

Hochschule für angewandte Wissenschaften Coburg  
Fakultät Elektrotechnik und Informatik

Studiengang: Informatik-Bachelor

Bachelorarbeit

Klassifikation von Verkehrsteilnehmern auf Basis realer Positionszeitreihen mit Verfahren des maschinellen Lernens

Lennart Köpper

Abgabe der Arbeit: 10. 09. 2023

Betreut durch:  
Prof. Dr. Thomas Wieland, Hochschule Coburg

# Abstract

# Inhaltsverzeichnis

Abstract 2

Inhaltsverzeichnis 3

Abbildungsverzeichnis 5

Tabellenverzeichnis 6

Programmcodeverzeichnis 7

Abkürzungsverzeichnis 8

1 Einleitung 9

2 Theoretischer Hintergrund 10

2.1 Global Navigation Satellite System 10

2.2 Map-Matching 10

2.3 Maschinelles Lernen und Klassifikation 10

2.4 Klassifikationsverfahren 10

2.4.1 Klassifikator 1, z.B. Support-Vector-Machine 10

2.4.2 Klassifikator 2, z.B. Entscheidungsbäume/Random-Forest 10

2.5 Klassifikation mit Künstlichen Neuronalen Netzen 10

2.5.1 Grundlagen 11

2.5.2 Feed-Forward-Netze 11

2.5.3 Rekurrente Neuronale Netze 11

2.6 Bewertungsmaße für Klassifikatoren 11

2.6.1 Konfusionsmatrizen 11

2.6.2 Accuracy / Genauigkeit 11

2.6.3 Precision / Relevanz 11

2.6.4 Recall / Sensitivität 11

2.6.5 F1-Score 11

2.7 Eingesetzte Technologien und Frameworks 11

2.7.1 Python 11

2.7.2 ScikitLearn 12

2.7.3 Keras 12

3 Verwandte Arbeiten 13

3.1 Vorangegangene Abschlussarbeiten 13

3.2 Vehicle Classification from Low-Frequency GPS Data with Recurrent Neural Networks 13

3.3 Vehicle Classification using GPS Data 13

4 Datengrundlage 14

4.1 Gewinnung der Ausgangsdaten 14

4.2 Beschreibung der Ausgangsdaten 14

4.3 Weiterverarbeitung zu Bewegungsdaten 14

5 Anforderungen und Gesamtkonzept der Klassifikation 15

6 Umsetzung des Map-Matchings 16

6.1 Einrichtung der Valhalla-Engine 16

6.2 Umsetzung und Evaluierung des Road-Snappings 16

6.3 Vorklassifikation zur Bestimmung des korrekten Matching-Modus 16

6.3.1 Erzeugung des Trainingsdatensatzes 16

6.3.2 Auswahl des Klassifikators 16

6.3.3 Evaluierung 16

7 Umsetzung der Klassifikation 17

7.1.1 Erzeugung der Trainingsdatensätze (inkl. Vorverarbeitungen) 17

7.1.2 Training und Optimierung der Modelle 17

7.1.3 Vorstellung der besten Klassifikatoren 17

8 Evaluierung und Diskussion 18

9 Zusammenfassung und Ausblick 19

Quellenverzeichnis 20

Anhang A 1. Test 21

Anhang A 2. Test 2 22

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Programmcodeverzeichnis

# Abkürzungsverzeichnis

# Einleitung

**Motivation und Projekthintergrund**

Brücke über das autonome Fahren (erweiterte Umfeldwahrnehmung) als mögliches Einsatzgebiet. 🡪 Forschung in diesem Gebiet auch Teil des Projektes 5GKC an welchem die Hochschule beteiligt ist.

**Zielsetzung und Forschungsfragen**

Ziel: Klassifikation von Verkehrsteilnehmern auf Basis ihrer sequenziell bereitgestellten Positionsdaten durch den Einsatz maschineller Lernverfahren.

Frage 1: Wie können reale Positionsdaten, die Ungenauigkeiten und Rauschen aufweisen, so vorverarbeitet werden, dass sie sich gut für den Einsatz maschineller Lernverfahren eignen? Frage 2: Welche Verfahren des maschinellen Lernens sind für die Klassifizierung von Verkehrsteilnehmern auf Basis von sequenziellen Positionsdaten geeignet?

**Aufbau der Arbeit**

Beschreibung der Gliederung.

