Comparison of Different Driving Style Analysis Approaches based on Trip Segmentation over GPS Information

Marco Brambilla
Politecnico di Milano, DEIB.
Italy
marco.brambilla@polimi.it

Paolo Mascetti
DataBiz Srl.
Italy
paolo.mascetti@databiz.it

Andrea Mauri
Politecnico di Milano, DEIB.
Italy
andrea.mauri@polimi.it

ABSTRACT

Over one billion cars interact with each other on the road every day. Each driver has his own driving style, which could impact safety, fuel economy and road congestion. Knowledge about the driving style of the driver could be used to encourage "better" driving behaviour through immediate feedback while driving, or by scaling auto insurance rates based on the aggressiveness of the driving style.

In this work we report on our study of driving behaviour profiling based on unsupervised data mining methods. The main goal is to detect the different driving behaviours, and thus to cluster drivers with similar behaviour. This paves the way to new business models related to the driving sector, such as Pay-How-You-Drive insurance policies and car rentals.

Driver behavioral characteristics are studied by collecting information from GPS sensors on the cars and by applying three different analysis approaches (DP-means, Hidden Markov Models, and Behavioural Topic Extraction) to the contextual scene detection problems on car trips, in order to detect different behaviour along each trip. Subsequently, drivers are clustered in similar profiles based on that and the results are compared with a human-defined groundtruth on drivers classification. The proposed framework is tested on a real dataset containing sampled car signals. While the different approaches show relevant differences in trip segment classification, the coherence of the final driver clustering results is surprisingly high.

1. INTRODUCTION

According to the global status report on road safety conducted by the World Health Organisation (WHO) in 2013, 1.24 million traffic-related fatalities occur annually worldwide, currently the leading cause of death for people aged between 15 and 29 years. In the majority of the cases accidents are caused by risky driving behavior[1]. Driving is essentially a multi-factor cognitive task based on the underlying road layout, traffic, weather, and social context. Drivers modeling is based on on-road behavior analysis and it allows users' segmentation into categories based upon their driving style [8]. Exploiting this segmentation can bring a great impact in road safety and on business models related to driving, such as Pay-How-You-Drive insurance policies and car rentals.

Current driver characterization methods are mainly based on the process of identification and recognition of patterns defined by prior research studies and adapted to a specific context through supervised learning techniques [7]. However, this kind of analysis lacks in generalization, misses the selection phase of the best set of features to consider, and implies complex human labeling of multivariate time series for the definition of classes or sets of rules for the automatic identification.

In this work we propose an approach aimed to the identification of patterns characterizing driving behaviours independently from prior knowledge concerning the driving process. In this way the relation between features (e.g. cognitive, operational and contextual) can be fully investigated. Hence we propose three unsupervised approaches: a clustering based segmentation, a sequences analysis segmentation, and a behavior characterization obtained with a soft-clustering technique. Through these techniques we classify trip segments, and based on this we apply a second clustering step on trips (and thus drivers). We compare the different techniques and we assess the correctness of drivers clustering against a human-defined ground truth.

The paper is structured as follows: Section 2 describes the applied methods, Section 3 reports on our experiments and discussion, Section 4 describes the related work and finally Section 5 concludes.

2. CONTEXTUAL SCENE EXTRACTION

In this section we describe three different unsupervised approaches, namely clustering methods, to extract observed recurrent patterns (named *contextual scenes*) in the behaviour of different drivers. This allows the extraction of a reduced

[1] http://www.who.int/gho/publications/world_health_statistics/2016/whs2016_AnnexA_RoadTraffic.pdf?ua=1&ua=1

SAC'17, April 3-7, 2017, Marrakesh, Morocco
Copyright 2017 ACM 978-1-4503-4486-9/17/04...$15.00
http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

representation of the original multivariate time series describing the trips of drivers.

2.1 DP-Means for Contextual Scenes Clustering

The first method is the DP-Means algorithm, a Bayesian non-parametric clustering approach [4]. We apply this algorithm with the aim to group observation points into contextual scenes representing a behavioral pattern in driving trips. This algorithm infers the number of groups considering data similarity based on euclidean distance measured between processed points. In fact, a new cluster is formed whenever a point is farther than λ away from every existing cluster centroid. The parameter λ controls the trade-off between traditional k-means objective function and the cluster penalty term introduced by DP-means formulation. Higher values of λ discourage the creation of more clusters.

Since this analysis is accomplished in an unsupervised setting, to determine the best number of cluster we rely on an *well known evaluation metrics*, namely the *Silhouette* similarity coefficient, which computes how tightly grouped are the points in each cluster. Assuming to have k clusters, for each datum i, let $a(i)$ be the average dissimilarity of i with all other data in the same cluster and let $b(i)$ be the lowest dissimilarity of i to any other cluster, of which i is not member. The *silhouette* index of i is defined as:

$$s\left(i\right) = \frac{b\left(i\right) - a\left(i\right)}{max\{a\left(i\right), b\left(i\right)\}} \qquad -1 \leq s(i) \leq 1 \tag{1}$$

If $s\left(i\right)$ is near to 1, object i has been assigned to an appropriate cluster. Thus, the average $s\left(i\right)$ consider the appropriateness of clustering of all data points. This allows to determine the optimal number of clusters k, by running the clustering algorithm several times with different parameter values and choosing k which yields the highest silhouette.

