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

Koch, M.

CHAPTER 1

Introduction

1.1 Background

The recent progress on machine learning topics has sparked many research projects in science and industry. Many science projects focus on the development of new algorithms or methods and the industry often applies those but tries to improve and automatize them in order to save resources and to create additional business values. Examples are the use of historical car sharing customer data to place shared cars in cities in more efficient ways, or quality assessments with image recognition methods in assembly lines. Numerous industrial applications are imaginable, however, a limiting factor can be a lack of the necessary data and the huge effort to generate it. Nevertheless, many industrial machines or devices already generate data with their equipped sensors which are usually implemented for, e.g., machine monitoring. The data of such sensors are often recorded over time as a so-called time series. Using time series for machine learning can be a tedious and time consuming procedure as they can be and often are of highly complex nature. Next to time series, machine learning models can be used with very different types of data like images, videos, audio, writing, categorical features and numerical features. This results in a huge variety of required knowledge needed for such projects. Due to this variety of knowledge combined with the lack of data experts, industry often calls for simple but good performing modeling techniques which can also be applied by non-experts. In other words, industry requires simple, comprehensible and reliable techniques which solve real-world problems with good performances in an automated manner. Such automated methods belong to the group of so-called Automated Machine Learning (AutoML).

When developing a machine learning model in an industrial context, it is important to consider that the machine learning model itself is only one part of a probably long process chain. Therefore, it needs to be implemented seamlessly in the end-to-end process to be able to exploit all its potentials. This can be challenging because it often requires an adaption or recreation of existing processes.

1.2 Objectives

The industry often demands automated approaches which contain sophisticated machine learning techniques for data-driven projects. Automated methods which are easy-to-use and applicable to real world problems are rare, especially for processing time series. Since devices or machines are often equipped with sensors, time series are being recorded already. The recordings of more than one time series from, e.g., one acceleration sensor and one pressure sensor, is a called multivariate time series. Whereas the data of a single sensor recording is called univariate time series.

This work is mainly devoted to multivariate time series. Therefore, an extensive study of methods for multivariate time series is conducted. Based on this, novel time series classification pipelines with more or less complex methods are developed and verified, as well as other state-of-the-art methods like using AutoML methods as part of the pipeline. Pipelines can build optimized models in an automated approach and enable dealing with machine learning topics without time-consuming modeling, even for non-experts. Beyond that, this work proposes novel techniques based on genetic programming and neural networks. The algorithms are tested on both academic and real-world problems.

Another main objective of this work is to develop an end-to-end methodology to build data-driven services in an automotive environment. This methodology earmarks the implementation of earlier mentioned pipelines. As a real-world problem in the automotive industry, we investigate a data-driven service to assess the damage caused by an accident, i.e. based on vehicle on-board data this model can compute the damaged parts of a vehicle involved in a crash. Amongst others, in the development of the data-driven model of this service the proposed time series modeling algorithms are tested on this real-world problem. The relevant research questions (RQ) of this work are stated in the following:

RQ1 Can we develop efficient time series classification approaches by using feature-based techniques? (in terms of the trade-off between classification quality and computation time)

RQ2 Can AutoML methods be applied to solve time series classification tasks?

RQ3 Are state-of-the-art feature selection methods suitable for time series classification tasks based on a massive number of extracted features?

RQ4 Can the classification quality be increased when combining features with genetic programming?

RQ5 Can we develop a systematic approach for efficiently deploying data-driven services in the industry?

These questions are answered in this dissertation with algorithmic contributions and the linking of most promising existing methods. All methods are applied on academic data sets for a solid comparison. In order to evaluate the industrial use, the methods are also applied on industrial data sets.

1.3 Outline of the Dissertation

In this section, the outline of this dissertation with its motivation and content is briefly described. After each chapter the corresponding publications are listed.

Chapter 2 provides an overview of an industrial project called automated damage assessment. The objective of this project is to estimate the damaged parts caused by a low speed vehicle crash without any additional sensors, i.e. it is only based on already existing sensor data like acceleration signals. Such a digital damage assessment service would support the vehicle customer in the settlement of an accident. The business motivation and the technical challenges are presented.

