전리층 총 전자량 데이터에 적용한 LSTM 기반의 지진 이상현상 탐지

조건우, 박동건, 김홍국
광주과학기술원

joungju257@gist.ac.kr, dongkeon@gist.ac.kr, hongkook@gist.ac.kr

LSTM-based Earthquake Anomaly Detection Applied to Total Electron Current Data

Geon Woo Cho, Dong Keon Park, Hong Kook Kim
Gwangju Institute of Science and Technology

요 약

본 논문에서는 전리층 총 전자량(TEC, Total Electron Current) 데이터에서의 이상현상을 long short-term memory (LSTM) 기반 모델로 탐지하는 기법을 제안한다. 또한 제안된 방식은 Gaussian mixture model, K-means clustering, 그리고 support vector machine 의 기계학습 기반의 기법과 비교한다. 실제 지진 데이터와 비교하여 이상현상 탐지 성능을 비교한 결과, LSTM 기반의 이상현상 탐지 성능이 기존의 기계학습 기반의 성능과 비교하여 F1-score 로 약 20% 향상됨을 확인하였다.

Ⅰ. 서 론

전리층 총 전자량(TEC, Total Electron Current) 데이터는 지구 전리층에 있는 전자의 밀도를 나타내는 지표로, 지진과의 상관 관계는 꾸준히 논의되어 왔다[1]. 하지만, 규모 6.0 이상의 지진 중 특정한 사례에 대해서만 분석된 것이 대부분이었다[2, 3]. 본 논문에서는 시계열 데이터 학습에 사용되는 지도 학습 모델 중 하나인 long short-term memory (LSTM)[4] 기반의 모델을 이용한 TEC 데이터에서의 이상현상을 탐지하는 기법을 제안한다. 제안된 기법의 성능은 2016 년 1 년간 미국에서 지진이 가장 많이 발생한 경도 -117°~-120°, 위도 35°~40° 사이의 TEC 데이터를 이용하여 실제 지진 데이터와 비교하여 평가된다. 또한, 머신러닝 기법 중에 Gaussian mixture model (GMM), K-means clustering, 그리고 support vector machine (SVM) 기반의 기법과도 성능을 비교한다.

Ⅱ. 본론

2.1 TEC 데이터

TEC 데이터는 NOAA 에서 2016 년 1 월 1 일 0 시부터 12 월 31 일 24 시까지 15 분 간격 (총 35040 개), 위도와 경도를 1 도 간격으로 관측한 것을 사용하였으며[5], 위도, 경도 별로 관측된 TEC 데이터들에 평균을 취한 후 이용하였다. 단위는 $10\ \mathrm{TECU} = 10^{17}\ electrons/m^2$ 이다. TEC 데이터는 태양 활동에 영향을 크게 받아 계절별 분포 차이가 크기 때문에, 학습 정확도 향상을 위해 1~3 월, 4~6 월, 7~9 월, 10~12 월 4 분기로 나누어 사용하였고, 전체 데이터를 학습시킨 결과와 비교하였다.

전체 TEC 데이터의 70%를 학습 데이터, 나머지 30%를 평가 데이터로 분리하였다. [3]의 연구에서는 지진 발생 이전 3 일 안에 TEC 데이터에서 두드러지는 변화가 나타남을 보인 반면, 본 연구에서는 규모 4.5 이상의 지진에 대한 정보를 이용하였기에 본진이 끝난 후에도 여진이 있을 것을 감안하여 지진 전후로 3 일을 이상현상으로 레이블링하였다.

2.2 방법론

2.2.1 K-means Clustering

전체 데이터에 대해 군집의 수를 1 개부터 늘려가며 elbow point 로 군집의 개수를 6 개로 정하였다. K-means clustering 기법을 통해 6 개의 군집을 형성한 후, 군집들의 중심으로부터 가장 먼 거리에 위치한 군집에 속한 데이터들을 이상현상으로 분류하였다.