2.2 Hidden Markov Model based Segmentation

The second method is Hidden Markov Model (HMM), applied to extract contextual scenes. The proposed model is a statistical latent process model which assumes that the observed driving behaviour is governed by sequence of hidden (unobserved) activities. HMMs introduce the concept of sequence and relationship between consequent observations and hidden states. We employ this models to perform unsupervised segmentation of gathered trip points in order to learn behavioral patterns described by their latent structure. In particular we used continuous emissions HMMs where hidden states generative process is described by multivariate Gaussian distributions.

The overall process is composed mainly by three phases: model initialization based on clustering results; model training aimed to estimate the model parameters; and sequence decoding.

HMMs unsupervised training process requires knowledge of the hidden structure; more specifically, the number of the latent variables and the initialization parameters need to be specified. We perform model initialization based results obtained by DP-means algorithm. Starting from DP-clustering results, Multivariate Gaussian Distribution parameters (means and covariances) are estimated for every hidden state. Once hidden state parameters and transition behavior are learnt, we perform the tagging of each observed sequence to assign each observed sample to the correspondent generative hidden state. In particular, the model parameters are estimated making use of *Baum-Welch* algorithm that uses a Expectation Maximization (EM) algorithm to find the maximum likelihood of the parameters of the HMM given a set of feature vectors.

Once the HMM's parameter are estimated by the learning algorithm, our goal is to tag each observation point according to the corresponding generative hidden state. To achieve sequence labeling we used the *Viterbi algorithm* [19].

2.3 Hierarchical Dirichlet Processes for Behavioural Topics Extraction

So far we analyzed methods producing an hard clustering assignment for each observed data point and the ulterior clusters' distribution has been used to characterize trip belonging to different users. Now we apply a Hierarchical Dirichlet Processes to model topical concepts belonging to a set of documents.

Straub et al. in [15] proved that HDPs are able to obtain descriptive topics about road-states considering a small set of car signals. Starting from this points we used the same approach to compare driver models belonging to different drivers to allow a clustering based on their driving behaviour and habits.

Most of the time, the number of topics for a corpus is unknown. In that case, a non-parametric model is a good choice since the number of parameters (such as the number of topics) is not set a priori, but learned from data. This model can adapt the number of topics based on the data. The overall process is composed mainly by three phases: (1) data discretization, (2) documents and corpus creation, and finally (3) topic extraction phase.

2.3.1 Data Discretization

The first step consists in discretizing the continuous features into categorical ones. Our goal is to represent each observation point with a string of length D where D is the number of features. Each symbol in the string represents a discretized feature value. In this transformation the number of features taken into account is fundamental because defines the length of each word of ghe corpus and may affect the model performance. Furthermore, it is essential to define the number of categorical values associated to each signal.

Two methods can be used for feature discretization: (i) *Clustering:* in this technique each features is grouped using clustering algorithm; or (2) *Binning:* in this alternative each signal spanning range is divided in categorical bins: Each bin is defined by two threshold values: starting point and ending point. The size of each bin may be of different length based on signal's distribution (equal frequencies) or it may be constant (equal intervals) in order to easily recognize outliers. In our case we opted for binning with equal intervals.

2.3.2 *Corpus Creation*

Once each observation signal has been quantized, we built a text corpus on which to train our model. Quantized observations representation (which we will call *words*), are grouped into *documents* based on their trip membership.

2.3.3 *Topic Analysis*

The topics are analyzed using a Hierarchical Dirichlet Process (HDP) which applies Sethuraman' s stick breaking construction twice, as described in [14]. This kind of construction, unlike the original ones derived by Teh *et al.* [17], allows the derivation of an efficient and scalable stochastic variational inference, as proposed by Wang *et al.* [20]. These approximation assumes that all the variables involved in the process are independent and it truncates the stick-breaks to T on the corpus level and to K on the document level. This fact does not affect the results since truncation level can be set high enough to allow the HDP to adapt to the complexity of the data.

3. EXPERIMENT

In this section we describe how we performed the analysis described in the previous sections on a real world dataset.

3.1 Dataset Description

In this work we used the XSens[2] dataset, a collection of observations retrieved during an evaluation study of a driving behavior collection system. Each observation represents an observed driving trip and each sample represents a set of signals' observation.

3.1.1 *Data Collection*

The data have been preprocessed by the collection device Xsens MTi-G-710, a 3D motion tracking device, which performs an initial filtering process including speed estimation based on GPS positioning. The used coordinate system used is known as ENU and is the standard in inertial navigation for aviation and geodetic applications.

3.1.2 *Sampling Frequency and Down-Sampling*

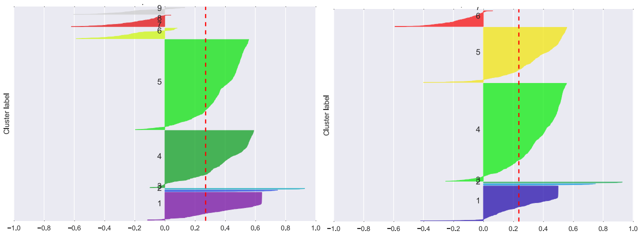
The retrieved signals have their own sampling frequency that can vary from one device and sensor to another. For instance, there is a different sampling frequency between GPS positioning (1Hz) and inertial measurements (100Hz). To overcome this problem observations are grouped using a temporal window size of 1 seconds, decision based on the slower sampling frequency. In subsampling, for the inertial features we consider the mean value, while for the GPS coordinates we consider the median (to avoid generating "fake" mean GPS positions).

