EGU24-8410, updated on 20 May 2024
https://doi.org/10.5194/egusphere-egu24-8410
EGU General Assembly 2024

Leveraging the Power of Graph Neural Networks in Environmental Time Series Anomaly Detection

Elżbieta Lasota[1], Julius Polz[1], Timo Houben[2,3], Lennart Schmidt[2,3], David Schäfer[2,3], Jan Bumberger[2], and Christian Chwala[1]

[1]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany

[2]Research Data Management (RDM), Helmholtz-Centre for Environmental Research (UFZ), Leipzig, Germany

[3]Department of Monitoring and Exploration Technologies, Helmholtz-Centre for Environmental Research (UFZ), Leipzig, Germany

Efficient quality control (QC) of time series data from environmental sensors is crucial for ensuring data accuracy and reliability. In this work, we turn to machine learning, specifically Graph Neural Networks (GNN), to elevate QC efficiency for large datasets originating from sparsely distributed sensors. Our proposed model, specifically tailored for anomaly detection as a vital aspect of QC, combines graph convolution (GC) and Long Short-Term Memory (LSTM) layers to capture both spatial dependencies and temporal patterns in the time series data. The focus on anomaly detection enables the identification of deviations or irregularities in the signal, providing insights into important events, faults, or disturbances within the data.

We conducted experiments using two distinct types of labeled data: three months of data in 2019 from 20 Commercial Microwave Links (CML) distributed around Germany and a 2.5-year period (June 2014 to December 2016) of soil moisture data from the TERENO SoilNet network in Hohes Holz, Germany. These datasets, encompassing an impressive 2.5 million samples, pose challenges in QC due to diverse dynamics, signal anomalies, and variations in temporal resolution and spatial densities of observations.

The classification results demonstrated satisfactory performance, with Matthews Correlation Coefficients of over 0.6 and 0.8 for the CML and SoilNet datasets, respectively. To evaluate the added value of processing the spatial information provided by neighboring sensors, we also compared the results of our final GNN with a baseline model that uses the same LSTM layers but disregards the GC layer, which integrates the neighboring information. The GNN model exhibited improved performance, as evidenced by 5-fold cross-validation mean Area Under the Receiver Operating Characteristic Curve (AUC) values of 0.934 and 0.971 for the CML and SoilNet data, respectively. In contrast, the baseline model yielded mean AUC values of 0.877 and 0.950, highlighting the effectiveness of incorporating the information from neighboring sensors via the GC layers to enhance anomaly detection for environmental sensor time series data.