2.2.2 One-class Support Vector Machine

One-class SVM 에 기반한 학습 기법을 적용해 정상 학습 데이터에 대한 경계를 학습한 후, 전체 데이터를 넣어 학습된 경계 밖에 있는 데이터들을 이상현상으로 판별하였다.

2.2.3 Gaussian Mixture Model

전체 혼합 개수를 3 개로 설정한 후, expectation-maximization (EM) 알고리즘을 이용하여 정상 데이터에 대한 확률 분포를 추정하였고, 문턱치보다 높은 확률 값을 가지는 데이터들을 이상현상이라고 탐지하였다.

2.2.4 Long Short-Term Memory Model

70%의 학습 데이터 중 정상 데이터만 선택하여 LSTM 이 정상일 때의 분포를 예측할 수 있도록 학습시켰다. 학습된 파라미터를 이용해 10 단계 예측값을 도출하였고, 예측된 결과를 바탕으로 다음과 같은 Mahalanobis distance 기반의 anomaly score 를 정의하였다.

$$\text{Anomaly Score} = (e^{(i)} - \mu)\Sigma^{-1}(e^{(i)} - \mu)^T \quad (1)$$

여기서, μ와 Σ 는 평균 및 공분산을, e^i 는 i(범위: 0~35039)번째 index 에서의 오차를 의미한다. 식 (1)의 anomaly score 가 문턱치 값보다 높은 값을 가지고 있으면 이상현상이라고 판별하였다.

2.3 실험 및 성능 평가

데이터들을 학습시킬 때, 잡음이 약간 섞인 데이터에 대해서도 비슷한 결과를 얻을 수 있도록 아래와 같은 데이터 증강 기법을 적용하였다.

표 1. 잡음 인가 기반의 데이터 증강 기법

데이터 증강 기법
1: Total data = {Original data}
2: Noise ~ $N(0,1)$, $i = 0$
3: **While** $i < 0.05$
4: Total data ← Original data + $\sigma \times$ Noise $\times i$ (σ 는 Original data 의 표준 편차)
5: $i = i + 0.0005$

LSTM 의 학습 결과에 대한 예시는 그림 1 과 같다. 초록색 그래프는 1 단계 예측에 대한 결과이고, 파란색 그래프는 이전 단계에 대한 값을 주지 않고 재귀적으로 예측한 결과로, 정상 데이터에 대한 모델의 학습 정도를 확인할 수 있다.

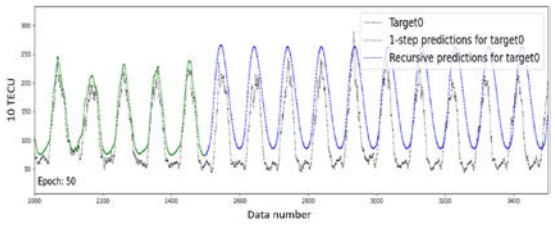

그림 1. LSTM 기반의 TEC 데이터 예측 (target0: 봄 데이터, 전체 데이터 번호의 범위: 0~35039)

LSTM, K-means clustering, SVM, GMM 을 이용해 이상 현상을 탐지한 결과에 대한 예시는 그림 2 와 같다.

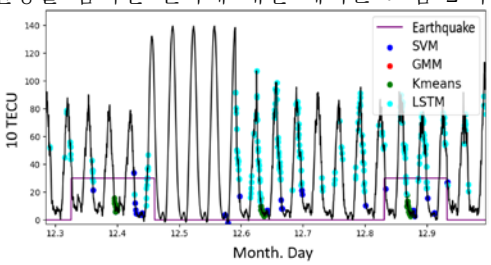

그림 2. LSTM, K-means clustering, SVM, GMM 기반의 이상현상 탐지 성능 비교

전체 데이터 35,040 개 중 이상현상이라고 표기된 값은 총 5,159 개로, 이상현상의 개수에 비해 정상일 때의 개수가 많은 불균형 데이터이기에, 정확도로는 예측 결과에 대한 분석을 정확하게 할 수 없었다. 따라서 precision 과 recall 의 조화 평균에 해당하는 F1-score 를 이용해 예측 결과를 분석하였다. LSTM 에서의 F1-score 는 anomaly score 에 대해 임의의 문턱치 값을 잡았을 때 얻어지는 F1-score 중 가장 큰 값으로 하였다. 예측 결과는 표 2 와 같다.

