COMP09012 - Machine Learning - Group Project

Mark Breen, Phillip Garrad, Gerard Comerford

Institute of Technology, Sligo

Abstract

Report for the COMP09012 group project

Keywords: Imaging, Image Processing, Machine Vision, etc. (Maximum five)

1 Introduction

The goal of this paper is to predict the value of multiple classes relating to the road surface, traffic congestion and driving style of a vehicle.

2 Methods

This section will highlight the methods used for data cleaning and preparation, along with a description of the machine learning models and metrics we used.

2.1 Data

Data was received in the form of four text file of comma separated value (CSV) extension. The delimiter encoded in the four CSV files is a semi-colon (;). The data contains information recorded for two trips in two vehicles, the vehicles being an Opel Corsa 1.3 HDi (95 CV) and Peugeot 207 1.4 HDi (70 CV). Since there is two recorded trips per vehicle, this means there is a total of four CSV data files. The data was recorded from the vehicle on-board diagnostics (OBD) and micro-devices embedded in the smartphone of the driver operating the vehicle at the time.

A breakdown of the number of rows in each vehicle dataset can be found in Table 1

Table 1: Number of rows in each of the raw data files

Filename	Vehicle	Number of rows	Number of columns
opel_corsa_01.csv	Opel Corsa 1.3 HDi (95 CV)	7392	17
opel_corsa_02.csv	Opel Corsa 1.3 HDi (95 CV)	4328	17
peugeot_207_01.csv	Peugeot 207 1.4 HDi (70 CV)	8614	17
peugeot_207_02.csv	Peugeot 207 1.4 HDi (70 CV)	4623	17

A description of the seventeen different columns available in the dataset is shown in Table 3

Due to large variations in the value and units of measurement across the numerical columns it was decided to scale the numerical columns to be within the range $(0, 1)$. To do this equation was applied column-wise for each column j in the numerical subset of the data columns:

$$X_{j\ \text{scaled}} = \frac{X_j - \min(X_j)}{\max(X_j) - \min(X_j)} \tag{1}$$

Table 2: Data column descriptions

Column Name	Column Description	Data Type
AltitudeVariation	The altitude change calculated over 10 seconds	Decimal
VehicleSpeedInstantaneous	The current speed of the vehicle	Decimal
VehicleSpeedAverage	Vehicle average speed in the last 60 seconds	Decimal
VehicleSpeedVariance	Speed variance in the last 60 seconds	Decimal
VehicleSpeedVariation	Speed variation for every second of detection	Decimal
LongitudinalAcceleration	Longitudinal acceleration	Decimal
EngineLoad	Engine load as a percentage	Decimal
EngineCoolantTemperature	The engine coolant temperature in degrees celsius	Decimal
ManifoldAbsolutePressure	Manifold air pressure	Decimal
EngineRPM	Revolutions Per Minute (RPM) of the engine	Decimal
MassAirFlow	Mass Air Flow measured in grams per second	Decimal
IntakeAirTemperature	Intake air temperature at the engine entrance	Decimal
VerticalAcceleration	Vertical acceleration	Decimal
FuelConsumptionAverage	Average fuel consumption in litres per 100 km	Decimal
roadSurface	Road surface condition	String
traffic	Traffic congestion condition	String
drivingStyle	Driving style	String

Table 3: Percentage values missing in each column as a proportion of the number of rows of each dataset

Column Name	opel_corsa_01.csv	opel_corsa_2.csv	peugeot_207_1.csv	peugeot_207_2.csv
AltitudeVariation	0.73	0.83		
VehicleSpeedInstantaneous	0.09	0.18		
VehicleSpeedAverage	4.8	5.45		
VehicleSpeedVariance	4.8	5.45		
VehicleSpeedVariation	0.46	0.88		
LongitudinalAcceleration	0	0		
EngineLoad	0	0		
EngineCoolantTemperature	0	0		
ManifoldAbsolutePressure	0	0		
EngineRPM	0	0		
MassAirFlow	0	0		
IntakeAirTemperature	0	0		
VerticalAcceleration	0	0		
FuelConsumptionAverage	1.31	1.2		
roadSurface	0	0		
traffic	0	0		
drivingStyle	0	0		

2.2 Model Metrics

The F_1 score was used to assess the accuracy of the model. The F_1 score is defined as the harmonic mean of the precision and recall. Precision is defined as:

$$\text{Precision} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}} \tag{2}$$

Recall is defined as:

$$\text{Precision} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}} \tag{3}$$

Then, the F_1 score is defined as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

Acknowledgments

The Acknowledgments section, if included, follows the main body of the text and is headed “Acknowledgments,” printed in the same style as a section heading, but without a number.

References