# Theoretischer Hintergrund

## Global Navigation Satellite System

Bildet technische Grundlage für das gesamte Vorhaben (keine Positionsdaten ohne GNSS). Gewisse Umstände, wie die nach wie vor existierenden Ungenauigkeiten in GNSS-Daten bilden Grundlage für den Einsatz von Vorverarbeitungen wie Map-Matching.

## Map-Matching

Vielversprechender Ansatz in der Datenvorverarbeitung. Damit die Idee klar wird, sollte der Ansatz theoretisch erläutert werden.

## Maschinelles Lernen und Klassifikation

Was ist maschinelles Lernen? Wo ist mein Problem darin einzuordnen? -> Klassifikationsproblem; überwachtes Lernen

## Klassifikationsverfahren

Für den Fall, dass „einfachere“ Klassifikationsverfahren im Zuge der Umsetzung (insbesondere bei der Vorklassifizierung) und damit auch Evaluierung zum Einsatz kommen.

### Klassifikator 1, z.B. Support-Vector-Machine

### Klassifikator 2, z.B. Entscheidungsbäume/Random-Forest

## Klassifikation mit Künstlichen Neuronalen Netzen

Angestrebte Art der Klassifikation (insbesondere Hoffnungsträger bei sequenzieller Klassifikation motorisierter Fahrzeuge) in dieser Arbeit. Funktionsweise von KNNs ist alles andere als trivial und sollte deshalb erläutert werden.

### Grundlagen

### Feed-Forward-Netze

### Rekurrente Neuronale Netze

## Bewertungsmaße für Klassifikatoren

Um die Qualität eines Klassifikationsverfahrens messbar zu machen und durch Training und Parameteroptimierung zu steigern, ist es wichtig Bewertungsmaße zu definieren, die die Vorhersageleistung eines Klassifikators widerspiegeln. In diesen Kapitel sollen solche Maße definiert und erläutert werden, wobei zum besseren Verständnis in allen Fällen schrittweise von einer binären Klassifikation auf ein n-klassiges Problem verallgemeinert wird.

### Konfusionsmatrizen

Ein verbreitetes Werkzeug, um die Vorhersageleistung eines Klassifikators zu visualisieren, ist die *Konfusionsmatrix*. Hierbei handelt es sich um eine Tabelle, bei welcher für eine Menge von Datenpunkten, die Schnittmengen aller vorhergesagten und tatsächlichen Klassen ausgezählt werden. Hierbei stehen typischerweise die Zeilen der Matrix für die tatsächlichen Klassen, wohingegen die Spalten die Vorhersagen repräsentieren [vgl. Géron 2020, 94]. Abb. X zeigt ein Beispiel für eine Konfusionsmatrix zur Auswertung eines binären Klassifikators auf Basis von 65 Datenpunkten.



Abb. : Beispiel für eine zweiklassige Konfusionsmatrix

Konfusionsmatrizen können eine beliebige Anzahl von Klassen umfassen. Wichtig ist lediglich, dass deren Anordnung in den Zeilen und Spalten identisch ist, damit alle korrekt klassifizierten Datenpunkte auf der Hauptdiagonalen der Matrix widergespiegelt werden. Ein guter Klassifikator enthält also außerhalb dieser Hauptdiagonalen eine möglichst geringe Anzahl an Datenpunkten [vgl. Grandini et al. 2020, 2].

### Accuracy / Genauigkeit

Die *Accuracy* (dt. Genauigkeit) ist eine häufig genutzte Metrik bei Klassifikationsproblemen. Sie misst den Anteil der korrekt klassifizierten Datenpunkte an der Gesamtzahl aller betrachteten Datenpunkte. Für eine binäre Klassifikation ergibt sich somit die folgende Formel:

Bei n-klassigen Klassifikationsproblemen kann man die Accuracy berechnen, indem man zunächst die Accuracy-Werte für alle Klassen im Einzelnen und anschließend das arithmetische Mittel dieser Werte berechnet. Hiermit ergibt sich folgende Formel:

Die Accuracy ist ein sehr intuitives Maß. Unabhängig von der Anzahl der Klassen liegt sie (wie auch alle nachfolgenden Maße) immer im Intervall [0, 1], wobei ein Klassifikator umso besser ist, desto näher der Wert an 1 liegt. Allerdings sollte man vorsichtig sein, die Accuracy einzusetzen, falls der zugrundeliegende Datensatz unausgeglichen ist, da sie dazu tendiert große Klassifikationsfehler in unterrepräsentierten Klassen zu verbergen [vgl. Grandini et al. 2020, 3f.]. Für die Matrix in Abb. X berechnet sich eine Accuracy von 0,8.

