A Large Scale Examination of Vehicle Recorder Data to Understand Relationship between Drivers' Behaviors and Their Past Driving Histories

Daisaku Yokoyama and Masashi Toyoda
The University of Tokyo
Meguro-ku Komaba 4-6-1, Tokyo, Japan, 153-8505
Email: {yokoyama, toyoda}@tkl.iis.u-tokyo.ac.jp

***Abstract*—We propose an analysis method of driving behaviors based on large-scale and long-term vehicle recorder data to support fleet driver management by classifying drivers by their skill, safety, physical/mental fatigue, aggressiveness, and so on. Previous studies rely on precise data with small number of drivers, which are difficult to extrapolate to general drivers. In this study, we examine ability of a dataset that is sparse but large-scale (over 100 fleet drivers) and long-term (10 months' worth). We focus on classifying drivers recently involved in accidents, and examine correlation with driving behaviors. We propose two models for the classification; entropy-like model and KL divergence model that aim to emphasize the behavioral difference from average drivers. From experiments, we will show some informative findings on behaviors that might cause accidents.**

I. Introduction

We propose a method for analyzing relationships between vehicle drivers' properties and their driving behaviors based on large-scale and long-term vehicle recorder data. Our purpose is to support fleet driver management by classifying drivers by their skill, safety, physical/mental fatigue, aggressiveness, and so on. There have been several studies [1][2][3] that analyze usual driving behaviors. They, however, rely on detailed and precise data of small number of drivers, and it is difficult to extrapolate the results to general drivers. Recently many transport companies started to introduce dashcams or vehicle data recorder systems that keep track of GPS trajectories, velocity, and acceleration. Due to the limitation of storage size, the data tend to be sparse, but it can be collected from large number of drivers.

Our method classifies drivers using long-term records of their driving operations (braking, wheeling, etc.) with several attributes (max speed, acceleration, etc.). It is based on the assumption that the distributions of these attributes would differ driver to driver by their properties. In this work, we focus on classifying drivers recently involved in accidents, and examine correlation with driving behaviors. The results might be useful for education and for preventing further accidents.

Our main contributions are:

- We intensively examine a large-scale vehicle recorder data and confirm the effectiveness of our method to analyze drivers' properties.
- We design two methods to characterize driver's behaviors for classifying drivers; entropy-like model, and KL divergence model. We apply machine learning technique to select effective features and successfully find some informative ones.

II. Related Work

There are many researches that try to utilize vehicles' recorded data, such as velocity or location, for many purposes[4][5]. Several researches are focusing on the driving behavior and utilize them to understand or classify the drivers' characteristics. Ly et al. [6] shows that two similar drivers behavior can be classified using the records of vehicle's internal sensors from the CAN bus.

Some researches try to classify drivers by the aggressiveness of their driving behavior, which aims to achieve driving safety in future. Higgs et al. [1] analyze three drivers' behavior during car-following period, and show the differences among them. Dang et al. [2] focus on the lane change characteristics in highway, and reveal that some properties such as lane change frequency are different among 12 drivers. Castignani et al. [3] utilize the sensors embedded in current smartphones to score the driving aggressiveness among five drivers. These researches try to distinguish the behaviors that are assumed to be clearly different. Our research tries to find an uncertain difference among drivers. The dataset size of our work is larger than in the existing researches.

III. Classification of drivers

A. Dataset

We use large scale real driving records which are collected by Sagawa Express Co., Ltd., which is one of the largest transportation companies in Japan providing a door-to-door delivery service, in cooperation with Datatec Co., Ltd. It consists of about 1400 drivers who assigned to Tokyo prefecture, worth about 10 months (from 21 July 2014 to 4 June 2015). Data is recorded by a multifunctional vehicle recorder, developed by Datatec Co., Ltd., [1] that has longitudinal accelerometer, lateral accelerometer, gyro compass, and GPS.

Figure 1 shows some statistics of the drivers in the dataset. The upper histogram shows the distribution of driving days for

[1] "SRVideo", http://www.datatec.co.jp/seiftyrecorder/srcomm.html (in Japanese)

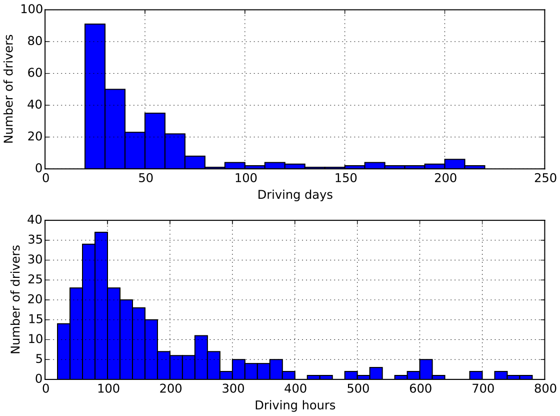


Fig. 1. Drivers' statistics for whole dataset

TABLE I
OPERATION RECORD STATISTICS

operation type	# of records per driver (min, max)	# of records (total)
braking	(114, 41320)	1624248
wheeling	(239, 40943)	2255959
turning	(121, 19147)	989253
stopping	(337, 36257)	1770696

each drivers, and the lower shows the driving hours for each drivers. The driving hour does not contain the time when a car engine is turned off. Drivers who work less than 20 days or drive less than 20 hours in total are eliminated from the data.

