UE 3

Merkmalsextraktion, binäre Klassifikation

Überblick

Das Ziel dieser Übung ist die Integration der in der Vorlesung kennengelernten Merkmalsextraktionsverfahren in eine Klassifikationspipeline. Dabei werden wir uns zunächst einmal der Unterscheidung von jeweils **zwei** Klassen widmen, also binären Klassifikationsproblemen. Wir werden den GTSRB Datensatz verwenden (***R**ecognition* statt ***D**etection*) um die einzelnen Verkehrsschildklassen zu unterscheiden. PCA und Lernkurve werden benutzt um das Problem zu veranschaulichen und genauer zu untersuchen.

Verglichen mit den Übungen 1 und 2 ist diese Übung anders gestaltet. Es gibt diesmal keine Jupyter-Notebook-Vorlage. Ihre Aufgabe besteht darin, selbstständig ein Jupyter Notebook anzulegen, die nötigen Pakete zu importieren und die Aufgaben basierend auf dem Wissen aus den vergangenen Übungen und den Hinweisen in diesem Übungsblatt zu lösen. Als Grundlage dient die abgebildete Pipeline.

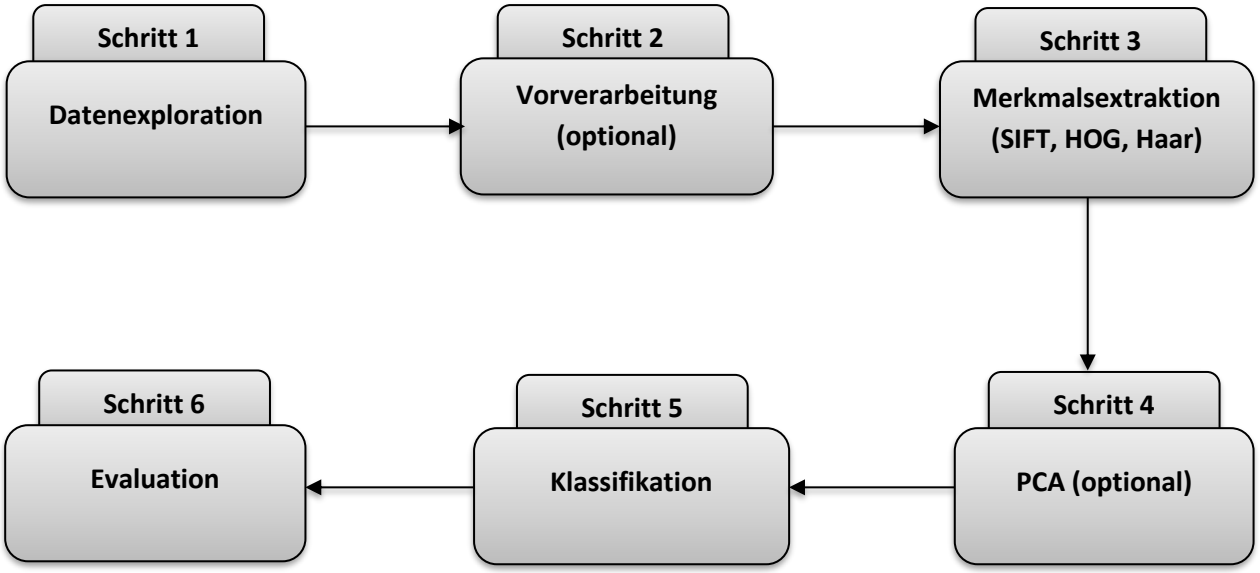

Abbildung 1: Schematische Darstellung der Klassifikationspipeline

Aufgabe 1 – Datenexploration

Die Daten (*official training data* und *official test dataset*) findet ihr unter folgendem Link. Hier sind sowohl die Datensätze mit den Bildern abgelegt als auch die Datensätze mit vorverarbeiteten Features (HOG, Haar-like, Hue Histograms). Ihr könnt sowohl die bereits berechneten Features-Datensätze nutzen als auch eure eigenen Features-Datensätze erstellen.