As part of this dissertation ideas of the invented damage assessment system are summarized in the following patent application:

> Koch, M. and Hundt, W. and Malotta, J. and Godau, R. and Geiger, M. and Krieger, J. (13.06.2019). A Method of Determining Damage Occurring in an Accident between a Vehicle and a Collision Partner on the Vehicle. *World Patent Application WO2019110434A1.*

Chapter 3 introduces the used methods in this dissertation. Traditional machine learning models, neural networks, state-of-the-art AutoML methods, as well as evolutionary algorithms are described. Furthermore, the challenges of feature engineering when dealing with time series, its methods and hyperparameter optimization are discussed. Parts of this chapter are published in the following article:

> Koch, M. and Wang H. and Bürgel R. and Bäck, T. (2018). A Comparison of Hand-Crafted and Automated Machine Learning Approaches for Multivariate Time Series Classification. submitted.

Chapter 4 proposes novel methods for time series classifications. A first method called plain hand-crafted pipeline represents a practical approach which aims at creating good performing models with high efficiencies. Another pipeline method is based on genetic programming with the objective to create complex high level features from many low level features. Furthermore, hand-crafted and state-of-the-art neural network designs are introduced for this purpose. Parts of this chapter are published in the following article:

> Koch, M. and Wang H. and Bürgel R. and Bäck, T. (2018). A Comparison of Hand-Crafted and Automated Machine Learning Approaches for Multivariate Time Series Classification. submitted.

Chapter 5 shows three applications of the plain hand-crafted pipeline on real-world data sets, namely on two automotive on-board data sets and one medical data set. The data set of the first application contains recorded vehicle on-board data from crash tests with the locational crash information on the car. The objective is to estimate the impacted location. In the second application, based on recorded vehicle on-board data from crash tests which were carried out within this work, damaged parts caused by crash events are estimated. In the third application, data from Parkinson's disease patients is used to identify the usefulness of a certain surgery to enhance the patients well-being.

These three applications of the plain hand-crafted pipeline were previously published in:

> Koch, M. and Bäck, T. (2018). Machine Learning for Predicting the Impact Point of a Low Speed Vehicle Crash. In *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications*, ICMLA '18, Orlando, USA, pp. 1432–1437.

Koch, M. and Wang, H. and Bäck, T. (2018). Machine Learning for Predicting the Damaged Parts of a Low Speed Vehicle Crash. In *Proceedings of the 13th International Conference on Digital Information Management*, ICDIM '18, Berlin, Germany, pp. 179–184.

Koch, M. and Geraedts V. and Wang H. and Tannemaat M. and Bäck, T. (2018). Automated Machine Learning for EEG-Based Classification of Parkinson's Disease Patients. IEEE Big Data '19, Los Angeles, USA, pp. 4845–4852 (2019).

Chapter 6 shows the comparison of all proposed methods for multivariate time series classification, as well as promising state-of-the-art AutoML approaches implemented in the pipeline. As basis for the comparison, 18 publicly available multivariate time series data sets and the data set containing on-board data from crash events with its damaged parts are used. Especially with the latter mentioned data set the industrial uses of the methods are investigated. The methods are compared based on the performance and the efficiency (computation time). This extensive comparison is based on the following publication:

Koch, M. and Wang H. and Bürgel R. and Bäck, T. (2018). A Comparison of Hand-Crafted and Automated Machine Learning Approaches for Multivariate Time Series Classification. submitted.

Chapter 7 proposes a novel method to develop data-driven services in the automotive industry. The damage assessment system is used as an exemplary data-driven service. The proposed methodology aims at giving an overview of key points regarding business motivations, as well as the technical challenges when creating successful services. This work has been published in this article:

Koch, M. and Wang, H. and Bürgel and Bäck, T. (2020). Towards Data-driven Services in Vehicles. In *Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems*, VEHITS '20.

Next to the mentioned publications above, another publication of the author regarding this topic is:

Geraedts V. and Koch M. and Contarino M. and et al. (2020). Automated EEG-based Machine Learning for Cognitive Profiling during the DBS Screening in Parkinson's Disease Patients. submitted.