Random Forest-Based Approach for Physiological Functional Variable Selection for Driver's Stress Level Classification

Jean-Michel Poggi [1] & Neska El Haouij [2] & Raja Ghozi [3] & Sylvie Sevestre Ghalila [4] & Mériem Jaïdane [5]

[1] *Paris Descartes Univ., France & Paris Sud Univ., France*
Jean-Michel.Poggi@math.u-psud.fr
[2] *CEA-LinkLab, Tunisia & Tunis El Manar Univ., Tunisia & Paris Sud Univ., France*
elhaouij.nsk@gmail.com
[3] *Tunis El Manar Univ., Tunisia & CEA-LinkLab, Tunisia, rjghozi@yahoo.com*
[4] *CEA-LinkLab, Tunisia, sylvie.sevestre-ghalila@cea.fr*
[5] *Tunis El Manar Univ., Tunisia & CEA-LinkLab, Tunisia, meriem.jaidane@planet.tn*

Résumé. Avec l'urbanisation croissante et les progrès technologiques, la conduite automobile urbaine est une tâche complexe qui exige un niveau élevé de vigilance. Ainsi, la charge mentale du conducteur doit être optimale afin de gérer des situations critiques dans de telles conditions de conduite. Les études antérieures sur les performances du conducteur reposaient sur l'utilisation de mesures subjectives. La nouvelle technologie de capteurs portables et non intrusifs, fournit non seulement une surveillance physiologique en temps réel, mais enrichit également les outils de surveillance des états affectifs et cognitifs humains.

Cette étude se concentre sur les changements physiologiques du conducteur mesurés à l'aide de capteurs portatifs dans différentes conditions de circulation urbaines. Plus précisément, l'activité électrodermale (EDA) mesurée en deux endroits différents : main et pied, l'électromyogramme (EMG), la fréquence cardiaque (HR) et la respiration (RESP) sont enregistrés lors de dix expériences de conduite sur trois types de routes. Les données de conduite considérées sont issues de la base de données physiologiques *drivedb*, disponible en ligne sur le site PHYSIONET.

Plusieurs études ont été réalisées sur la reconnaissance du niveau de stress à partir de signaux physiologiques. Classiquement, la stratégie consiste dans l'extraction par des experts de descripteurs de signaux physiologiques et la sélection des caractéristiques les plus pertinentes dans la reconnaissance du niveau de stress par une méthode statistique classique.

Le présent travail fournit une méthode basée sur la forêts aléatoires pour la sélection de variables physiologiques fonctionnelles afin de classer le niveau de stress au cours de l'expérience de conduite. La contribution de cette étude est double : sur le plan méthodologique, elle considère les signaux physiologiques comme des variables fonctionnelles et adapte une procédure de traitement et de sélection de variables de telles données. Du côté appliqué, la méthode proposée fournit une procédure "aveugle" de classification

du niveau de stress du conducteur qui ne dépend pas des études d'experts des signaux physiologiques.

Mots-clés. Forêts aléatoires, Sélection de variables, Données fonctionnelles, Signaux physiologiques

Abstract. With the increasing urbanization and technological advances, urban driving is bound to be a complex task that requires higher levels of alertness. Thus, the drivers mental workload should be optimal in order to manage critical situations in such challenging driving conditions. Past studies relied on drivers performances used subjective measures. The new wearable and non-intrusive sensor technology, is not only providing real-time physiological monitoring, but also is enriching the tools for human affective and cognitive states monitoring.

This study focuses on a drivers physiological changes using portable sensors in different urban routes. Specifically, the Electrodermal Activity (EDA) measured on two different locations: hand and foot, Electromyogram (EMG), Heart Rate (HR) and Respiration (RESP) of ten driving experiments in three types of routes are considered: rest area, city, and highway driving issued from physiological database, labelled *drivedb*, available online on the PHYSIONET website.

Several studies have been done on driver's stress level recognition using physiological signals. Classically, researchers extract expert-based features from physiological signals and select the most relevant features in stress level recognition. This work aims to provide a random forest-based method for the selection of physiological functional variables in order to classify the stress level during real-world driving experience. The contribution of this study is twofold: on the methodological side, it considers physiological signals as functional variables and adapts a procedure of data processing and variable selection. On the applied side, the proposed method provides a "blind" procedure of driver's stress level classification that do not depend on the expert-based studies of physiological signals.

Keywords. Random Forests, Variable Selection, Functional Data, Physiological Signals

1 Introduction

This paper aims to provide a random forests-based method for the selection of physiological functional variables in order to classify the stress level experienced during real-world driving. For that, we present first the context of our work which concerns the affective computing aspects with a summary of the study introducing the physiological database *drivedb*. Then, methods on functional data, variable selection using random forests and grouped variables importance are addressed. The contribution of this study is twofold: on the methodological side, it adapts the scheme proposed by [6] to take advantage of the functional nature of the physiological data and offers a procedure of data processing

and variable selection. On the applied side, the proposed method provides a blind (i.e. without prior information) procedure of driver's stress level classification that does not depend on the extraction of expert-based features of physiological signals. This allows automatic exploration of promising signals to be included in statistical models for driver's state recognition.

2 Stress level recognition while driving

Many research groups tried to provide solutions and tools to vehicles and roadway users in order to improve safety, efficiency and quality in the sector of transport. [14] points out that according to the American Highway Traffic Safety Administration, high stress levels impact negatively drivers reactions especially in critical situations. It is one of the most prominent causes of vehicle accidents such as intoxication, fatigue and aggressive driving. In real world driving, human affective state monitoring can offer useful information to avoid traffic incidents and provide safe and comfortable driving.

With the increasing urbanization and technological advances, the new wearable and non-intrusive sensor technology, is not only providing real-time physiological monitoring, but also is enriching the tools for human affective and cognitive states monitoring. In particular, several studies have been reported the last years in the field of driver's stress monitoring. In this paper, we base our analysis on the study of [10] where they presented a protocol of physiological data collection in real-world driving conditions in order to detect stress levels. Specifically, physiological signals such as Electrodermal Activity (EDA), Electrocardiogram (ECG), Electromyogram (EMG) and Respiration (RESP) were captured for 24 driving experiences.

Features derived from non-overlapping segments of physiological signals taken from rest, highway and city of the driving experiences. The first analysis aiming to classify the stress levels allows to distinguish between the three levels of driver stress with an accuracy of 97%. The second analysis concerns the study of the correlation between extracted features from physiological signals and a stress levels metric created from the video tape. In this study, [10] reported that there is a correlation between driver's affective state quantified by the stress levels metric and the physiological signals, the highest correlation is with the EDA and HR. They have partially released their physiological database, labeled "*drivedb*", on-line on the PhysioNet website[1]. The data used in our work were extracted from the *drivedb* database which has a clear annotation of the several driving periods for each experience, allowing an easy exploitation of the information. Apart its availability on-line, various studies were based on this database which constitutes a main reference on stress level recognition in highway and city driving.

[1] http://physionet.org/

3 Functional Variable Selection

The main issue of variable selection methods is their instability where a set of selected variables may change when perturbing the training sample. The most widely used solution to solve this instability consists in using bootstrap samples where a stable solution is obtained by aggregating selections achieved on several bootstrap subsets of the training data. Random forests algorithm, introduced by [1], is one of these methods based on aggregating a large collection of tree-based estimators. These methods have good predictive performances in practice and they work well for high dimensional problems. Their power is shown in several studies summarized in [15]. Moreover, random forests provide several measures of the importance of the variables with respect to the prediction of the outcome variable. It has been shown that the permutation importance measure introduced by Breiman, is an efficient tool for selecting variables ([2, 5, 7]).

The standard approach in Functional Data Analysis (FDA) (see for example [13, 3]) consists in projecting the functional variables into a space spanned by a functional basis such as splines, wavelets, Fourier. Several regression and classification methods were the focus of studies in two situations: with one functional predictor and recently for several functional variables.

