Gesture UI Project 1 - 2024

Ronan Noonan, G00384824

Atlantic Technological University

Abstract

To predict driving styles from vehicle sensor data, this study assesses the performance of three classification algorithms: Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN). Extensive data preprocessing was conducted on datasets from Opel Corsa and Peugeot 207 cars, which included feature scaling, categorical encoding, and imputation for missing values. According to results, SVM is the most effective model for correctly detecting driving styles. This highlights the importance of careful model selection and data preparation for machine learning projects.

Introduction

The ability to understand driving habits through sensor data has massive implications for improving road safety and customizing vehicle designs to suit various driving ideas. In order to differentiate between "Even Pace" and "Aggressive" driving styles, this study uses machine learning, utilizing 14 different variables that are captured by vehicle sensors. Real-world data obtained from a Kaggle dataset need complex feature engineering and data cleaning in order to render the information suitable for algorithmic analysis.

Because of their unique advantages in modeling complex data interactions, Support Vector Machines (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) are being asked to be used. SVM is valued for its ability to manage nonlinear and high-dimensional data, making it suitable for complex sensor patterns associated with driving philosophies (Burges, 1998). A clear framework for feature impact analysis is provided by logistic regression, which is perfect for binary classifications like determining driving styles (Hosmer Jr., Lemeshow, Sturdivant, 2013). KNN, on the other hand, is well suited for this complex task since it makes use of data point similarities to capture driving behaviors through sensor data proximity (Cover and Hart, 1967). When combined, these classifiers offer a thorough analysis of the dataset, ranging from linear to nonlinear modeling, improving our comprehension of the predictive power of each model.

2. Methodology

This project follows a detailed and organized method that includes preparing the data, labeling, scaling, analyzing it thoroughly, and using visualization. This process helps in effectively training and evaluating the models.

2.1 Data Pre-Processing

The dataset has standard problems upon first inspection, including missing values, a variety of data types, and features with varying scales. Because mean imputation is simple and effective, it was selected to handle the missing values with the expectation that the absences would not significantly alter the results. This method works very well with datasets that have randomly missing data, meaning there are no predictable gaps (ProjectPro, 2023).

For the features "roadSurface" and "traffic," which are categorical, one-hot encoding was used. This method turns categories into a format algorithm can easily use, without wrongly assuming some categories are more important than others (Scikit-Learn., n.d.).

2.2 Data Labelling

The goal variable "drivingStyle" was encoded into numeric values, which made it easier to use different classification methods. This phase was vital because it converted subjective assessments of driving styles into quantitative results that could be predicted by models, enabling a direct comparison of algorithmic classification efficacy in driving behaviors.

2.3 Data Scaling

To normalize the range of independent variables, or features, in the data, feature scaling was used. Standardization (scaling to a mean of 0 and a standard deviation of 1) was used in light of the variability of sensor data scales to ensure that all characteristics contributed equally to the model's performance and to remove the disproportionate effect of features with wider ranges. This was particularly important for distance-based algorithms like K-Nearest Neighbors (KNN), which could be biased by features on larger scales.

2.4 Data Analysis and Visualisation

The distribution of the data and the relationships between the various pieces of information were visualized using histograms and correlation heatmaps. Selecting the best features for the model was simple with the use of these tools, which helped see which aspects were overly similar or useless. Additionally, they brought up the problem of inconsistent data across different driving styles, which provoked the use of stratified sampling to guarantee model training fairness and prevent bias toward more typical results.

Challenges and Solutions in Model Training

Model training faced difficulties with hyperparameter adjustment as well as overfitting and underfitting. In order to prevent overfitting, cross-validation was used to evaluate the model using unseen data. The approach for underfitting was to explore complex models and improve data features. GridSearchCV was crucial in helping to balance model complexity and accuracy while optimizing hyperparameters for SVM and KNN models. This method emphasizes the careful examination and adaptability required to successfully use machine learning for complicated datasets.

3. Experiments and Results

This section we will see the performance of the three classification algorithms on vehicle sensor data to predict driving styles.

3.1 Classifier Models

- **Logistic Regression**: This model showed potential in identifying linear relationships within the dataset. The logistic regression model got an accuracy of 84% on the test data. Its precision, recall, and F1-score for the "Aggressive" driving style were 85%, 98%, and 91% , indicating a high ability to correctly identify aggressive driving but with limitations in detecting even-paced driving due to the complexity of the dataset.
- **SVM:** Excelling in handling the complex patterns present within the driving style dataset, the SVM model, configured with parameters {'C': 10, 'gamma': 1, 'kernel': 'rbf'}, achieved a standout accuracy of 92%. Its performance was highlighted by precision and recall values of 93% and 98% for the "Aggressive" driving style, showcasing its accuracy in classifying driving behaviors.
- KNN: Configured with 'n_neighbors' set to 3, KNN utilized local data patterns to predict driving styles. The model achieved an accuracy of 91% on the test data, with precision and recall for the "Aggressive" driving style at 94% and 96%, respectively. This underscores KNN's capability in capturing driving behaviors through sensor data proximity, despite some challenges related to feature scaling and the data's high dimensionality.

3.2 Results

- The SVM model was notably effective, demonstrating dominant performance in distinguishing between driving styles, which was clearly shown by its accuracy and the consistency of its predictions across different data splits. The model's high mean cross-validation score of approximately 0.897 and a lower standard deviation compared to other models underline its efficiency and reliability.
- While the Logistic Regression and KNN models provided valuable insights into the data, their performance metrics underscore the importance of selecting appropriate models based on the specific characteristics and complexities of the dataset to achieve optimal outcomes.

Limitations and Potential Biases:

- The varied performances of the models across the dataset's complexity and feature scaling emphasize the essentials for model and hyperparameter selection. This variability suggests that different models may excel under specific conditions, reinforcing the need for a customized approach to model choice based on the dataset characteristics.
- The potential for data imbalance raises concerns about biases and calls for future investigations into ensemble approaches or advanced deep learning techniques in order to develop more reliable and objective predictive models.

4. Conclusion

The comparative analysis of driving style prediction highlighted the unique benefits and drawbacks of SVM, KNN, and Logistic Regression. The SVM model stood out for its exceptional recall and precision, highlighting the vital importance of careful data preprocessing and thoughtful algorithm selection in raising model accuracy. This investigation revealed how crucial feature selection and data quality are to the effectiveness of predictive models, particularly in fine-grained applications like driving behavior research.

4.1 Learning Outcomes and Future Research Directions:

- The importance of choosing correct machine learning models based on the specific characteristics of the dataset and the prediction task at hand.

- The significant impact of data preprocessing techniques in improving model performance.
- The utility of visual tools like histograms and heatmaps in identifying relevant features and assessing data quality.

As a result, this work has improved the ability to anticipate driving styles from sensor data while also highlights the ongoing effort to improve computational models in order to better comprehend and predict human behavior.

References

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. –

https://www.di.ens.fr/~mallat/papiers/svmtutorial.pdf

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons. - https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Applied+logistic+regression+%283rd+ed.%29.+John+Wiley+%26+Sons.&btnG=

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. –

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1053964

ProjectPro. (2023). How to Impute Missing Values with Mean in Python. - https://www.projectpro.io/recipes/impute-missing-values-with-means-in-python

Scikit-Learn. (n.d.). Preprocessing data. Retrieved from -

https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features