### Precision / Relevanz

Die *Precision* (dt. Relevanz) misst die Genauigkeit der positiven Vorhersagen. Für eine binäre Klassifikation berechnet sie sich aus der folgenden Formel [vgl. Géron 2020, 95]:

Für n-klassige Klassifikationsprobleme haben sich verschiedene Herangehensweisen herausgebildet, die Precision zu berechnen. Die in dieser Arbeit verwendete, ist die sogenannte *Macro-Precision*, welche sich analog zur berechnet [vgl. Grandini et al. 2020, 6f.]:

Da die Precision allein sehr einseitig ist, geht Sie in der Regel mit dem *Recall* einher. Für die Matrix in Abb. X beträgt auch die Precision 0,8.

### Recall / Sensitivität

Der *Recall* (dt. Sensitivität), auch als Trefferquote bezeichnet, ist der Anteil der tatsächlich positiven Datenpunkte, die vom Klassifikator als solche vorhergesagt wurden. Sie berechnet sich aus der folgenden Formel [vgl. Géron 2020, 95]:

Auch für den Recall haben sich im Hinblick auf n-klassige Klassifikationsprobleme verschiedene Herangehensweisen gebildet. In dieser Arbeit wird der *Macro-Recall* verwendet, der sich analog zur und berechnet [vgl. Grandini et al. 2020, 6f.]:

Auch der Recall ist für sich genommen sehr einseitig, deshalb wird er in der Regel gemeinsam mit der Precision ausgewertet. Für die Matrix in Abb. X beträgt der Recall lediglich ca. 0,55.

### F1-Score

Da Precision und Recall oft gemeinsam ausgewertet werden kann es bequem sein diese zu einer einzigen Metrik zusammenzufassen, dem sogenannten *F1-Score*. Dieser berechnet sich aus dem harmonischen Mittelwert von Precision und Recall (bzw. Macro-Precision und -Recall bei n-klassigen Klassifikationen) [vgl. Géron 2020, 96]:

Im Gegensatz zur Accuracy eignet sich der F1-Score auch zur Beurteilung von unausgeglichenen Datensätzen. Für die Matrix in Abb. X beträgt der F1-Score ca. 0,65.

## Eingesetzte Technologien und Frameworks

### Python

Kurze Anführung der verwendeten Python-Version und der genutzten Standardbibliotheken (insb. Pandas und numpy).

### ScikitLearn

### Keras

# Verwandte Arbeiten

## Vorangegangene Abschlussarbeiten

Dieser Arbeit gehen drei Abschlussarbeiten voraus, welche sich ebenfalls mit dem Themenfeld *Klassifizierung von Verkehrsteilnehmern basierend auf Positionsdaten* befasst haben. Nachfolgend werden die wichtigsten Erkenntnisse aber auch die Grenzen dieser Arbeiten aufgezeigt.

Die Bachelorarbeit von Torlak aus dem Jahr 2022 bildet im Wesentlichen den Grundstein für die späteren Arbeiten. Die Arbeit setzt sich mit verschiedenen Möglichkeiten der Datengewinnung auseinander, wobei der Fokus auf der Abfrage von Positionsdaten aus der Verkehrssimulation *CARLA* liegt, für die im Zuge der Arbeit auch eine Python-API umgesetzt wurde. Auch Möglichkeiten reale Daten zu gewinnen und simulierte Daten realistischer zu machen, werden in der Arbeit aufgeführt. Außerdem beschäftigte sich Torlak mit der Berechnung von Bewegungsinformationen aus den ermittelten Positionsdaten, welche später zur Klassifikation genutzt werden sollen. Die Klassifikation selbst ist jedoch kein Teil der Arbeit [siehe Torlak 2022].

Sohl knüpft in seiner Bachelorarbeit, ebenfalls aus dem Jahr 2022, direkt an die Arbeit von Torlak an. Er nutzt hierbei die aus der Verkehrssimulation CARLA gewonnenen Daten um drei verschiedene Klassifikationsverfahren zu evaluieren. Bei diesen handelt es sich um die Verfahren *Support-Vector-Machine*, *Decision-Tree* und *K-Nearest-Neighbours.* Hierbei sollten die Klassifikatoren auf Basis verschieden stark verrauschter Positionsdaten bis zu fünf Fahrzeugklassen zu unterscheiden lernen, allerdings stellte sich im Verlauf der Arbeit heraus, dass auf Basis der Simulation höchstens drei Klassen unterschieden werden können. Vor der Klassifikation wurden die über ein festes Zeitintervall gesammelten Positionsdaten in lange Eingabevektoren weiterverarbeitet, die die Bewegung der Verkehrsteilnehmer innerhalb dieses Intervalls repräsentieren. Eine Merkmalsreduktion fand nicht statt. Die besten Ergebnisse wurden dabei durch die SVM erzielt, welche eine Genauigkeit von 93,9% auf unverrauschten Daten erreichen konnte. Ein Verrauschen der Daten, um diese vermeintlich realistischer zu machen, führte in allen Fällen zu einem starken Abfall der Qualität der Klassifikation [siehe Sohl 2022].

