Data," Proceedings of Symposium of the Korean Institute of communications andInformation Sciences, pp. 98-99, Aug. 2020.

[2] Jongseok Kim, Ohyun Jo "MuShAug: Boosting Sequence Signal Classificiation via Multishape Augmentation" IEEE Internet of Things Journal, pp. 32585 - 32597, Oct 2024

[3] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, Mingsheng Long, "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis," Proceedings of the 11th International Conference on Learning Representations (ICLR), Kigali, Rwanda, pp. 1 - 23, May 2023.

[4] Seungwoo Kang, Ohyun Jo "Multivariate Time-Series Imagification with Time Embedding in Constrained Environments (Student Abstract)" The Thirty-Eighth AAAI Conference on Artificial intelligence(AAAI-24), Vancouver, Canada, pp.23535 - 23536, Mar 2024