3.1.3 *Features Selection*

The features used in the experiments have been chosen relying on prior knowledge regarding their relevance in Driving Behavior Modeling [10, 6]. Besides the ones already available, a few others have been computed, such as the difference in orientation (yaw) with respect to the previous instant. The selected features are: **acceleration** (on Y and X axes), **speed** (on Y and X axes) and the **difference in yaw**.

Table 1: Silhouette Coefficients depending on clustering parameters

Parameter	Number of Clusters	SC
λ =5	36	0.182
λ =7	18	0.198
λ =10	9	0.270
λ =12	7	0.250
λ =13	2	0.927



(a) SC for λ = 10 (b) SC for λ = 12

Figure 1: Silhouette Coefficient representation. The vertical dashed red line represents average SC.

3.2 DP-Means for Contextual Scenes Clustering

We run DP-means algorithm over the dataset to assign to each observed trip point the corresponding Contextual Scene. In the datasets data points belonging to each trip are ordered according to the timestamp and trips are ordered by collection sequence. It is important to highlight this structural ordering of clustering dataset because DP-means algorithm results depend on the order in which data points are processed. A different ordering of data points or driving trip may produce different clustering results.

We performed Silhouette Coefficient (SC) Analysis to tune the clustering parameters when applying DP-means. The results are summarized in Table 1 and Figure 1. The optimal SC values correspond to values of λ equal to 10 or 12 (we discard λ=13 because it yields only 2 clusters, with all trips but one in the same cluster). We decided to use λ=12 since the corresponding SC is very similar to the one for λ=10 and it generates less clusters. The centroids are represented in Table 2: for instance Cluster 1 identifies straight-line proceeding at low speed while Cluster 7 groups points of harsh maneuvers with high lateral and longitudinal accelerations.

To support the visual exploration of clusters and centroids, Figure 2 reports a graphical representation of the features distribution within each cluster allowing one to understand what kind of behavior or Contextual Scene can be associated to each cluster, based on the values of the features and the allocation of clustered points. The visualization makes use of histograms where each graph represents the distribution of a feature across the clusters.

Furthermore, each trip and consequently each driver can be characterized by the distribution of the points belonging to

[2] http://selfracingcars.com/blog/2016/8/22/xsensfairchild-data-available

Table 2: DP-Means clustering centroids with $\lambda = 12$.

	Acc_X	**Acc_Y**	**Speed_X**	**Speed_Y**	**Diff_Yaw**
Cluster 1	-0.2509	-0.0401	1.6041	1.2607	1.4984
Cluster 2	0.0699	-1.3350	5.4737	0.7912	224.8210
Cluster 3	-0.0219	1.0751	7.0265	0.7751	200.4068
Cluster 4	-0.0505	-0.0417	7.8803	9.2362	0.0483
Cluster 5	-0.0218	0.4173	9.7849	3.0444	0.7412
Cluster 6	0.7957	-0.3374	4.0299	4.0781	1.3649
Cluster 7	2.3836	-4.0015	5.2456	4.5716	-2.3374

each cluster. The percentage of each cluster's points within each trip can be used as characterizing feature of individual driving style. These information about cluster distribution in each trip are summarized in Table 3.

Table 3: The percentage of points belonging to each cluster have been computed considering every observed driving trip. Similar trips present a similar clusters distribution.

Trip_ID	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
t1	0.17	0.01	0.00	0.48	0.27	0.08	0.00
t2	0.49	0.01	0.01	0.15	0.20	0.15	0.00
t3	0.51	0.01	0.01	0.15	0.21	0.12	0.00
t4	0.31	0.01	0.00	0.34	0.17	0.15	0.01
t5	0.43	0.01	0.00	0.18	0.25	0.13	0.00
t6	0.41	0.01	0.01	0.18	0.24	0.15	0.00
t7	0.36	0.01	0.01	0.31	0.16	0.14	0.02
t8	0.33	0.02	0.01	0.32	0.17	0.14	0.01
t9	0.44	0.01	0.01	0.20	0.22	0.13	0.00
t10	0.39	0.01	0.01	0.21	0.24	0.13	0.00
t11	0.45	0.01	0.01	0.17	0.24	0.12	0.00
t12	0.46	0.01	0.01	0.14	0.20	0.18	0.00
t13	0.43	0.01	0.01	0.18	0.24	0.13	0.00
t14	0.34	0.02	0.01	0.32	0.16	0.14	0.02
t15	0.17	0.00	0.00	0.52	0.24	0.06	0.00
t16	0.49	0.01	0.01	0.16	0.20	0.14	0.00
t17	0.48	0.01	0.01	0.15	0.23	0.12	0.00
t18	0.47	0.01	0.01	0.16	0.21	0.13	0.00
t19	0.48	0.01	0.00	0.14	0.22	0.14	0.00
t20	0.30	0.02	0.00	0.33	0.18	0.15	0.01
t21	0.50	0.01	0.01	0.14	0.21	0.13	0.00
t22	0.48	0.01	0.01	0.16	0.21	0.13	0.00
t23	0.35	0.01	0.00	0.31	0.16	0.15	0.02
t24	0.56	0.01	0.00	0.13	0.18	0.12	0.00
t25	0.43	0.01	0.01	0.18	0.25	0.12	0.00
t26	0.53	0.00	0.00	0.15	0.15	0.16	0.00
t27	0.50	0.01	0.00	0.16	0.20	0.13	0.00