Die Masterarbeit von Fischer …

## Vehicle Classification from Low-Frequency GPS Data with Recurrent Neural Networks

Bildet Grundlage für meinen angestrebten Ansatz (Sequenzielle Klassifikation mit RNNs). Datenbasis ist jedoch nicht vergleichbar (GPS-Punkte mit stark variierenden und großen zeitlichen Abständen).

[Matteo Simoncini et al. 2018]

## Vehicle Classification using GPS Data

Ähnlich zur Arbeit von Sohl. Liefert gute Erkenntnisse darüber, welche Merkmale zur Klassifikation von Verkehrsteilnehmern genutzt werden können. Auch hier ist die Datenbasis (bildet lediglich Fahrten von PKWs und LKWs auf Hauptverkehrsadern ab) und bei genauerem Blick auch die Aufgabenstellung (nicht-sequenzielle Klassifikation) eine völlig verschiedene zu meiner Arbeit.

[Sun et al. 2013]

# Datengrundlage

## Gewinnung der Ausgangsdaten

Wie wurden die Daten gewonnen? -> MotionTrace

Unter welchen Umständen und unter welchen Einschränkungen wurden die Daten erzeugt?

## Beschreibung der Ausgangsdaten

Wie umfangreich ist die Datenbasis?

Wie sehen die Rohdaten aus?

Sind die Daten ausgeglichen?

Probleme in den Daten, welche beachtet werden müssen?

## Weiterverarbeitung zu Bewegungsdaten

# Anforderungen und Gesamtkonzept der Klassifikation

Teil 1: Wie genau sieht das zugrundeliegende Problem aus und welche konkreten Aufgaben/Anforderungen ergeben sich daraus?

Teil 2: Wie könnte man dieser Aufgaben Herr werden? Welche Ansätze und Ideen werden gewählt?

# Umsetzung des Map-Matchings

## Einrichtung der Valhalla-Engine

## Umsetzung und Evaluierung des Road-Snappings

## Vorklassifikation zur Bestimmung des korrekten Matching-Modus

### Erzeugung des Trainingsdatensatzes

### Auswahl des Klassifikators

### Evaluierung

# Umsetzung der Klassifikation

### Erzeugung der Trainingsdatensätze (inkl. Vorverarbeitungen)

### Training und Optimierung der Modelle

### Vorstellung der besten Klassifikatoren

Weitere Unterkapitel ggf. abhängig von schlussendlich verwendeten Verfahren.

# Evaluierung und Diskussion

Weitere Unterkapitel abhängig von schlussendlich verwendeten Verfahren. Dieses Kapitel wird aber auf jeden Fall einen Vergleich zwischen gematchten und ungematchten Daten bei der finalen Klassifikation enthalten.

# Zusammenfassung und Ausblick

…

# Quellenverzeichnis

Fischer, D. 2023. *Verwendung von Positionsdaten zur automatisierten Klassifizierung von Verkehrsteilnehmern mittels maschinellen Lernverfahren.* Masterarbeit.

Géron, A. 2020. *Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow. Konzepte, Tools und Techniken für intelligente Systeme.* 2. Auflage. Heidelberg: O'Reilly.

Grandini, M. et al. 2020. *Metrics for Multi-Class Classification: an Overview* [online]. Verfügbar unter https://arxiv.org/pdf/2008.05756.

Matteo Simoncini et al. 2018. Vehicle Classification from Low-Frequency GPS Data with Recurrent Neural Networks [online]. *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2018.03.024.

Sohl, M. 2022. *Klassifizierung der Bewegungsmuster von Mobilfunkteilnehmern zur erweiterten Umfeldwahrnehmung autonomer Fahrzeuge.* Bachelorarbeit.

Sun, Z. und Ban, X. 2013. Vehicle classification using GPS data [online]. *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2013.09.015.

Torlak, R. 2022. *Detektion der Bewegung von Verkehrsteilnehmern aus Positionsdaten.* Bachelorarbeit.

1. Test
2. Test 2