TECHNISCHE UNIVERSITÄT DARMSTADT

The Machine Learning Applications Data Challenge 21/22

Intelligent use of movement and shock data from freight wagons

Duration: 03.12.2021 – 21.02.2022

Background

With an increasing amount of sensor hardware onboard train vehicles, not only the shear amount of data, but also the level of detail of the gathered information is rising. DB Cargo as one of the leading rail network operators in Europe equips their trains with datalogger-devices that provide valuable information of various kinds. The analysis of the data can lead to detailed insights on the wagon's usage and the condition of trains and railroads. Also, the effectiveness of the trains' operation can be evaluated to answer specific business questions of DB Cargo.

However, data in the industrial sector is often stored in an unstructured manner and in large quantities which leads to difficulties in gaining a manual overview of it. Thus, advanced data mining approaches are needed in order to clean, process, and analyze data to transform the data into knowledge. Subsequently, the extracted smart data can be used to make decisions aimed at further increasing effectiveness and reducing costs of operation.

Datasets

The ***shunting data*** contains movement data of wagons with shunting operations in the Maschen train station within a 30 day period. In addition to the position data, information on the position acquisition, the wagon and train number (which can change during operation) and the movement status are given.

The ***long-term data*** contains movement data of various wagon types over the course of 52 weeks. The wagon types are linked to the given wagon numbers through an additional mapping sheet. The data provides an insight into the usage patterns of different types of wagons. The quality of the position data depends on the GNSS and mobile network quality and coverage.

Institut für Flugsysteme und Regelungstechnik

Institute of Flight Systems and Automatic Control

Prof. Dr.-Ing. Uwe Klingauf

Otto-Berndt-Str. 2
64287 Darmstadt

mla@fsr.tu-darmstadt.de

WS 2021/2022

TECHNISCHE UNIVERSITÄT DARMSTADT

The Data Challenge

As a team you have to work on all three parts of the data challenge (I, II and III).

Part I: Data Exploration and Statistics

In the first part you will explore the *long-term dataset* and answer several questions on its basic structure. Find out the meaning of the various variables, their structure and boundaries as well as the main statistical parameters. This step will provide you with the necessary overview of the several datasets.

Main objective: Identify the usage profile of different wagon types within the *long-term dataset*. Apply explorative data analysis and statistics. Get started with the following questions:

- How often are the wagons moved? Which differences can be identified between the wagon types?
- Are there seasonal dependencies?
- How much distance do the wagons travel loaded, how much empty? How far do they travel in different countries?
- Does the battery capacity of the telematic device change during the year, what might be the reasons?
- Use your understanding of the data to derive at least one further insight and explain its significance for DB Cargo.

Helpful tools for Part I might be:

- Statistics / Explorative data analysis (max/min/mean/deviations…)
- Plots: timeseries, distributions, boxplots…
- Heat maps

Try to automate your analysis toolset as much as possible by writing functions and scripts that aggregate the information of various data files. One of the core challenges will be the handling of the large volume of data!

Part II: Data Mining

In the second part you will analyze the *long-term dataset* and apply more sophisticated data mining techniques. Try to combine the question with your understanding of the DB Cargo business. How can the answer to the task help DB Cargo specifically?

Main objective: Analyze the Mobile network and GNSS quality and coverage and show its structure in different regions in Europe within the *long-term data*.

- Identify and combine relevant features (e.g. time stamps of data generation and transmission) and position data for your analysis
- Analyze and plot the statistical behavior of your features
- Be creative on the creation of new features as indicators for the Mobile network quality
- Fuse the data with geographic information and display your results in a geo-specific heatmap
- Be aware of appropriate aggregation and clustering of the high volume data
- Evaluate the errors and uncertainties of your analysis
- Interpret anomalies and trends in the analysis results
- *Optional:* The integration of OpenRailway maps can help to interpret your results
- If data-related difficulties or impossibilities arise concerning the task, describe reasons and possible solutions

TECHNISCHE UNIVERSITÄT DARMSTADT

Part III: Machine Learning Application

In the third part you will concentrate on the *shunting-data* and you will focus on the identification of the different shunting phases: parking, shunting or driving. The provided data has no labels attached – therefore you have to apply unsupervised Machine Learning and use your gained knowledge of the wagon's usage from Part I and II.

Main objective: Cluster / classify the operational states of the shunting process: those could include whether a wagon is incoming, parked, shunted, positioned on the target track or outgoing, but it's up to you to find the most useful categories for your classification model.

- Analyze the existing database
- Extract relevant features for your clustering model
- Develop a clustering model: you might gain a first overview with features depending on the geofence or the wagon moving state
- Evaluate the errors, uncertainty and robustness of the models and try to validate the models
- Is your algorithm also transferable on the *long-term dataset*? Describe the challenges, problems and critically discuss the quality of the results
- If data-related difficulties or impossibilities arise concerning the task, describe reasons and possible solutions