Data-driven machine learning and optimization pipelines for real-world applications

Koch, M.

Author: Koch, M.
Title: Data-driven machine learning and optimization pipelines for real-world applications
Issue Date: 2020-09-01

English Summary

Machine Learning is becoming more and more a substantial technology for industry. It allows to learn from previous experiences in an automated way to make choices based on the learned experience. This principle can be used to optimize and / or automatize processes in industry applications. A suitable process from the car industry is for example the quality assessment in assembly lines. Instead of assessing the quality in a time-consuming manual process, machine learning based image recognition methods could be applied to detect failures. Next to enhancing the existent, machine learning enables the development of completely new products like autonomous driving or services which are purely driven by data. The development of such new data-driven products is often a long procedure. Even the application of machine learning algorithms to specific problems is mostly not straightforward. To illustrate this, we introduce a data-driven service from the automotive industry called Automated Damage Assessment. Based on the gained experience from such data-driven service developments, we propose a methodology to develop data-driven services in an accurate and fast manner.

Especially data-driven services in the automotive domain could be based on sensor data, i.e. data which is recorded from on-board sensors over time in a so-called time series. In many cases, time series from more than one sensor can be used which is a so-called multivariate time series. The existent methods to solve multivariate time series classification-problems are often complex and developed to solve a specific problem without being scalable to solve various problems. To overcome this, suitable approaches with different complexities, applied on multiple publicly available data sets, as well as medical and industrial data sets are proposed in this work. As starting point, we have designed a plain feature-based pipeline and applied and enhanced it on several real-world data sets from the automotive and medical domains. Recently, numerous AutoML methods have been proposed. AutoML

aims at building optimized models in an automatized way. We have empowered those promising AutoML methods with another method to solve multivariate time series problems with the result that some of those techniques are suitable for this task. We have compared all those approaches on several publicly available data set with the results that especially two AutoML approaches, namely GAMA and ATM, as well as our PHCP approach are most suitable to solve this particular multivariate time series problems.

In order to consider the interaction of features from different time series, an approach is developed based on the evolutionary algorithm technique called genetic programming. It uses the principle of biological evolution to compute combination of features. Based on our results we can state, that the assumption, that combining several features can improve the performance instead of using the pure features, is true.