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

Koch, M.

CHAPTER 8

Conclusions and Outlooks

8.1 Conclusions

This thesis deals with several important aspects regarding automated machine learning techniques for time series in depth, as well its end-to-end implementation into data-driven industrial application.

To answer the research question (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)*, in this work hand-crafted machine learning techniques for time series data are developed with the result that feature-based methods can solve such problems in a competitive way. The use of feature-based techniques is sometimes required by the industry because of the need of tamper-proof techniques, i.e. a machine learning model must be explainable. In this context feature-based techniques are very suitable due to the descriptive nature of the features.

Next to hand-crafted approaches, in this work the practical use of state-of-the-art AutoML methods are investigated. This leads to the next research question (RQ2), *Can AutoML methods be applied to solve time series classification tasks?* with the result that, based on an extensive comparison, especially some of those methods are suitable for solving time series classification tasks in an efficient way.

When using generalized and feature-based approaches for time series classification, a massive number of time series features is computed. From this feature space the most significant features are selected in the following step. In order to generate high performing models, it is key to find the most relevant features in the massive feature space. Compared to common machine learning tasks, the feature space

for such time series problems is much larger due to the massive extraction which complicates the search problem for significant features. This leads to the next research question (RQ3): *Are state-of-the-art feature selection methods suitable for time series classification tasks based on a massive number of extracted features?*. As result of automatically computed features, especially one method (Boruta) can cope with this problem, even when we believe that some adaptions and optimizations of this methods to this purpose would enhance the performance.

When approaching multivariate time series classification with feature-based methods, it is often beneficial to consider also combinations of features from different time series. This is a challenging problem, because due to the large feature space many different combinations are possible. This issue leads to the next research question (RQ4): *Can the classification quality be increased when combining features with genetic programming?*. To answer this question a tree-based genetic programming approach was developed to combine several features to one complex feature which results in an enhancement of the model performance measures.

Beyond the research questions related to machine learning topics, this work deals with the application of the associated algorithms to the automotive environment. This leads to the next research question (RQ5): *Can we develop a systematic approach for efficiently deploying data-driven services in the industry?* From experience we present the challenges and key points when developing data-driven services and based on this, we developed a methodology to efficiently build such services. This methodology is proposed in this work.

8.2 Outlook

Some of the developed and shown methods show solid results on several academic and real-world data sets. When following the objectives of this work to develop high performing methods with reasonable computation times, parts of each pipeline can be further developed. The neural network CNN+LSTM architecture offers an efficient approach when accepting to sacrifice explainable models. The neural network avoids the whole feature engineering process. The architecture can be further developed, maybe with an even deeper structure when dealing with larger data sets.

The combination of features from different time series is investigated in this work with a genetic approach. Using the massive extracted feature space from the time series is not really feasible due to overwhelming number of feature combinations. Therefore, we reduce the initial feature space by PCA. Even when PCA is an efficient way, information can get lost. Hence, other ways to reduce the dimensions of the initial features without dropping critical information need to be investigated. A promising way could be a neural network.