3.3 Hidden Markov Models Based Segmentation

HMM is trained using the dataset and each trip's observation have been tagged to assign the corresponding hidden state. The result is similar to clustering process, where each point is assigned to the nearest cluster, but HMM has intrinsic information about the probability of all the possible state changes. This property of the system is expressed by the transition matrix as described in Table 5 where it has been computed considering an HMM model initialized with seven hidden states.

High values of self transition represent behaviors that tend to last in time for long period and instead lower values of self-transition probability characterize behavioral patterns that have short term.

Similarly to clusters' centroids, in the extracted HMM we evaluated mean and covariance of Multivariate Gaussian distributions belonging to latent states. In Table 4 are represented means vector of Multivariate Gaussian Distribution belonging to each latent state. Furthermore we computed the distribution of each hidden state within each observed driving trip as described in Table 6. Similar trips present a similar clusters distribution.

An example of clustering is shown in Figure 3, where the colors represents the points belonging to different clusters.

Table 4: HMM Gaussian emission means with k=7. In the Table are represented Multivariate Gaussian features' means associated to the identified hidden state.

	Acc_X	**Acc_Y**	**Speed_X**	**Speed_Y**	**Diff_Yaw**
Cluster 1	0.0002	-0.0061	0.0779	0.0803	0.0034
Cluster 2	0.2071	-1.4373	4.8023	2.3062	-69.1111
Cluster 3	-0.5347	1.0825	5.4776	1.3215	68.0753
Cluster 4	0.0054	-0.01035	8.5401	6.9008	0.0125
Cluster 5	-0.1031	0.4770	7.9902	2.9251	-0.0544
Cluster 6	0.1472	-0.3185	5.0797	6.3760	1.3183
Cluster 7	-0.0218	-0.0479	1.5448	1.7622	3.8423

Table 5: Standard-HMM Transitions Matrix initialized with $k = 7$.

	State 1	**State 2**	**State 3**	**State 4**	**State 5**	**State 6**	**State 7**
State 1	0.9501	∼0	∼0	∼0	∼0	∼0	0.0499
State 2	∼0	0.6176	0.0375	∼0	0.0940	0.1509	0.1000
State 3	∼0	∼0	0.6101	∼0	0.1972	0.0199	0.1728
State 4	0.0001	0.0006	∼0	0.9217	0.0348	0.0392	0.0036
State 5	∼0	0.0021	0.0492	0.0759	0.8491	0.0164	0.0073
State 6	∼0	0.0524	∼0	0.1123	0.0123	0.7943	0.0286
State 7	0.0515	0.0365	0.0079	0.0020	0.0045	0.0320	0.8645

Table 6: Hidden States assignments distribution in trips.

Trip_ID	State1	State2	State3	State4	State5	State6	State7
t1	0.06	0.01	0.01	0.59	0.11	0.14	0.07
t2	0.10	0.02	0.02	0.40	0.15	0.16	0.15
t3	0.27	0.02	0.02	0.32	0.12	0.12	0.13
t4	0.11	0.06	0.03	0.31	0.20	0.10	0.19
t5	0.17	0.02	0.02	0.32	0.15	0.14	0.17
t6	0.13	0.02	0.02	0.37	0.13	0.13	0.19
t7	0.13	0.05	0.04	0.28	0.19	0.09	0.21
t8	0.10	0.06	0.04	0.28	0.19	0.11	0.22
t9	0.22	0.03	0.02	0.34	0.11	0.15	0.13
t10	0.21	0.02	0.02	0.32	0.12	0.15	0.13
t11	0.17	0.02	0.02	0.31	0.13	0.13	0.20
t12	0.22	0.02	0.02	0.24	0.12	0.16	0.20
t13	0.26	0.02	0.02	0.29	0.12	0.14	0.14
t14	0.06	0.05	0.03	0.28	0.20	0.12	0.26
t15	0.09	0.00	0.01	0.64	0.07	0.12	0.08
t16	0.24	0.02	0.02	0.27	0.14	0.12	0.18
t17	0.26	0.02	0.02	0.31	0.13	0.15	0.11
t18	0.25	0.02	0.02	0.28	0.13	0.14	0.16
t19	0.18	0.02	0.02	0.37	0.12	0.14	0.16
t20	0.11	0.09	0.03	0.26	0.20	0.14	0.18
t21	0.28	0.02	0.02	0.25	0.14	0.16	0.12
t22	0.26	0.02	0.02	0.30	0.13	0.11	0.16
t24	0.33	0.02	0.02	0.21	0.10	0.12	0.21
t23	0.09	0.05	0.03	0.28	0.19	0.11	0.25
t25	0.20	0.03	0.02	0.31	0.13	0.14	0.17
t26	0.20	0.02	0.01	0.24	0.10	0.20	0.23
t27	0.27	0.02	0.02	0.25	0.14	0.13	0.17

