Driver Identification by Driving Style

Zhan Fan Quek, Eldwin Ng
zfquek, eldwin at stanford dot edu

Abstract

From optimizing a car's performance, encouraging safer driving to accurate insurance pricing, it is desirable to be able to identify a driver by his driving style. Using a car data supplied by MetroMile Inc., driving styles were analyzed and core features were extracted. Two learning models were investigated and evaluated:support vector machine (SVM) and multinomial logistic regression. High accuracies of around 90% were achieved for identifying a particular driver from a group of up to 6 drivers.

I. Introduction

In the world today, there are over one billion cars with different drivers interacting with each other on the roads. Each driver has their own driving style, which could impact safety, fuel economy, and road congestion, among many things. The precise relationships between driving style and their effects have not been well characterized, although there is some general consensus that "aggressive" driving (e.g. speeding, hard braking, tailgating) has a mostly negative impact (except perhaps on driver enjoyment). For example, research done by Honda in 2012 has shown that hard braking can increase road congestion and decrease fuel efficiency [1]. Knowing the driving style of the driver could be used to encourage "better" driving styles: combined with driver records or ideal driving style models, one's driving style could be compared and used as an immediate in-dash feedback while driving, or by scaling auto insurance rates commensurate to the aggressiveness of one's driving style, in rental cars for example – such discounts for good driving habits could help promote safer driving on the roads. The identification of driving style could also be used in conjunction with known classifications, such as "safe driver", "aggressive driver", or "good fuel economy driver" to optimize the performance of the vehicle based on the driver's style, for example, adjusting the transmission and other powertrain parameters to conform to the driver's preferred style.

The goal of our project is hence to analyze the driving style and to accurately identify a driver from a group of drivers based on the driving style. The dataset provided by MetroMile consists of the speed, heading, and instantaneous gas mileage of 18 vehicles as recorded using a driving sensor plugged into the car. The hypothesis is that each driver will have a particular driving style, and though this could depend on vehicle, route, road conditions, driver's mood, weather, etc., we expect that there will be some common elements which could be used to identify the driver. From the dataset provided, we can extract out features and a model using the driving data from one vehicle (driver) against the data from a few other drivers, and use it to identify the driver, evaluating the accuracy of the data features and the learning methods.

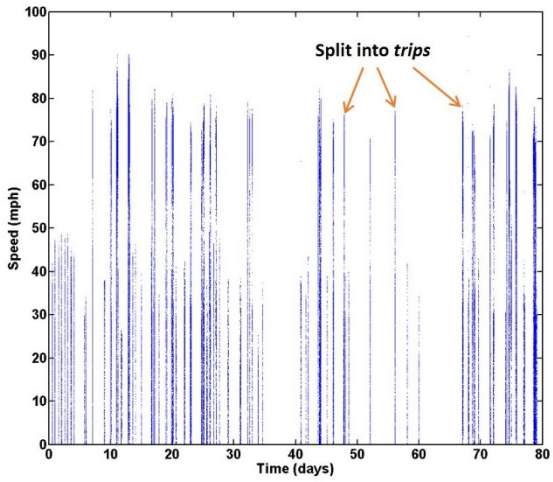

Figure 1. Raw data indicates that the vehicle was driven at intervals, and these were split into separate trips by time. Each trip forms an example, with an associated feature vector.

II. Analyzing MetroMile Data

The dataset provided by MetroMile Inc. recorded speed, heading, and instantaneous gas mileage for 18 vehicles. Unfortunately, not every vehicle's data had sufficient time resolution, or even data points, making it difficult to obtain accurate data features. A dataset from 6 vehicles was ultimately used in our analysis.

a. Raw data

Fig. 1 plots the raw speed data from MetroMile Inc., and it is observed that the data occur in intervals, each corresponding to a trip. As such, we broke up the raw data into different trips, and analyzed each individual trip data. Each trip was then taken to be a single example data point with an associated feature vector.

b. Features

Some features that were hypothesized to be indicative of driving behavior are the acceleration / deceleration

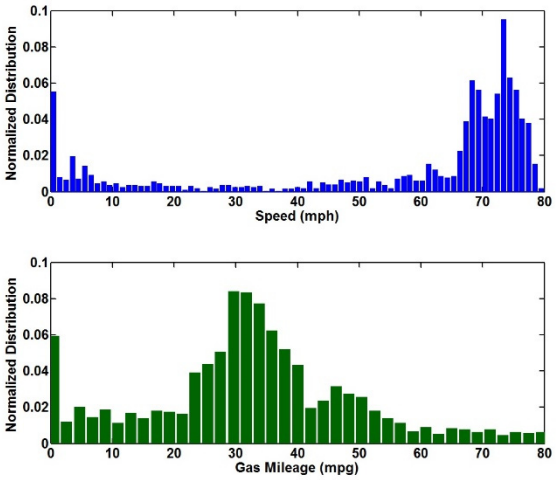

Figure 2. A normalized distribution of speed (top) and instantaneous gas mileage (bottom) for a trip (example).

