Machine Learning Assignment 1 – Group Classification Algorithms and Cleaning Data

Group A - Ben Fitzgerald (s00244687)

15/11/2023

Abstract

This project undertakes the analysis of classification algorithms using a dataset derived from vehicle sensors. The dataset includes features for classifying three distinct labels: Driving Style, Road Surface Condition, and Traffic Congestion. A data cleaning process is performed, addressing formatting inconsistencies and missing values while also looking at imbalanced datasets, feature scaling, changing classes to numeric values, and determining what features are useful. The study focuses on the classification of Road Surface Conditions. Two classification models, Logistic Regression and Support Vector Machine (SVM) are trained and evaluated. This research provides insights into the performance of classification algorithms on real-world sensor data, specifically in the context of road surface classification.

1 Introduction

This project explores the application of classification algorithms to data obtained from vehicle sensors, with the goal of classifying Road Surface Conditions into three different classifications: Smooth Condition, Uneven Condition, or Full of Holes Condition. The dataset includes features from different vehicles and journeys, introducing challenges such as data inconsistencies and missing values. The first step is to make sure that the data is cleaned effectively, as this will help us provide the most accurate model.

2 Data Cleaning

Data cleaning is a fundamental process in this research, aimed at preparing the dataset for analysis and modeling. The dataset, sourced from vehicle sensors, faced real-world data challenges that required careful handling. The data cleaning steps included data integration, handling missing values, addressing class imbalance, label encoding, and feature scaling.

Data integration involved consolidating four separate datasets representing different vehicles and journeys into a single dataframe to ensure a comprehensive approach to data analysis. Once this was done, there were 23,775 rows of data but some features had missing values. Since there were not many rows that contained missing values, it was easier to just exclude these data points from the dataset, which left the dataset with 23,762 rows of data, so there was little to no impact in removing these missing values.

To make sure the data was consistent across the two vehicles and between the two runs of the same vehicle, an analysis was done among the four original dataframes comparing all 14 features using boxplots to see if any particular vehicle or run of the data contained any outliers. The 2nd run of the Opel Corsa seems to provide some outliers in the Vertical Acceleration and Longitudinal Acceleration columns.
To focus on only one class column, Driving Style and Traffic were removed from the dataset and the only focus was identifying the Road Surface Condition and which class it fell into between Smooth Condition, Uneven Condition, and Full of Holes Condition.

Examining these 3 classes further to see if this was an imbalanced dataset, this was the result;

- Smooth Condition: 14,235
- Uneven Condition: 6,280
- Full of Holes Condition: 3,247

This dataset is imbalanced, so the best approach is to undersample the data which requires removing rows at random from the majority classes and leaving all 3 classes with the same number of samples in each, 3,247. The main advantage of undersampling is that artificial observations are not added to the dataset. This is beneficial because duplicating existing observations might make it seem like patterns are more widespread than they are in reality. This can lead to overfitting to specific patterns. [3]

The dataset is now balanced but the classes must be changed to a numeric value to fit the model. This was done by encoding the Road Surface Condition with different values:

- Smooth Condition: 2
- Uneven Condition: 1
- Full of Holes Condition: 0

The features in the dataframe are all using different scales, for example, the EngineRPM has a mean value of 1,367.5 whereas VehicleAcceleration has a mean value of 0.1. The best way to combat this was to use feature scaling using StandardScalar, which gave every feature a mean of 0 and a standard deviation of 1. This process is essential for machine learning algorithms, as it prevents features with larger numerical values from dominating those with smaller values and helps make fair comparisons between different features.

3 Methodology

The research methodology primarily focuses on the model selection, training, and evaluation for classifying 'Road Surface Condition' accurately. The dataset, derived from vehicle sensors, has undergone thorough data cleaning to ensure data quality and reliability.
Two classification algorithms, Logistic Regression and Support Vector Machine (SVM) were selected for analysis. The dataset was split into training and testing sets to facilitate model training and evaluation. The train-test data split ratio was set to 80:20. This split allows the models to learn from most of the data and then be tested to assess how predictive each model is and to compare the results of each model. The two models chosen are detailed below;

- Logistic Regression:

  Logistic Regression is a widely used classification algorithm that is particularly suited for binary and multiclass classification problems. It models the probability that a given input belongs to a particular class. The logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables [1]

- Support Vector Machines (SVM):

  Support Vector Machine is another powerful classification algorithm that is known for its ability to handle both linear and non-linear classification tasks. The objective of the support vector machine algorithm is to find the hyperplane in an N-dimensional space that distinctly classifies the data points. [2]

With a cleaned and prepared dataset, the project transitioned to model selection, training, and evaluation, with a focus on the accuracy of classifying 'Road Surface Condition.' The detailed findings and insights from the experimental results are discussed in the following section.