Eine detaillierte Beschreibung der Datensätze ist unter folgendem Link zu finden. Die README-Dateien in den jeweiligen Ordnern enthalten ebenfalls nützliche Informationen. Für unser binäres Klassifikationsproblem werden nur die Datensätze aus dem *official training data*-Block benötigt.

Daten in Trainings- und Validierungsdaten unterteilen:

1. Festlegen, welche Klassen eingelesen werden (sich zwei aussuchen). Dies könnt ihr im Laufe der Übung variieren, um schwierigere und leichtere Klassifikationsprobleme zu untersuchen, je nachdem ob die Schilder sehr ähnlich sind oder sich sehr unterscheiden.

2. Daten der entsprechenden Klassen einlesen. Nutzt dafür euer Vorwissen aus vergangenen Übungen. Da ihr nur die Datensätze aus dem *official training data*-Block verwendet, sollten die eingelesenen Daten zusätzlich in Trainings- und Validierungsdaten unterteilt werden. Ein mögliches Verhältnis wäre z.B. 75% Trainingsdaten und 25% Validierungsdaten. Dafür könnt ihr beispielsweise die train_test_split-Funktion aus der scikit-learn-Bibliothek nutzen. Bevor die Funktion genutzt werden kann, solltet ihr das scikit-learn-Paket in eurer Entwicklungsumgebung installieren (detaillierte Beschreibung zum Installieren von Paketen könnt ihr in der Anleitung zur Einrichtung der Entwicklungsumgebung oder hier nachschlagen).

Aufgabe 2 – Merkmalsextraktion

Zur Vorbereitung auf diese Aufgabe sollt ihr euch bereits im Vorfeld mit der Verwendung der dafür benötigten Funktionen vertraut machen.

- HOG-Features, SIFT
  - Wie erfolgt die Berechnung? Pixel, Auflösung **角点，**边界，平面 Eckepunkte, Kanten,
  - Wie viele Merkmale sind bei einer bestimmten Konfiguration im Merkmalsvektor enthalten?
  - OpenCV: SIFT (Theorie und Implementierung)
  - Scikit-Image: HOG (Beispiel einer Implementierung)

Als erstes könnt ihr euch die verlinkten Beispiele anschauen. In dem darauffolgenden Schritt könnt ihr die Verfahren mit einem beliebigen Verkehrszeichenbild nachimplementieren (siehe Beispiel in Abbildung 2).

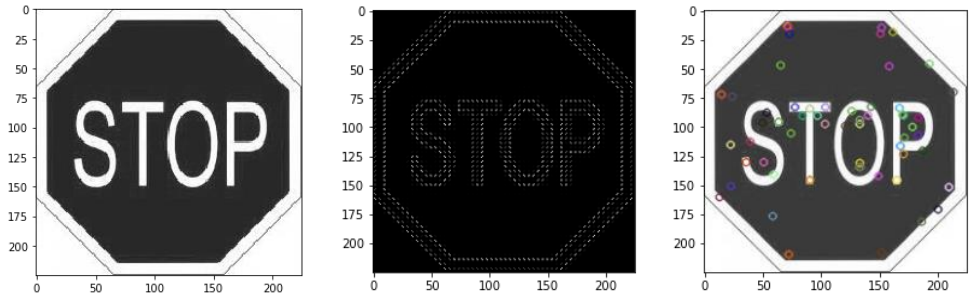

Abbildung 2: Graustufenbild (links), Visualisierung der HOG-Features (mittig), Graustufenbild mit SIFT-Keypoints (rechts)

Anschließend solltet ihr euer Code so anpassen, dass ihr Features aus allen Bildern den von euch ausgewählten Klassen extrahiert.

Aufgabe 3 – PCA

Beantworte zuerst folgende Fragen:

- Wie kann eine PCA in OpenCV oder in scikit-learn durchgeführt werden?
- In welchem Format werden die Daten dafür benötigt? Multidimensional data set X of dimension p
- Wie kann ich neue Daten (z.B. bei der Validierung) entsprechend projizieren?

Hinweise zur PCA: 对训练数据进行PCA，然后将数据投影到新的特征空间。这将告诉你，你的分类问题有多难，或者说你的特征能够多好地区分类别。
Führt eine PCA auf den Trainingsdaten aus und projiziert die Daten dann in den neuen Merkmalsraum. Das gibt euch Aufschluss darüber, wie schwierig euer Klassifikationsproblem ist beziehungsweise wie gut eure Features geeignet sind, um die Klassen zu unterscheiden.

Visualisiert euch die Features von unterschiedlichen Klassen von Verkehrsschildern. (Ihr könnt hier auch mehr als nur zwei Klassen auswählen, um euch mehr Klassen auf einmal zu visualisieren.) Ihr könnt sowohl die OpenCV- als auch die scikit-learn-Bibliothek dafür nutzen.

Aufgabe 4 – Klassifikation

Beantwortet zuerst folgende Fragen:

- Welche Klassifikatoren kennt ihr bereits?
- Wie kann ich die Klassifikatoren trainieren?
- In welchem Format werden die Daten benötigt?
- Wie mache ich Vorhersagen für neue Daten?
- Welche Fehlermaße bieten sich an, um den Klassifikator auszuwerten?
- Was ist die Lernkurve eines Klassifikators?
- Wie kann ich diese Lernkurve erstellen?
- Wie kann mir die Lernkurve weiterhelfen, um die Eignung meines Klassifikators einzuschätzen?

Hinweise zur Klassifikation:

Für die Klassifikation könnt ihr SVM-Klassifikator einsetzen.

OpenCV:

- OpenCV Tutorial: https://docs.opencv.org/3.4/d1/d73/tutorial_introduction_to_svm.html
- Informationen zur Verwendung des SVM-Klassifikators könnt ihr auch im OpenCV 3 computer vision with Python cookbook finden.

Scikit-learn: https://scikit-learn.org/stable/modules/svm.html

Aufgabe 5 – Evaluation

Hinweise zur Evaluation:

Zur Beurteilung der Klassifikationsleistung des Klassifikators könnt ihr die CCR auf den Validierungsdaten berechnen und euch die Konfusionsmatrix anschauen. Dafür könnt ihr beispielsweise die accuracy_score-Funktion, die confusion_matrix-Funktion, f1_score-Funktion der scikit-learn-Bibliothek nutzen. Experimentiert ein wenig und vergleicht die Ergebnisse für verschiedene Klassen und Konfigurationen eures Klassifikators.

Erstellung der Lernkurve: Das Ziel ist, zu schauen, wie sich die Fehler auf den Trainings- und Validierungsdaten mit der Menge der genutzten Trainingsdaten verändern. Dies wird Aufschluss darüber geben, wie ihr eure Klassifikationsleistung am effektivsten verbessern könnt. Man kann daraus ablesen, ob z.B. die Features ungeeignet sind oder das Modell des Klassifikators zu simpel ist, oder ob z.B. mehr Trainingsdaten helfen würden um die Klassifikationsleistung zu verbessern. Zur Erstellung der Lernkurve könnt ihr die learning_curve-Funktion der scikit-learn-Bibliothek nutzen.

02.06.2022: Da noch nicht alle Themen in der Vorlesung behandelt wurden, werden wir in der heutigen Übung mit den ersten 4 Schritten der Klassifikationspipeline beschäftigen.

其目的是看训练和验证数据的误差如何随着使用的训练数据量而变化。这将使人们深入了解如何最有效地提高你的分类性能。你可以看到，例如，特征是否不合适或分类器模型是否过于简单，或者，例如，更多的训练数据是否有助于提高分类性能。为了创建学习曲线，你可以使用scikit-learn库的learning_curve函数。