COMP09012 - Machine Learning - Group Project

Mark Breen, Phillip Garrad, Gerard Comerford

Institute of Technology, Sligo

Abstract

There are hundreds of inputs which are important to consider when making decisions in autonomous vehicles. The ability to quantify these factors is of the upmost importance since quantitative data is ultimately what will need to be fed into any decision making algorithm. Three of these factors will be considered in this paper. Three such factors: driving style, road surface condition and traffic are investigated in this paper.

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:

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%	0.73%	0.58%
VehicleSpeedInstantaneous	0.09%	0.18%	0.10%	0.43%
VehicleSpeedAverage	4.79%	5.45%	4.82%	3.83%
VehicleSpeedVariance	4.79%	5.45%	4.82%	3.83%
VehicleSpeedVariation	0.46%	0.88%	0.91%	0.82%
LongitudinalAcceleration	0.00%	0.00%	0.00%	0.00%
EngineLoad	0.00%	0.00%	0.06%	0.00%
EngineCoolantTemperature	0.00%	0.00%	0.06%	0.00%
ManifoldAbsolutePressure	0.00%	0.00%	0.06%	0.00%
EngineRPM	0.00%	0.00%	0.06%	0.00%
MassAirFlow	0.00%	0.00%	0.06%	0.00%
IntakeAirTemperature	0.00%	0.00%	0.06%	0.00%
VerticalAcceleration	0.00%	0.00%	0.00%	0.00%
FuelConsumptionAverage	1.31%	1.20%	1.11%	0.89%
roadSurface	0.00%	0.00%	0.00%	0.00%
traffic	0.00%	0.00%	0.00%	0.00%
drivingStyle	0.00%	0.00%	0.00%	0.00%

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

2.2 Classification Models

2.2.1 Random Forest Classifier

A random forest classifier works on the basis of decision trees and is essentially an ensemblement of decision tree classifiers. A random forest algorithm works drawing bootstrapped samples from the original dataset and fitting a decision tree using a random sample of the predictors. In order to predict using new data, an aggregate prediction is calculated from n_{tree}, the number of trees. A majority concensus is what's used to determine the predicted classification label.

2.2.2 K Neighbours Classifier

The K-Nearest Neighbours algorithm differs from the Random Forest Classifier in that instead of inferring a model from the data, the data is used directly. The K-Nearest Neighbours algorithm essentially works as follows (for a hyper-parameter k):

- Retrieve the k most observations
- Take the mode of k votes to determine which class the observation belongs to

2.3 Model Metrics

The F_1 score was used to assess the accuracy of the model. The reason why the F_1 score has been chosen as a model metric here is due to imbalanced classes being observed in the dependent variables, meaning that the number of false positives and false negatives is crucial to consider when comparing models. 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:

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

Since the F_1 score is calculated on a per-class basis and there are multiple classes in two of the dependent variables, an average F_1 score was calculated to in order to assess model accuracy.

3 Results

3.1 Random Forest Classifier

3.2 K Neighbours Classifier

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