Automated Machine Learning für Zeitreihen am Beispiel der Klassifikation von sequenziellen Fahrzeugdaten

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Shamil Nabiyev

Hochschule Darmstadt
Fachbereiche Mathematik und Naturwissenschaften & Informatik

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Motivation

Moderne, vernetzte Autos generieren laufend große Mengen an Mobilitätsdaten, die das Fahrverhalten, die Emissionen, den Standort und die Umweltbedingungen beschreiben. Die großen Datenmengen werden auf dem Auto mit Hilfe des Fahrzeugdiagnosesystems ausgewertet und für die weitere Diagnose von technischen Problemen zwischengespeichert. Diagnose-Fehlercodes (englisch: *Diagnostic Trouble Codes*, kurz *DTCs*) in sequenziellen Fahrzeugdaten weisen auf technische Probleme hin. Diagnosesysteme im Fahrzeug protokollieren jedoch nicht alle Fehlercodes. Eine weitere Schwierigkeit besteht darin, dass nicht alle aufgetretenen Fehlercodes mit einem Zeitstempel versehen werden. In diesem Fall ist es nicht möglich zu nachzuvollziehen, wann ein bestimmter Fehler im Fahrzeug aufgetreten ist. Die Analyse von solchen sequenziellen Automobil-Daten ist jedoch zeit- und ressourcenaufwendig. Sie erfordert außerdem jahrelange Erfahrung und Expertise in diesem Bereich. Um diesen Aufwand zu minimieren, können moderne Machine-Learning-Methoden verwendet werden. Insbesondere das Automated Machine Learning (AutoML) bringt große Potenziale für die automatisierte Verarbeitung von Daten und bietet die Lösung für Regressions- und Klassifikationsprobleme. In der vorliegenden Arbeit soll untersucht werden, welchen Beitrag eine ausgewählte AutoML-Lösung bei der Auswertung von sequenziellen Fahrzeugdaten leistet.

Ziel

Ziel der vorliegenden Arbeit ist es, zunächst durch Literaturrecherche einen Überblick über die existierenden AutoML-Lösungen für Zeitreihen zu schaffen. Als Nächstes soll basierend auf der Literaturrecherche eine ausgewählte AutoML-Lösung für die Zeitreihen-Klassifikation in die bestehende Data-Science-Plattform, Ontologie-basiertes Meta AutoML[a] [1, 4], integriert werden. Dabei geht es darum, ein neues Software-Modul für die Plattform zu konzipieren und das Konzept umzusetzen. Zusätzlich sollen die bestehenden Software-Module der Plattform erweitert und angepasst werden. Abschließend soll die Data-Science-Plattform für die Klassifizierung von sequenziellen Fahrzeugdaten angewandt werden.

[a] MetaAutoML, GitHub Repository (2022). URL: https://github.com/hochschule-darmstadt/MetaAutoML

Ergebnisse

Durch die Literaturrecherche wurden existierende AutoML-Lösungen für die Zeitreihen-Klassifikation, -Prognose und -Clustering gefunden. Das Ergebnis zeigt, dass drei open-source AutoML-Lösungen die Klassifikation von Zeitreihen ermöglichen: AlphaD3M, EvalML und Mcfly [2, 3]. Weitere zwei kommerzielle AutoML-Lösungen unterstützen die Zeitreihen-Klassifikation: Amazon SageMaker AutoPilot und DataRobot AI Cloud. Der Großteil von theoretisch verglichenen AutoML-Lösungen ermöglichen die Zeitreihen-Prognose. Durch die Literaturrecherche wurde nur eine AutoML-Lösung für die Zeitreihen-Clustering gefunden: DataRobot AI Cloud.
Das Mcfly AutoML [3] wurde in die bestehende Data-Science-Plattform, Ontologie-basiertes Meta AutoML, integriert. Somit ermöglicht die Plattform die Klassifikation von Zeitreihen.
Abschließend wurde die Plattform für die Klassifikation von sequenziellen Fahrzeugdaten angewandt. Bei dem Experiment auf dem Teildatensatz VIN4 wurde ein F1-Score von 0,70 erzielt. Obwohl die 94% der Diagnose-Fehlercodes mit Hilfe von AutoML detektiert wurden (Recall von 0,94), war die Falsch-Positiv-Rate in dem Experiment hoch (Precision von 0,57).

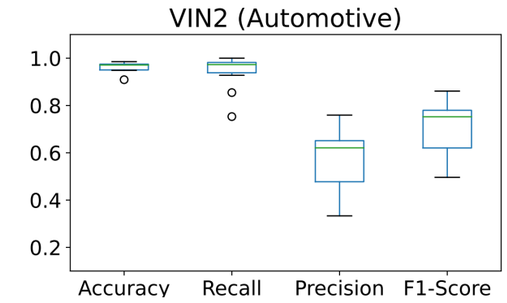

Abb. 1: Performance-Metriken von Machine Learning Pipelines für den Automotive-Datensatz. Die Boxplots repräsentieren Metriken aus der 10-fachen Kreuzvalidierung

Fazit

Durch die Anwendung der Data-Science-Plattform kann der Aufwand bezüglich der Erstellung von Machine-Learning-Modellen für die Zeitreihen-Klassifikation sowie der Hyperparameteroptimierung verringert werden. Die in die Plattform integrierte AutoML-Lösung basiert auf Deep-Learning Netz-Architekturen. Daher ist längere Laufzeit für die Optimierung von Machine Learning Pipelines zu erwarten.

Literatur

[1] Humm, B. G., and Zender, A. An Ontology-Based Concept for Meta AutoML. In *ARTIFICIAL INTELLIGENCE APPLICATIONS AND INNOVATIONS*, I. Maglogiannis, J. Macintyre, and L. Iliadis, Eds., vol. 627 of *IFIP Advances in Information and Communication Technology*. Springer Nature, [S.l.], 2021, pp. 117–128.

[2] Van Kuppevelt, D., Meijer, C., Huber, F., Van der Ploeg, A., Georgievska, S., and Van Hees, V. T. Mcfly: Automated deep learning on time series. *SoftwareX 12* (2020), 100548.

[3] Van Kuppevelt, D., Meijer, C., Huber, F., Van Hees, V., Solino Fernandez, B., Bos, P., Spaaks, J., Kuzak, M., Hidding, J., and Van der Ploeg, A. mcfly: deep learning for time series, 2020. URL: https://zenodo.org/record/3968518.

[4] Zender, A., and Humm, B. G. Ontology-based Meta AutoML. *Integrated Computer-Aided Engineering* (2022), 1–16.

Vorgelegt von Shamil Nabiyev
E-Mail: shamil.nabiyev@stud.h-da.de

Referent: Prof. Dr. Bernhard G. Humm
Korreferent: Prof. Dr. Christoph Becker