Classification based on several possibly functional variables has also been considered using the CART algorithm for similar driving experiences in the study of [12], using SVM in [16] work. Variable selection using random forests was achieved in the study of [4]. In our study, multiple FDA using random forests and the grouped variable importance measure proposed by [6] are used.

3.1 Variable Selection using Random Forest-based Recursive Feature Elimination

In this study, Random Forests-based Recursive Feature Elimination (RF-RFE) is used. The RF-RFE algorithm, proposed by [6], was inspired from [9] introducing Recursive Feature Elimination algorithm for SVM (SVM-RFE). At the first step, the dataset is randomly split into a training set containing two thirds of the data and a validation set containing the remaining one third. The procedure fits the model to all explanatory variables using Random Forests. Then, the variables are ranked using their importance measure. The grouped VI is computed only on the training set. The less important predictor is eliminated, the model is refit and the performance is assessed by a prediction error computed on the validation set. The variable ranking and elimination is repeated until no variable remains. The final model is chosen by minimizing the prediction error. It should be noted that at each iteration, the predictors importance is recomputed on the model composed by the reduced set of explanatory variables.

In the case of functional variables, the selection is performed using the algorithm on two different types of groups, thanks to the definition of importance of groups of

variables. This allows to consider a group of variables as a whole, for example the group of the wavelet coefficients of a given signal, and to quantify its relative importance with respect to the other functional variables.

3.2 Our procedure: Variable selection using iterative RF-RFE

The proposed approach in this work aims to first eliminate the irrelevant physiological variables in the stress level classification task and then select among each kept variable the most relevant wavelet levels. In this study, the number of variables is very large (20480), compared to the number of the observations (68), thus the procedure is not stable. In order to reduce the variability of the selection, the procedure is repeated 10 times.

3.3 Variable selection results

The objective of variable selection is first to eliminate physiological signals that do not contribute significantly in the stress level classification, then for the retained physiological variables, the most relevant wavelet levels will be selected.
When applying our procedure to the drivedb database, we perform at a first stage functional variables decomposition using the Haar wavelet which is considered as the simplest one. We pick 12 as the decomposition level which corresponds to the maximum level compatible with the $4096 = 2^{12}$ samples.
To achieve this work, we use the `R` software, with the `randomForest` package proposed by [11] and `RFgroove` packages developed by [8].

The proposed "blind" approach performs as the expert-based approach in terms of misclassification rate. This procedure offers moreover, additional information such as the physiological variables ranking according to their importance and the list of the relevant variables in stress level classification. The obtained results suggest that *EMG* and the *HR* are not very relevant when compared to the EDA and the respiration signals. This may help to investigate the list of physiological sensors that can be proposed to the smart vehicles designers, in order to determine the stress level.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):1–13, 2006.

[3] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] R. Genuer, J.-M. Poggi, and Tuleau-Malot. Vsurf: An r package for variable selection using random forests. *The R Journal*, 7(2):19–33, 2015.

[5] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225 – 2236, 2010.

[6] B. Gregorutti, B. Michel, and P. Saint-Pierre. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics and Data Analysis*, 90:15–35, 2015.

[7] B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, pages 1–20, 2016.

[8] Gregorutti, B. Rfgroove: Importance measure and selection for groups of variables with random forests. *https://CRAN.R-project.org/package=RFgroove, R package version 1.1, (2016)*, 2016.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, March 2002.

[10] J.-A. Healey and R.-W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, June 2005.

[11] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[12] J.-M. Poggi and C. Tuleau. Classification of objectivization data using cart and wavelets. *Proceedings of the IASC 07, Aveiro, Portugal*, pages 1–8, 2007.

[13] J.-O. Ramsay and B.-W. Silverman. *Functional Data Analysis*. Springer-Verlag New York, 2005.

[14] R.-G. Smart, E. Cannon, A. Howard, P. Frise, and R.-E. Mann. Can we design cars to prevent road rage? *International Journal of Vehicle Information and Communication Systems*, 1(1-2):44–55, 2005.

[15] A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, 2011.

[16] K. Yang, H. Yoon, and C. Shahabi. A supervised feature subset selection technique for multivariate time series. *Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning with Statistics*, pages 92–101, 2005.