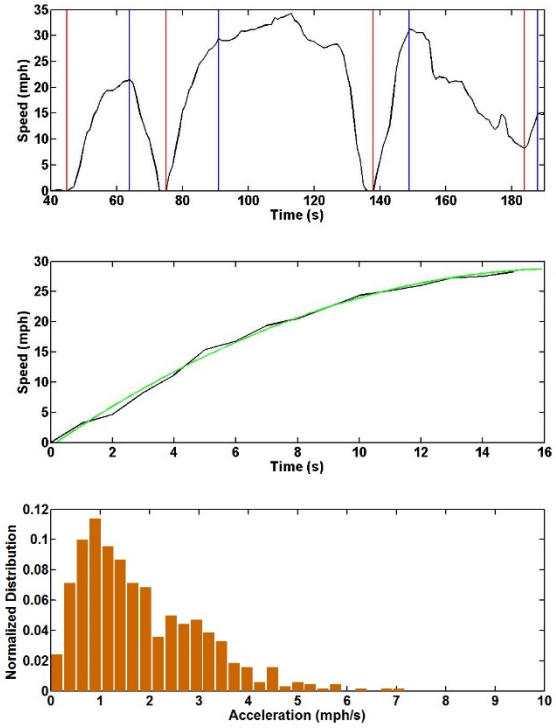

Figure 3. (a) Regions with significant acceleration were distinguished from the speed profile. Red: start of acceleration region; Blue: end of acceleration region. (b) The speed profile was fitted (green) with a 3rd order polynomial to reduce the noise. Acceleration data was then derived from the fitted speed curve. (c) Distribution of acceleration points (taken at each second from the fitted curve) for the acceleration regions of an entire trip.

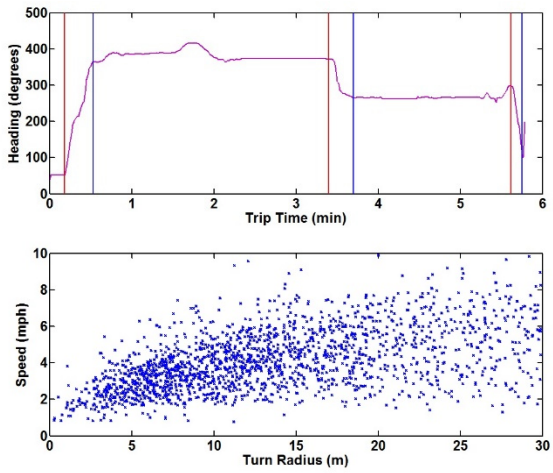

Figure 4. (a) Vehicle heading, as a function of trip time. Regions where the heading vary significantly are indicative of turning. (b) The speed vs. turning radius are plotted for the entire dataset of a driver.

profile, turning speed vs. radius, speed, and gas mileage. While some of these features have a dependence on external factors such as road conditions and the type of vehicle used, we assumed here in our work that these features capture more information with regards to the driving style of individual drivers rather than on these external factors. Such assumptions are necessary given the absence of data that was taken with a controlled route and vehicle.

i. Speed and Gas Mileage

The speed and gas mileage data points for each trip were binned into 80 and 40 groups respectively, and the normalized distribution is then used as a feature vector for the trip. The normalized distribution for an example trip is shown in Fig. 2. This distribution can be interpreted as a probability distribution in which a driver drives with a particular speed/gas mileage for a single trip.

ii. Acceleration / Deceleration Profile

To obtain the acceleration / deceleration profiles for each trip, we zoomed in on the regions where significant acceleration / deceleration was observed (- acceleration regions are shown in Fig. 3a). Because of the noisy data points, we fitted a 3rd order polynomial curve to the speed data at these regions (Fig 3b), and through differentiation of the fitted polynomial curve, obtain the curve for the acceleration. Using the fitted curve, the acceleration data is then resampled and binned into 40 groups (Fig 3c). The normalized distribution was then used as a feature vector for the machine learning algorithm. This normalized distribution can again be interpreted as a probability distribution in which a driver drive with a particular acceleration for a single trip.