표에서 보는 바와 같이, F1-score 측정 결과, LSTM 기반의 기법의 F1-score 가 기존의 기계학습 기반의 기법보다 높고, 계절별 학습에 대해서는 LSTM 기반 기법이 이상현상을 F1-score 로 약 20%만큼 잘 예측하고, 전체 데이터에 대한 학습에서는 이상현상을 약 10%만큼 잘 예측함을 확인할 수 있었다. 또한, LSTM 을 이용해 학습시킬 시, 전체 데이터를 모두 학습시킨 기법에 비해 데이터를 계절별로 따로 학습시킨 기법이 이상현상을 약 15%만큼 잘 예측함을 확인할 수 있었다.

표 2. 각 기법별 단일모델과 계절별 모델의 F1-score 비교

Method	단일 모델	계절별 모델			
		spring	summer	autumn	winter
K-means	0.000	0.016	0.005	0.017	0.011
GMM	0.033	0.031	0.001	0.011	0.013
SVM	0.031	0.009	0.000	0.004	0.032
LSTM	0.121	0.197	0.218	0.223	0.378

Ⅲ. 결론

본 논문에서는 TEC 데이터에서의 이상현상 탐지를 위해 제안된 LSTM 기반의 기법의 F1-score 가 GMM, K-means clustering, 그리고 SVM 에 비해 좋음을 확인하였다. 또한, TEC 데이터의 계절별 분포 차이가 있음을 각 계절별 F1-score 를 통해 간접적으로 확인할 수 있었다. 하지만, LSTM 기반 기법의 F1-score 가 상대적으로는 높지만 절대적인 수치를 봤을 때 이상현상 탐지를 효율적으로 하지는 못함을 확인할 수 있었다. 이는 disturbance storm time (Dst) 지수, K(George) 지수에 의해 발생하는 잡음이 지진에 의해 발생하는 TEC 데이터의 변화보다 더 크거나 불규칙하게 발생할 때가 많아 학습이 제대로 이루어지지 않아 그런 것으로 판단된다. 추후 규모가 큰 지진이라는 제약을 두어 잡음의 영향을 상대적으로 줄이고, Dst, K 지수 등을 활용한 잡음제거 기법을 도입하는 등의 학습 방법을 도입해 LSTM 기반의 이상현상 탐지 기법의 성능을 높이고자 한다.

ACKNOWLEDGMENT

본 연구는 광주과학기술원 전기전자컴퓨터공학부 오디오지능연구실 인턴십의 결과이며, 2020 년도 광주과학기술원 GRI(GIST 연구원)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] M. Hayakawa, “Earthquake prediction with electromagnetic phenomena,” in *Proc. AIP Conference,* vol. 1709. no. 1. p. 020002, 2016.

[2] W. Liu and L. Xu, “Statistical analysis of ionospheric TEC anomalies before global M w ≥ 7.0 earthquakes using data of CODE GIM,” *Journal of Seismology*, vol. 21, no. 4, pp. 759-775, 2016.

[3] J. Y. Liu, Y. I. Chen, Y. J. Chuo, and H. F. Tsai, “Variations of ionospheric total electron content during the Chi-Chi earthquake,” *Geophysical Research Letters*, vol. 28, no. 7, pp. 1383-1386, 2001.

[4] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, “LSTM-based encoder-decoder for multi-sensor anomaly detection,” arXiv:1607.00148, 2016.

[5] N. G. D. Center, “Real-time US-Total Electron Content: Vertical and Slant” *NOAA National Centers for Environmental Information* (NCEI), 11-Jul-2006. (https://www.ngdc.noaa.gov/stp/iono/ustec/products/)