3.4 Hierarchical Dirichlet Processes for Behavioural Topic Extraction

For topic extraction we used an already provided implementation of HDP model provided by *Gensim* library [11]. Gensim is licensed under the OSI-approved GNU LGPL license. We discretized the dataset using a binning method (Binning

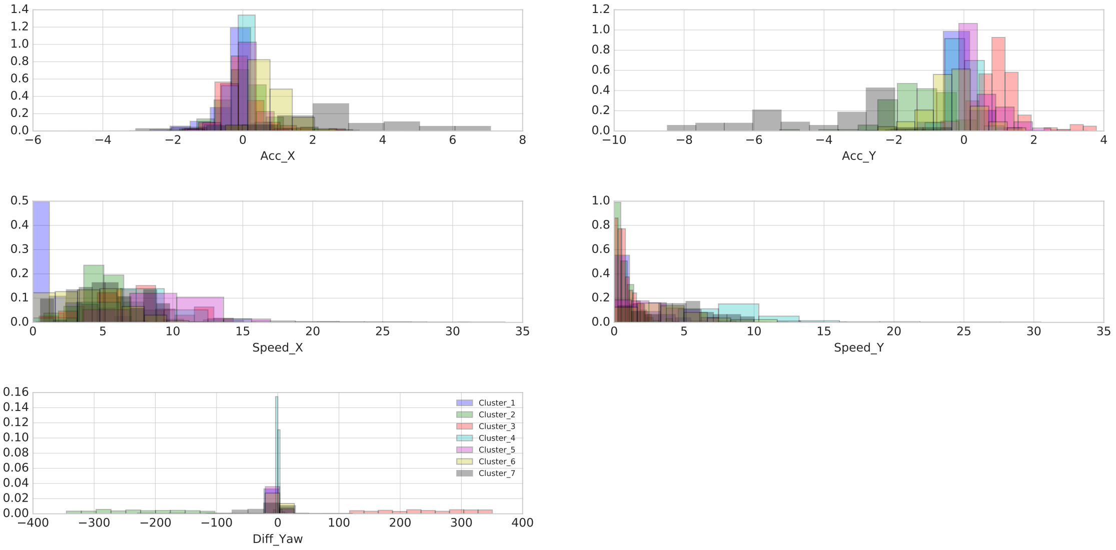


Figure 2: Features distribution on the identified clusters: each graph represents a feature used for clustering and each color represents a cluster. One can appreciate how elements of each cluster are spread across each feature values.

Table 7: Thresholds values defined in binning process.

	a	**b**	**c**	**d**	**e**
Acc_X	≤ -2.55	$-2.55 < v \leq -0.14$	$-0.14 < v \leq 2.27$	$2.27 < v \leq 4.67$	> 4.67
Acc_Y	≤ -6.02	$-6.02 < v \leq -3.57$	$-3.57 < v \leq -1.12$	$-1.12 < v \leq 1.33$	> 1.33
Speed_X	≤ 6.75	$6.75 < v \leq 13.50$	$13.50 < v \leq 20.25$	$20.25 < v \leq 26.70$	> 26.70
Speed_Y	≤ 6.10	$6.10 < v \leq 12.20$	$12.20 < v \leq 18.30$	$18.30 < v \leq 24.40$	> 24.40
Diff_Yaw	≤ -206.10	$-206.10 < v \leq -66.97$	$-66.97 < v \leq 72.17$	$72.17 < v \leq 211.30$	> 211.30

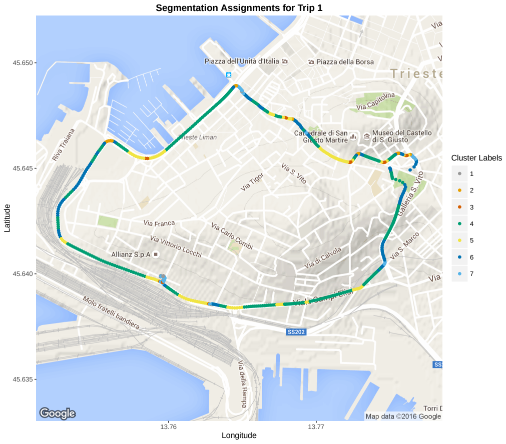


Figure 3: Geo-referenced representation of segmentation based on HMM for an example trip.

intervals are described in Table 7). Our goal is to obtain a soft clustering based on which similar trips can be identified analyzing topics distribution over each trip document.
For building the model we set T, top level truncation value, to 50; and K, second level truncation value, to 15.
In particular we retrieved two kinds of information: terms relevance in each identified topic (shown in Table 9), and the topic distribution over each document (see Table 8).

3.5 Validation and Discussion

So far our objective was to extract recurrent driving patterns from the trips in order to detect different behaviors along them. In this section we aim to categorize the whole trip using the information retrieved with the previous methods.