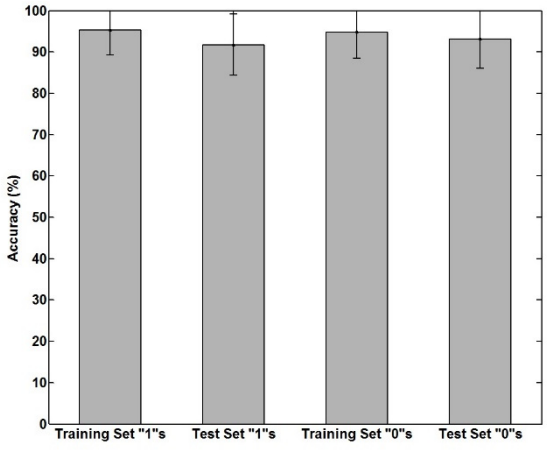

Figure 5. Accuracy of the SVM model for differentiating between two drivers. The driver to be identified has the label "1", while the other driver has the label "0". Error bars represent the different combinations of picking two drivers from a pool of six, and the random picking of examples to create a dataset with an equal number of positives / negatives.

iii. Turning speed vs. Radius of turn

Zooming in on significant changes in heading, and together with speed data, we extracted the turning radius, and Fig. 4 shows the speed as a function of the turning radius for a driver.

III. Modeling

a. Preliminaries

To assemble the labelled examples, the dataset for the driver to be identified was labelled positive ("1"), while that for the "Others" was labelled negative ("0"). One issue was that the datasets for positive "1"s and negatives "0"s were very different in size, because: 1) some driver data had a large number of trips and fine data points, while others had sparser data; 2) the dataset that is used for the "Others" category could consist of multiple drivers, leading to a larger set. This caused the decision boundary to shift in favor of the larger set, due to the objective to minimize the error, which was weighted heavily towards the larger set because of the difference in size. To reduce this problem, the sets for "1"s and "0"s were made to be the same size (size of the smaller set), by randomly picking examples from the original dataset to form a reduced dataset. Since there were multiple permutations of picking examples from the larger set, the simulation was re-run a series of times using different picked examples. A variation in the accuracy values is thus seen, and is captured by the error bars in the presented results.

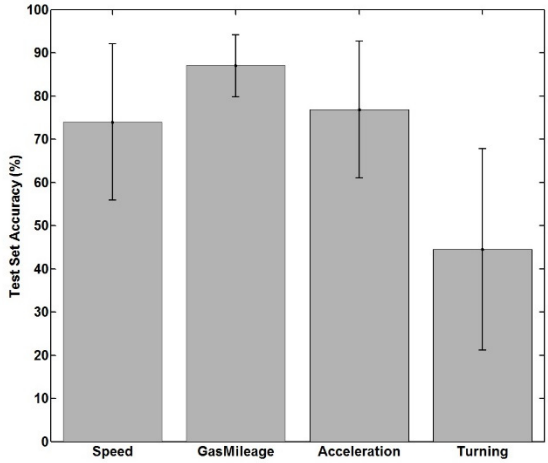

Figure 6. Accuracy of individual features of the SVM model for identifying between two drivers.

The labelled data was then divided up into a training and a test set, which consisted of 70% and 30% of the available data respectively, and support vector machine (SVM) and multinomial logistic regression models were trained and evaluated.

b. SVM

Implementing a linear SVM for classification, and a trip feature vector made up of acceleration/deceleration profiles, speed, and gas mileage, the results are shown in Fig. 5 for distinguishing between two drivers. Since data from six drivers were available, all possible combinations of two drivers were taken and this is represented by the error bars (in addition to the random picking of examples to create a dataset with an equal number of positives and negatives). Accuracy values are given for the model run against the training set itself and the test set, for both the driver to be identified "1", and the "Others" ("0") category. These accuracy values represent the fraction of the trips for which the driver was correctly identified using the model. High accuracy (>90%) is obtained.

While the above result is for the combined feature vector, one thing that we can look at is the accuracy of the individual features. Fig. 6 plots the accuracy of the individual features for identifying the driver of the test set. It is seen that the gas mileage feature gives the highest accuracy, followed by acceleration/deceleration, and speed features. The turning speed vs. turning radius feature is noticeably low in accuracy (~45%), doing worse than a random choice. This is likely to be due to a lack in data points, as most trips have under 10 turns – more data points per trip are required to boost the accuracy.

