Prediction of air pollution with machine learning

Christian Schmitz[1] Dhiren Devinder Serai[2] Tatiane Escobar Gava[3]

Abstract: Cities worldwide are facing air quality issues, leading to bans of vehicles and lower quality of life for inhabitants. We forecast the air quality for Stuttgart based on expected weather condition. For that purpose, we extract, cleanse, and integrate the DHT22 and SDS11 sensors' data to feed two different machine learning models for predicting the particulate matter values for the near future.

1 Introduction

According to the World Health Organization (WHO) [WH16], urban air pollution increased by more than 8% between 2008 and 2013, despite all efforts on improving air quality in many countries around the globe. Urban air pollution may lead to a number of diseases, including reduced lung function, respiratory infections, and aggravated asthma. We live in Stuttgart, the city with the highest air pollution in Germany. Thus, we are directly affected. However, we do not want to risk our health more than necessary. Therefore, we propose a solution to predict the air quality situation for the next hours and days in the city center. This information may support various applications of value to society, e.g., four route planning or adapting daily habits.

2 Approach

Our initial solution predicts air pollution by integrating two data sources. The first source is sensor data from [OK15]. In Stuttgart's downtown area, 25 sensors collect data on temperature, humidity, and particle matter. The second data source provides data on weather forecast OpenWeather [Op18], and is used as input for the machine learning models in order to predict the air quality for that given scenario (weather forecast, date, and time).

From the air pollution sensor dataset, we collect all DHT22 sensors measurements for temperature and humidity information and SDS011 sensors' measurements for air quality indices (particle matter values, to be precise). We filter the dataset, so that it only contains values from Stuttgart's downtown area. We integrate the weather information from the DHT22 sensors with the air quality indices from the SDS011 sensors based on the location and time. We notice that the dataset contains many implausible sensor values. That is why we remove those data items which were measured under unsupported weather conditions. In the end, we have an integrated dataset containing records with particle matter values and

[1] Universität Stuttgart, IPVS, Universitätsstr. 38, 70569 Stuttgart, st160269@stud.uni-stuttgart.de
[2] Universität Stuttgart, IPVS, Universitätsstr. 38, 70569 Stuttgart, st161906@stud.uni-stuttgart.de
[3] Universität Stuttgart, IPVS, Universitätsstr. 38, 70569 Stuttgart, st160427@stud.uni-stuttgart.de

doi:10.18420/btw2019-ws-34

weather information. We use this dataset to train, test, and fine-tune our machine learning models. We for instance use recurrent neural network regression techniques such as Long Short-Term Memory (LSTM).

Further, we obtain relevant weather forecast data such as the hourly humidity and temperature and feed our trained model with the forecast values, so that we finally obtain an hourly particle matter forecast for the next days. We present the final result using a visualization adapted to a chosen application, e.g., a plot to visualize the forecast for the next hours.

3 Cloud Services

Our solution makes use of multiple cloud services. For data cleaning and integration, we use Google's Cloud Dataproc service, which allows us to run Spark jobs. For machine learning, we employ Google's Cloud Machine Learning Engine. Further, we extract relevant weather forecast information from [Op18].

4 Insights

Based on our initial data cleaning, integration, and visualization, we get credible insights on relevant features for our machine learning model. In Fig. 1, we display the hourly pollution for a Sunday and Monday in spring and summer. Multiple factors affecting the air pollution are visible here, e.g., the seasonal (weather) effect, traffic during rush hour, or variation depending on day of week.

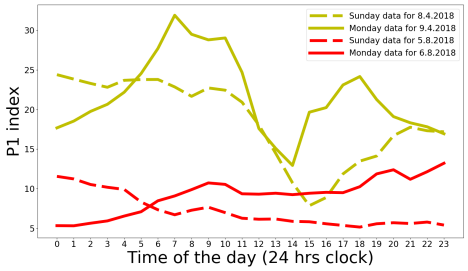

Fig. 1: Hourly air particle matter in summer and spring for a Monday and Sunday.

5 Next Steps

We currently have a cleaned and integrated dataset. We have also collected a list of important features for our machine learning approach. Nest, we will evaluate different recurrent neural network approaches such as LSTM and Gated Recurrent Unit (GRU) in order to find the most promising approach. The model for Stuttgart will then be used to implement a predictive application targeted to improve the quality of life of urban population.

Acknowledgements. We thank our advisors Ralf Diestelkämper and Melanie Herschel for their support on this project.

References

[OK15] OK Lab Stuttgart: Luftdaten Info, https://archive.luftdaten.info/csv_per_month/, Stand: 28.11.2018, 2015.

[Op18] OpenWeather: Weather API - OpenWeatherMap, https://openweathermap.org/api, Stand: 28.11.2018, 2018.

[WH16] WHO: Air pollution levels rising in many of the world's poorest cities, http://www.who.int/en/news-room/detail/12-05-2016-air-pollution-levels-rising-in-many-of-the-world-s-poorest-cities, Stand: 24.11.2018, 2016.