3.5.1 *Comparison of the Methods*

First, we compared the results of the different segmentation methods to understand how stable is the coherence of the obtained clusters across method. In order to do so we run k-Means clustering algorithm on the trips data obtained by the three methods. Each trip is characterized by its distribution of points among the identified clusters or topics. Using the Elbow method we set k equal to 6 for all the three methods.

Table 9: Terms relevance in top 7 extracted topics.

	Terms Distribution															
Topic	cdaac	bdaac	cdbbc	cdbac	cdabc	cccab	bdcdc	bdbbc	cdaad	cdaaa	ecabc	bdbac	cdcac	bdaac	bdaae	cdadc
Topic 1	0.315	0.111		0.095	0.048							0.047				
Topic 2	0.259	0.113	0.087		0.089							0.044				
Topic 3	0.201	0.077	0.040		0.032							0.029				
Topic 4	0.074		0.112	0.077				0.035					0.033			
Topic 5	0.101	0.023		0.031		0.030			0.026							
Topic 6	0.128	0.047		0.034						0.023					0.021	
Topic 7	0.032	0.026					0.024				0.024					0.022

Table 8: Topic relevance for observed driving trips.

	Topic Distribution
t1	(Topic 0, 0.0782), (Topic 1, 0.3130), (Topic 3, 0.6087)
t2	(Topic 0, 0.8662), (Topic 1, 0.1335)
t3	(Topic 0, 0.9751), (Topic 1, 0.0246)
t4	(Topic 1, 0.9989)
t5	(Topic 0, 0.7893), (Topic 1, 0.2104)
t6	(Topic 0, 0.8888), (Topic 1, 0.1109)
t7	(Topic 1, 0.9920)
t8	(Topic 1, 0.9992)
t9	(Topic 0, 0.6667), (Topic 1, 0.3330)
t10	(Topic 0, 0.6366), (Topic 1, 0.3631)
t11	(Topic 0, 0.9242), (Topic 1, 0.0755)
t12	(Topic 0, 0.6056), (Topic 1, 0.3941)
t13	(Topic 0, 0.7981), (Topic 1, 0.2016)
t14	(Topic 1, 0.9992)
t15	(Topic 0, 0.3015), (Topic 3, 0.0621), (Topic 6, 0.6358)
t16	(Topic 0, 0.6458), (Topic 1, 0.3539)
t17	(Topic 0, 0.9455), (Topic 1, 0.0543)
t18	(Topic 0, 0.7748), (Topic 1, 0.2249)
t19	(Topic 0, 0.9848), (Topic 1, 0.0149)
t20	(Topic 1, 0.9992)
t21	(Topic 0, 0.8960), (Topic 1, 0.1038)
t22	(Topic 0, 0.9142), (Topic 1, 0.0856)
t23	(Topic 1, 0.9988)
t24	(Topic 0, 0.9013), (Topic 1, 0.0985)
t25	(Topic 0, 0.8620), (Topic 1, 0.1377)
t26	(Topic 0, 0.7431), (Topic 1, 0.2568)
t27	(Topic 0, 0.8139), (Topic 1, 0.1859)

Thus we obtain trips clustered in 6 groups for each method (notice however that the 6 clusters may be different in the three methods).

At this point, we want to test the hypothesis that the trip clusters generated by the different methods are consistent. To check that, we need to identify the optimal mappings between clusters of the different methods. At this purpose, we build a set of confusion matrices for each pair of methods, where each cell contains the number of common elements between a cluster in a method and a cluster in the other. Since we don't know the optimal correspondence between clusters, we actually generate a combinatorial set of matrices, covering all the possible mappings (basically by changing the rows order according to all the possible combinations). We select the optimal mapping by considering the matrix with highest sum of elements on its main diagonal. This corresponds to the best configuration for the mapping and its value represent the number of elements that keeps the same grouping across the methods.

Considering DP-means and Hidden Markov Model, 74% of trips are grouped in the same way. If we consider the comparison between the aforementioned methods with the Topic Extraction method, the trips clustered in the same way is respectively only 44% and 48%, which probably account for the fact that, while DP-means and HMM are both clustering methods, Topics Extraction is a soft-clustering based on a totally different approach.

3.5.2 Ground-truth based validation

In order to validate our results, we asked a set of experts (knowledgeable about driving styles and driving paths recorded) to identify possible groups of trips in the dataset (i.e., considering factors such as signal distribution and driving routes). We were interested to see whether our method yield a classification coherent to the ones provided by the experts. The experts were able to highlight three clusters (shown in first two columns of Table 10). This accounts for a smaller set of clusters with respect to our analysis, but users weren't able to reach the level of details of six different categories of drivers.

The comparison with our clustering solutions therefore implied mapping our clusters the the human-generated ones and verifying their coherency. Table 10 reports the human-generated groundtruth in terms of clusters (first column: clusters A, B, C) and corresponding trips (second column). Then, the subsequent columns show, for every method, the best allocation of automatically calculated clusters (and corresponding trips), together with the number of wrongly assigned trips. For instance, experts assigned trips t1, t15 and t26 to Cluster B. For DP-means, the best correspondence is cluster C4, that contained only t1 and t15. Correspondingly, t26 results as one trip wrongly assigned.

The last row of the table shows the precision of each methods, computed as the ratio between the number of trips placed in the correct cluster and the total number of trips. Notice that we achieve the 96% of precision with all our approaches, thus demonstrate that the grouping defined by the proposed methods is coherent with the classification defined by the experts. The obtained results show the effectiveness of the proposed framework in profiling observed trips based on gathered information of vehicle status and driving behavior.

4. RELATED WORK

Driving behavior has been studied from different perspec-

Table 10: Comparison between the clusters identified by human and the results of our methods

Human-identified Clusters	**Groundtruth (trips in clusters)**	**DP**			**HMM**			**Topic**		
		Clusters	**Trips**	**Errors**	**Clusters**	**Trips**	**Errors**	**Clusters**	**Trips**	**Errors**
Cluster A	t2,t3,t5,t6,t9,t10,t11 t12,t13,t16,t17,t18, t19,t21,t24,t25,t27	C2,C6	t2,t3,t5,t6,t9,t10, t11,t12,t13,t16,t17, t18,t19,t21,t24,t25,t27 + t26	1	C1,C4,C5	t2,t3,t5,t6,t9,t10, t11,t12,t13,t16,t17, t18,t19,t21,t24,t25,t27 + t26	1	C1,C3,C4	t2,t3,t5,t6,t9,t10, t11,t12,t13,t16,t17, t18,t19,t21,t24,t25,t27 + t26	1
Cluster B	t4,t7,t8,t14,t20,t23	C1,C3,C5	t4,t7,t8,t14,t20,t23	0	C2,C3	t4,t7,t8,t14,t20,t23	0	C5	t4,t7,t8,t14,t20,t23	0
Cluster C	t1,t15,t26	C4	t1,t15	-1	C6	t1,t15	-1	C2,C6	t1,t15	-1
Precision		96%			96%			96%		

tives to reach different purposes such as: identification of driving behavior model [21], analysis of behavior variables impact in driving process, identification of driving maneuvers and driver's status, user classification based on driving behavior attitude and prediction of driver intentions.

Of particular importance are the studies regarding the identification of the driver's status. For instance works such as [22] and [13], use multiple sensors to provide intelligent information on the driver's physiological signals, which can include eye activity measures, the inclination of the driver's face, heart rate monitoring, skin electric potential, and electroencephalographic (EEG) activities. In [2] is proposed a novel and non-intrusive driver behaviour detection system using a context-aware system combined with in-vehicle sensors collecting information regarding to vehicle's speed, acceleration, the direction of driver's eyes, the position in lane and the level of alcohol in the driver's blood.

Another application of driver behavior modeling concern the motor insurance sector, that is interested in calculating their premiums based on statistical data through the evaluation of factors that are believed to impact expected cost of future claims. For instance [9] proposes to use of driver behaviour patterns and driving style classification to improve assessment of driver risk and insured risk using a smartphone as sensing platform.

Driver behavior modeling was also used to detect aggressive driving, a particular type of driving style, has long been studied due to its strong correlation with accidents and traffic safety hazards: by one estimate, it was influential in causing the majority of accidents in the United States from 2003 to 2007 [1]. In [18] has been proposed a theoretical framework in which acceleration behavior can be analyzed in order to detect aggressive behavior. The limits of these accelerations are related to the edges of the friction circle (also called ellipse of adherence), which depend on tire characteristics and road surface condition.

State of the art approaches in the attempt to model the driver behavior characteristics mainly employ models that are inspired by advanced neural network (NN), Hidden Markov Models (HMM), fuzzy control theory, Gaussian Mixture Models (GMM) and others models as stated by Meiring et. al. in [7]. Particular attention is paid to time series analysis used to mines behavioural data in order to achieve goals as driver profiling or maneuvers detection. Promising algorithms apply techniques belonging to a different scope, belonging to text processing and speech diarization with interesting and useful results. Takano et al. in [16] propose a hierarchical model with one HMM characterizing the short-term driving behaviors in the lower layer, and the other HMM characterizing the long term driving behaviors which are represented in the HMM space.This structure makes the vehicles intelligent by storing the knowledge of driving behaviors as the symbols of driving intention through observing the driving behavior given by expert drivers. Baum-Welch algorithm (a maximum likelihood estimation method) which trains parameters of HMMs is applied to optimize three HMMs driving straight, normal steering, and emergency steering [5]. In [12] Sathyanarayana et al. proposed a Driver Behavior Analysis and Route Recognition by Hidden Markov Models in two different approaches. The first (bottom-to-top) approach takes isolated maneuver recognition with model concatenation to construct a generic route, whereas the second (top-to-bottom) approach models the entire route as a phrase and refines the HMM to discover maneuvers. Only left turn (LT), right turn (RT) and lane change maneuvers are considered.
In [3] the authors present a new stochastic driver-behavior model based on Gaussian mixture model (GMM) framework. The proposed driver-behavior modeling is employed to anticipate car-following behavior in terms of pedal control operations in response to the observable driving signals, such as the own vehicle velocity and the following distance to the leading vehicle.