4 Results

The experimental results reveal the performance of the Logistic Regression and SVM models in classifying Road Surface Conditions.

Logistic Regression Model:

- Accuracy: 0.7573
- The Logistic Regression model achieved an accuracy of approximately 75.73%. The classification report for the Logistic Regression model provides detailed metrics for each class:
  - Full of Holes Condition: Precision - 0.75, Recall - 0.81, F1-Score - 0.78
  - Uneven Condition: Precision - 0.72, Recall - 0.68, F1-Score - 0.70
  - Smooth Condition: Precision - 0.80, Recall - 0.80, F1-Score - 0.80
- The weighted average F1-Score for this model is 0.76.

SVM Model:

- Accuracy: 0.8943
- The SVM model achieved a higher accuracy of approximately 89.43%. The classification report for the SVM model provides detailed metrics for each class:
  - Full of Holes Condition: Precision - 0.84, Recall - 0.96, F1-Score - 0.90
  - Uneven Condition: Precision - 0.89, Recall - 0.83, F1-Score - 0.86
  - Smooth Condition: Precision - 0.96, Recall - 0.90, F1-Score - 0.93
- The weighted average F1-Score for this model is 0.89.

5 Discussion

- Precision is a measure of how many of the positive predictions made are correct.
- Recall is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.
- F1-Score is a measure of combining both precision and recall. The F1-Score is not simply the arithmetic mean but it is more sensitive to one of precision or recall being quite low so both have to be high to get a good F1-Score. [4]

The results offer significant insights into the performance of the Logistic Regression and SVM models:

- The Logistic Regression model exhibits moderate accuracy, with an accuracy score of approximately 75.73%. While it achieves balanced performance across classes, its precision, recall, and F1-scores are slightly lower compared to the SVM model.
- In contrast, the SVM model outperforms Logistic Regression with an accuracy of around 89.43%. It demonstrates notably higher precision, recall, and F1-scores for all classes. The SVM model excels in capturing the nuances of Road Surface Condition classification, particularly in the challenging task of distinguishing between different conditions.
- The support vectors and dual coefficients of the SVM model reveal its robustness in accurately classifying road surface conditions. These support vectors play a crucial role in defining the decision boundary of the model.

The choice between Logistic Regression and SVM should consider the specific requirements of the application. The Logistic Regression model chosen performs reasonably well but the SVM model outperforms it by being more accurate and reliable. Where something such as road safety has little to no room for error, picking the correct model in instances like this is crucial, and the F1-score being significantly better in the SVM model shows it is the obvious choice in this scenario.

The 14 features used in the model also play a critical role in the results. The features that were most positively correlated with being classed in the Smooth Condition were VehicleSpeedAverage, EngineCoolantTemp, ManifoldAbsolutePressure, EngineRPM, and FuelConsumptionAverage. This makes intuitive sense as you would expect a car to be faster on a smooth road compared to one full of holes. Unsurprisingly, these features were negatively correlated with being classed as Full of Holes Condition. For the Uneven Condition, the features of the model sat in between Smooth Condition and Full of Holes Condition and none of them were too large in absolute value.

6 Conclusion

In conclusion, this project focused on the application of classification algorithms to real-world sensor data derived from vehicle sensors, specifically aiming to classify road surface conditions. A meticulous data cleaning process was conducted, addressing issues such as missing values, class imbalances, label encoding, and feature scaling. The analysis concentrated on two classification models: Logistic Regression and Support Vector Machine (SVM).

The results demonstrated that the SVM model outperformed the Logistic Regression model in accurately classifying road surface conditions. With an accuracy of approximately 89.43%, the SVM model exhibited higher precision, recall, and F1-scores for all classes. The F1-score, a balance between precision and recall, provided a comprehensive measure of model performance.

In practical applications, such as road safety, the choice of classification model is crucial. The SVM model, with its higher accuracy and robust performance, emerged as the preferred choice for this specific scenario. The findings emphasize the importance of selecting appropriate algorithms based on the specific requirements of the application.

References

[1] Wikipedia contributors. "Logistic regression." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 2 Nov. 2023. https://en.wikipedia.org/wiki/Logistic_regression

[2] Rohith Gandhi. “Support Vector Machine – Introduction to Machine Learning Algorithms”. Medium. 7 Jun. 2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[3] Christina Ellis. “Oversampling vs undersampling for machine learning”. Crunching the Data. Jan 21. 2023. https://crunchingthedata.com/oversampling-vs-undersampling/

[4] Teemu Kanstrén. “A Look at Precision, Recall, and F1-Score”. Medium. 11 Sep. 2020. https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec