An unsupervised approach for automotive driver identification

Poster Abstract

Nicholas Mainardi[*] Michele Zanella[*] Federico Reghenzani[*]

Niccoló Raspa[†] Carlo Brandolese[*]

ABSTRACT

The adoption of on-vehicle monitoring devices allows different entities to gather valuable data about driving styles, which can be further used to infer a variety of information for different purposes, such as fraud detection and driver profiling. In this paper, we focus on the identification of the number of people usually driving the same vehicle, proposing a data analytic work-flow specifically designed to address this problem. Our approach is based on unsupervised learning algorithms working on *non-invasive* data gathered from a specialized embedded device. In addition, we present a preliminary evaluation of our approach, showing promising driver identification capabilities and a limited computational effort.

CCS CONCEPTS

• **Computing methodologies → Cluster analysis**; • **Computer systems organization** → *Embedded and cyber-physical systems*;

1 INTRODUCTION

The spread of devices equipped with different type of sensors, as well as the continuous advancements in automotive electronics, increases the interest of researchers and companies in monitoring driving operations. This information can be used both (a) to develop more complex and adaptive Driver Assistance System (DAS) and (b) to gather data about drivers. The inferred knowledge can be employed in some scenarios (e.g., insurance companies) to profile drivers or detect possible frauds.

In this wide context, two similar problems are mainly found in literature: Driving Style Classification and Driver Identification. Both of them rely on raw data coming from sensors to infer knowledge about the driver. In the first case, the goal is to classify a driver according to predefined driving styles (e.g. calm/aggressive, good/bad) [8] in order to provide feedback to the driver with the extent of optimizing the energy usage of the car or improving the comfort of the ride. In the second case, the goal is to uniquely identify the driver for a given trip, which is useful for insurance purposes and anti-theft methods [1, 6, 9]. This work focuses on the identification of the number of people driving the same vehicle, a problem closer to Driver Identification. The data acquired from the sensors are bundled in *features*, which are either statistical-based (e.g. average speed) or event-based (e.g. braking events).

From a comprehensive survey of the literature [8], three main approaches can be identified: rule-based, model-based and learning-based. In particular, the increasing volume of data gathered by sensors has encouraged the usage of data-driven machine learning algorithms. Nevertheless, there are two relevant limitations in most of the previous works: they rely on *input data* mainly retrieved with invasive methodologies (i.e., reading from the vehicle Electronic Computed Board or the CAN bus [2, 4, 7]) and they leverage supervised techniques [2, 7, 10] (e.g., SVM, Random Forest Classifier, Neural Network), in turn requiring a labelled training data, which is generally harder to be obtained in real world application scenarios.

Our work tries to address the aforementioned issues, proposing a data analytic work-flow able to identify the number of drivers from unsupervised data collected with non-invasive methodologies.

2 THE PROPOSED APPROACH

In our approach, non-invasive monitoring device (plugged to OBD socket for power supply only) gathers motion data (3-axes accelerations and 3-axial angular velocity) and GPS data (position, altitude, speed). These data are enriched with reverse geocoded information about the type of the road (highway, urban etc.) and its speed limit. The gathered data are partitioned in *trips*, i.e. sets of measurements collected from the engine switch-on to the engine switch-off. Assuming there are no multiple drivers for the same trip, we can perform driver identification at trip level instead of measurement level. However, with this approach there is the need to devise a set of features for a trip to characterize the driving style. Then, the idea of our approach is that trips of the same driver should have similar values for these features: hence, by considering a trip as a point in an n dimensional space, where n is the number of features, then points representing trips of the same driver should be close, thus forming a cluster. Summing up, we identify the number of drivers by the number of clusters found in the set of trips.

From the data gathered by the monitoring device, we discard the GPS position and altitude, since they have no relation with the driving style. To extract the required features, we follow two approaches: (a) aggregate the motion data samples with some statistical measures – mean, variance, skewness and kurtosis – providing

[*] Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria <name.surname>@polimi.it.

[†] Politecnico di Milano, Scuola di Ingegneria Industriale e dell'Informazione.

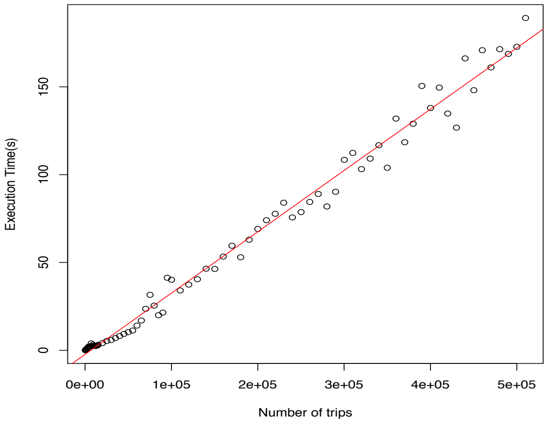

Figure 1: Execution time of PCA and clustering w.r.t. the number of trips. The red line shows the linear regression model, with equation $Time \approx 3.490 \cdot 10^{-4} \cdot Size$**.**