5. CONCLUSIONS

In this work the main goal was to propose a solution for driver behaviour modeling and driver profiling based on unsupervised methods. The main idea is to identify recurrent behaviours shared between drivers and characterize each driver according to the distribution of these behavioural patterns.

In order to achieve this goal we proposed three methods, which differ in terms of assumptions and implementation. However, they have in common the concept of identification of an underlying hidden structures: in clustering based segmentation this structure is identified in clusters grouping, using HMM is described by latent states and at last applying Hierarchical Dirichlet Processes behavioral topics have been extracted.

We applied these methods on a real world dataset and we compared the results between each other and with a ground truth built by experts from the car insurance industry. We found out that, even though the methods present relevant differences in clustering trips segment, they show high consistency in classifying whole trips.

For future activities, considering the encountered challenges and the critical tasks faced during this work, we propose to extend this analysis to a much larger collection of driving trips belonging to different drivers and different areas. In fact the experimental dataset used in our work represents just a small sample of a possible large scale data gathering and analysis process. The increase in dataset size can produce better results especially on the topic extraction process which requires a large collection of documents.

From a technical point of view we plan to relax our assumption of independence between observations in HMMs that can be too much restrictive. Regarding DP-means segmentation techniques and its dependence on data ordering we plan to investigate some reordering techniques to improve identification performance. The discretization phase in topic extraction process can be enhanced considering a variable number of bins for each of the considered features. For more precise evaluation, we plan to make use of camera recordings that can be useful in human labeling of Contextual Scenes.

6. REFERENCES

[1] Aggressive driving:research update. Technical report, AAA Foundation for Traffic Safety, 2009.

[2] S. Al-Sultan, A. H. Al-Bayatti, and H. Zedan. Context-aware driver behavior detection system in intelligent transportation systems. *IEEE Transactions on Vehicular Technology*, 62(9):4264–4275, Nov 2013.

[3] P. Angkititrakul, C. Miyajima, and K. Takeda. Modeling and adaptation of stochastic driver-behavior model with application to car following. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 814–819, June 2011.

[4] M. I. J. B.Kulis. Revising k-means: New algorithms via bayeian nonparametrics. In *Proceedings of the 29th international Conference on Machine Learning*, 2012.

[5] H. Lei, Z. Chang-fu, and W. Chang. Driving intention recognition and behaviour prediction based on a double-layer hidden markov model. *Journal of Zhejiang University SCIENCE C*, 13(3):208–217, 2012.

[6] N. Lin, C. Zong, M. Tomizuka, P. Song, Z. Zhang, and G. Li. An overview on study of identification of driver behavior characteristics for automotive control. *Mathematical Problems in Engineering*, 2014, 2014.

[7] M. G. A. Marthinus and M. H. Carel. A review of intelligent driving style analysis systems and related artificial intelligence algorithms. *Sensors*, 15(12):29822, 2015.

[8] C. Miyajima, Y. Nishiwaki, K. Ozawa, T. Wakita, K. Itou, K. Takeda, and F. Itakura. Driver modeling based on driving behavior and its evaluation in driver identification. *Proceedings of the IEEE*, 95(2):427–437, Feb 2007.

[9] J. Paefgen, F. Kehr, Y. Zhai, and F. Michahelles. Driving behavior analysis with smartphones: Insights from a controlled field study. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, MUM '12, pages 36:1–36:8, New York, NY, USA, 2012. ACM.

[10] Z. F. Quek and E. Ng. Driver identification by driving style. Technical report, technical report in CS 229 Project, Stanford university, 2013.

[11] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[12] A. Sathyanarayana, P. Boyraz, and J. H. L. Hansen. Driver behavior analysis and route recognition by hidden markov models. In *Vehicular Electronics and Safety, 2008. ICVES 2008. IEEE International Conference on*, pages 276–281, Sept 2008.

[13] S. Schneegass, B. Pfleging, N. Broy, F. Heinrich, and A. Schmidt. A data set of real world driving to assess driver workload. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '13, pages 150–157, New York, NY, USA, 2013. ACM.

[14] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[15] J. Straub, S. Zheng, and J. W. Fisher. Bayesian nonparametric modeling of driver behavior. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 932–938, June 2014.

[16] W. Takano, A. Matsushita, K. Iwao, and Y. Nakamura. Recognition of human driving behaviors based on stochastic symbolization of time series signal. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 167–172, Sept 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] R. Vaiana, T. Iuele, V. Astarita, M. Caruso, A. Tassitani, C. Zaffino, and V. Giofre. Driving behavior and traffic safety: An acceleration-based safety evaluation procedure for smartphones. *Modern Applied Science*, 8(1), 2014.

[19] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967.

[20] C. Wang, J. W. Paisley, and D. M. Blei. Online variational inference for the hierarchical dirichlet process. In *AISTATS*, volume 2, page 4, 2011.

[21] W. Wang, J. Xi, and H. Chen. Modeling and recognizing driver behavior based on driving data: A survey. *Mathematical Problems in Engineering*, 2014.

[22] L. Wei, S. C. Mukhopadhyay, R. Jidin, and C.-P. Chen. Multi-source information fusion for drowsy driving detection based on wireless sensor networks. In *Sensing Technology (ICST), 2013 Seventh International Conference on*, pages 850–857, Dec 2013.