Performance of machine learning algorithms for predicting air pollution parameters

Brianna Alexandra, Stan (School: Laude-Reut Educational Complex)

Air pollution is one of the world's most critical challenges today, generating numerous premature deaths and diseases and negatively impacting the environment and the overall quality of life. Furthermore, the development of commercial prototyping platforms has opened the way for citizen science projects in air quality monitoring and has offered the possibility for civic organizations to get involved. However, private networks often have the required hardware for monitoring, while the data analysis, identification of pollution sources, and pollution predictions based on long-term data are sparse or frequently missing. Therefore, forecasting air quality parameters is a solution to providing valuable information for local communities and authorities, helping them detect patterns, identify sources, and warn citizens against potential risks due to very high levels and sources.

This study analyzes and compares deep learning models like LSTM and GRU, which are recurrent neural networks generally used in this area, with ARIMA, SARIMAX, BVAR, VAR, and Prophet, not so common in the research literature about 24 hours air pollutants prediction. The project's goal was to test the performance of machine learning algorithms and develop a methodology for identifying the most accurate one aiming to integrate it with the urban community's air monitoring station networks for live prediction. The data used in the project contains measurements of different pollution parameters collected between 2001 and 2018 from 12 stations located in Madrid, Spain. In addition, a map of the station positions was generated using their coordination and OpenStreetMap API. All stations have identical batches of data saved in separate files and used in training to avoid biased results. The methodology started with one station by analyzing and cleaning the data and implementing a training model for each algorithm. Identifying the particularities of each model and the involved variables helped to reorganize the data to match all models. The implementing process continued by creating a unique dataset to be used by all models called processed data, changing each model implementation to use the same processed data on multiple stations, and training all models on the processed data. The source code is model-independent. Particular requirements of models for data packaging are generated in each model implementation.

The analysis of the prediction results of the models used MAPE, RMSE, MSE, and MAE as evaluation methods. The ARIMA and VAR models were the fastest to train but less accurate. Prophet model had the same medium duration training time but offered mixed results in terms of accuracy on PM2.5 and NO2. SARIMAX provides better results but has a long process to get the best combination of parameters, and BVAR was very time-consuming. The LSTM model was the most reliable for processing time-series datasets because of both long and short memory features. Another advantage of LSTM and GRU over other models was the possibility of being trained on GPU (graphics card).

Access to real-time information allows people and authorities to be informed about safe intervals for outdoor activities. Future work involves extending this research to other individual or combined algorithms, predicting other pollutants, and connecting the pollution predictions with climate data, including temperature, humidity, wind speed, and direction, for improved accuracy. Further improvements include launching a local air pollution monitoring network and generating prediction reports. Conducting accurate urban pollutants concentration forecasting is crucial for efficient air pollution prevention and control measures to protect citizens and involve communities in local action. Some of the limitations of this project enclose the hardware, which caused a lot of delays, the constraints of testing more algorithms, and the dimension of the datasets, causing a very long time to process them.