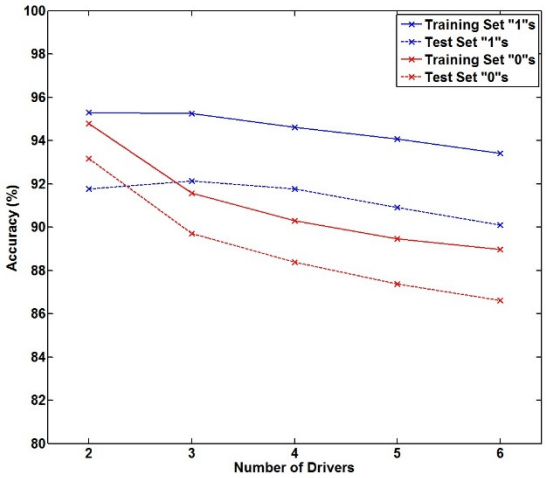

Figure 7. Accuracy of the SVM model as a function of the pool of drivers to identify a driver from. The "1" label indicates the driver to be identified, while the "0" label indicates the "Others" group of drivers.

Also investigated here is increasing the number of drivers in the "Others" group. The accuracy (Fig. 7) for the combined feature vector drops slightly, but is generally still around 90%.

c. Multinomial Logistic Regression

Using the speed, gas mileage, and acceleration probability distribution as feature vectors, we also implemented multinomial logistic regression. Unfortunately, with the full 160 elements of the feature vector, the multinomial logistic regression was unable to converge. We therefore performed Principal Component Analysis (PCA) on the dataset consisting of the six vehicles and identify the first 40 principle components. After performing this step, we proceed to implement the multinomial logistic regression with the 40 new features. Fig. 8 shows the result for the logistic regression when distinguishing between two drivers. As we increase the number of drivers in which to classify, the accuracy drops from 94% to 84%.

IV. Discussion

Our results show that the normalized distribution of speed, acceleration, and gas mileage are good features to be used for the machine learning task.

As we increase the number of drivers to be classified, the accuracy of the classification decreases. This decrease is expected, as with more drivers, there will be a greater overlap in the driving behaviors between different drivers. It is seen that using the SVM with two categories (driver-to-be-identified vs. "Others") generally gives better results than the multinomial logistic regression, especially when the number of drivers in the pool goes up. Depending on the application's required accuracy, our algorithm can be used to effectively classify within a small set of drivers.

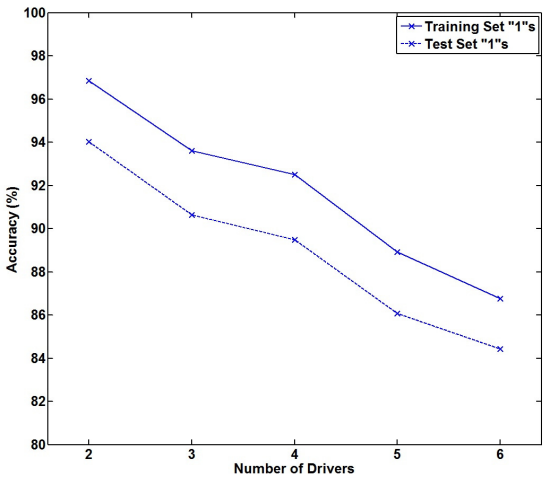

Figure 8. Accuracy of the multinomial logistic regression model as a function of the group of drivers to identify a driver from.

The feature vector currently includes speed, gas mileage, and acceleration probability distribution – information regarding the vehicle. It might be possible to improve our classification accuracy by using other information such as the amount of grip force that the driver exerts on the steering wheel, and the reaction time from the changing of the signal light to the actual movement of the car. This additional information can be included in the SVM or the logistic regression to improve the classification accuracy.

V. Conclusion

Presented is a model for learning the driving characteristics of a particular driver that could be used to identify him from amongst a pool of drivers. Using speed, gas mileage, and heading data, features of acceleration/deceleration profiles and turning speed vs. turn radius were extracted and binned to form a feature vector. SVM and multinomial logistic regression were performed and show that for a trip with a combined feature vector, a driver prediction accuracy of over 90% can be achieved, distinguishing amongst up to six drivers.

VI. Acknowledgements

We would like to thank Danny Goodman from MetroMile Inc. for supplying the dataset used in this work.

References

[1] Honda Motor Co., Online: http://world.honda.com/news/2013/4130321Congestion-Minimization-Technology/