a set of purely statistical domain-agnostic features; (b) design a set of specific features in order to characterize the driving style, such as the number of speed infringements, number of peaks in the acceleration in the three directions (denoting a less smooth driving style), and the percentage of samples with accelerations or speed higher than specific thresholds, aptly chosen with the aid of domain experts. At the end of this design process, we identified about 40 features. Thus, the clustering algorithm has to deal with points in a 40 dimensional space, surely yielding bad performances due to the well-known *curse of dimensionality* issue. Therefore, before performing clustering, we need to shrink the number of features to avoid this issue. To this purpose, we apply Principal Component Analysis (PCA), which is generally able to combine original features to derive a smaller set of new features sorted according to their statistical significance. In our case, by retaining 75% of the statistical significance of the data, we reduced the number of features from 40 to 5–7 features depending on the trip. After shrinking the dimension of the data, it is possible to perform clustering. In this work, we choose to employ density-based clustering, in particular the *DBSCAN* [3] algorithm, since it does not require to specify the number of clusters to be built as an input parameter (in our case it must be an output of the algorithm). In addition, DBSCAN does not cluster all the points: if some of them are not close enough to other points, they are labeled as noise points not belonging to any cluster. This may be useful for anomaly detection as a further development of our work. Apart from the input data, DBSCAN requires two additional input parameters, which can be estimated by a heuristic proposed in [3, Section 4].

3 PRELIMINARY EVALUATION

The evaluation of the proposed approach is focused on two key aspects: (a) the accuracy estimation of the algorithm; (b) the computational complexity and scalability analysis. For the former validation a labeled dataset is needed: each trip must be labeled with the associated driver. Unfortunately, the current public available datasets do not fit the needs of our evaluation, forcing us to resort to a driver-unlabeled dataset (gathered in UK region). Collecting a real labelled dataset for validation purpose is an ongoing work.

Accuracy analysis. In order to check if the proposed approach produces realistic result, we compute the average number of drivers per vehicle for the UK region, since this value strongly depends on the geosocial conditions. From the public available government data [5], we get the average number of adult people per vehicle and the number of drivers per adult people obtaining an average of 1.095 drivers per vehicle. Our results match this estimation: our approach identifies more than one drivers in about 10% of the vehicles.

Scalability analysis. Given that the computational complexity of PCA is linear in the number of trips n, the average-case complexity of the overall program is dominated by DBSCAN algorithm, i.e. $O(n \log n)$. To measure actual execution times, we perform an experimental evaluation on a high-end server machine [1]. Figure 1 shows that the actual value of execution time can be approximated by a linear trend, due to the minimal impact of the $\log n$ term.

4 CONCLUSIONS

In this paper, we propose a work-flow to identify the number of people usually driving the same vehicle, improving upon previous works by leveraging only non-invasive unlabeled data. Moreover, we provide some preliminary results, showing promising driver identification capabilities with limited computational effort.

ACKNOWLEDGMENT

This work was supported in part by the European Community under grant agreement no. 688201 (M^{2}DC – EU H2020).

REFERENCES

[1] Giovanni Agosta, Alessandro Barenghi, and et al. 2016. V2I Cooperation for Traffic Management with SafeCop. In *2016 Euromicro Conference on Digital System Design, DSD 2016, Limassol, Cyprus, August 31 - September 2, 2016*. 621–627. https://doi.org/10.1109/DSD.2016.18

[2] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno. 2016. Automobile driver fingerprinting. *Proceedings on Privacy Enhancing Technologies* 2016, 1 (2016), 34–50. https://doi.org/10.1515/popets-2015-0029

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad (Eds.). AAAI Press, 226–231. http://www.aaai.org/Library/KDD/1996/kdd96-037.php

[4] U. Fugiglando, E. Massaro, P. Santi, S. Milardo, K. Abida, R. Stahlmann, F. Netter, and C. Ratti. 2017. Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment. *CoRR* abs/1710.04133 (2017). arXiv:1710.04133

[5] UK Government. 2016. Driving licence holding and vehicle availability (NTS02). National Travel Survey.

[6] B. I. Kwak, J. Woo, and H. K. Kim. 2017. Know Your Master: Driver Profiling-based Anti-theft Method. *CoRR* abs/1704.05223 (2017). arXiv:1704.05223

[7] M. Van Ly, S. Martin, and M. M. Trivedi. 2013. Driver classification and driving style recognition using inertial sensors. In *2013 IEEE Intelligent Vehicles Symposium (IV)*. 1040–1045. https://doi.org/10.1109/IVS.2013.6629603

[8] C. Marina Martinez, M. Heucke, F. Y. Wang, B. Gao, and D. Cao. 2018. Driving Style Recognition for Intelligent Vehicle Control and Advanced Driver Assistance: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 19, 3 (March 2018), 666–676. https://doi.org/10.1109/TITS.2017.2706978

[9] Ariel Oleksiak and Michal Kierzynka et al. 2017. M2DC - Modular Microserver DataCentre with heterogeneous hardware. *Microprocessors and Microsystems* 52 (2017), 117 – 130. https://doi.org/10.1016/j.micpro.2017.05.019

[10] V. Vaitkus, P. Lengvenis, and G. Žylius. 2014. Driving style classification using long-term accelerometer information. In *2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR)*. 641–644. https://doi.org/10.1109/MMAR.2014.6957429

[1] A single-core R implementation on AMD Opteron 8435 cores and 128 GB of RAM