The vehicle data recorder automatically detects four basic driving operations; braking, wheeling, turning, and stopping operation. Several statuses such as maximum speed or acceleration during the operation are recorded. Table I shows the statistics of recorded operation for each drivers.

With the help of transport company, we can access to the drivers' history about accidents from 10 years before, 1st January 2004. We split drivers into two groups, accident drivers and no accident drivers, whose driving experience is at least 5 years worth. There are 38 drivers who had at least one accident within 98 drivers.

B. Classification by machine learning technique

We design several features using frequency of the digitized properties, and try to evaluate their ability to characterize drivers by using Support Vector Machine (SVM) with Gaussian kernel [7]. We first select 6 basic properties which are related to velocity to separate operation records, and combine other properties such as acceleration, jerk (the derivative of acceleration with respect to time), yaw velocity, etc. to make features. Basic properties are binned into several intervals which have 5km/h width. Other properties are binned into 10 intervals; the maximum and minimum bins' breakpoints are chosen by hand, and other bins are defined with the same interval width. The number of bins are 540 in total. The operation frequency of each bins are accumulated for every drivers, thus every driver has 540 values of frequency N_i^f, where f means the property combination and i means the bin id. We then normalize N to compute each driver's occurrence probability P, and also compute the average probability Q for all drivers.

TABLE II
PERFORMANCE COMPARISON (LABELED BY 5 YEARS HISTORY)

method	c	γ	precision	recall	f-measure
probability	2^{11}	2^{-11}	0.66	0.58	0.62
entropy-like	2^9	2^{-7}	0.69	0.56	0.62
KL divergence	2^1	2^{-3}	0.62	0.58	0.60
no information			0.37		

probability method Use P itself.

entropy-like method The value Q represents the distribution of the average driver. We consider that one occurrence have $-\log(Q)$ information, and each drivers' distribution can be represented as $-P\log(Q)$. This kind of technique often be used in the anomaly detection, because it emphasize the occurrence of rare case.
This indicator ignores whether P is greater than Q or not, thus we introduce the value $sign$, which equals to 1 when $P > Q$, or -1 otherwise. Finally we use the value $-P\log(Q) \times sign$ as the feature.

KL divergence method We try to describe the difference between two distributions, P and Q. KL divergence [8] is one of the representative definition for the distance between two distributions, as: $\sum_i P_i \log(P_i/Q_i)$

C. Experiments

We evaluate the performance of these features by 10-fold cross validation. Meta parameters of the Gaussian kernel (c, γ) are searched to achieve the best f-measure. We apply feature selection method [9] and use top 8 features to evaluate each method. The selected features are expected to have larger information of the drivers' accident history. Experimental results are shown in Table II. The baseline precision by random classification is also shown as the *no information* case.

We change the bias of SVM and investigate the relationship between precision and recall of these methods. Due to the page limitation, we only show the result of the entropy-like method. The red line in Figure 2 shows the precision (Y-axis) and recall (X-axis) relationship using the entropy-like method. The baseline precision is $36/98 = 0.37$ (described as the horizontal line in Figure 2), which can be achieved by the random classification. The precision continuously grows when the recall becomes small. The feature in this method seems to have some information to classify accident history. F-measures are also plotted as the green line.

1) Discussion of the classification experiments: According to the f-measure, the probability method achieves comparable performance with other methods. However the precision-recall relationship shows that its features' does not seems to be informative. Other two proposed methods shows promising results. Our dataset is larger than the one in other researches, but 36 accident drivers are not enough to discuss the accuracy.

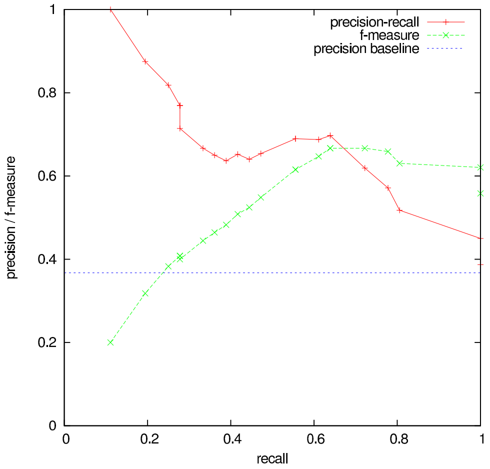
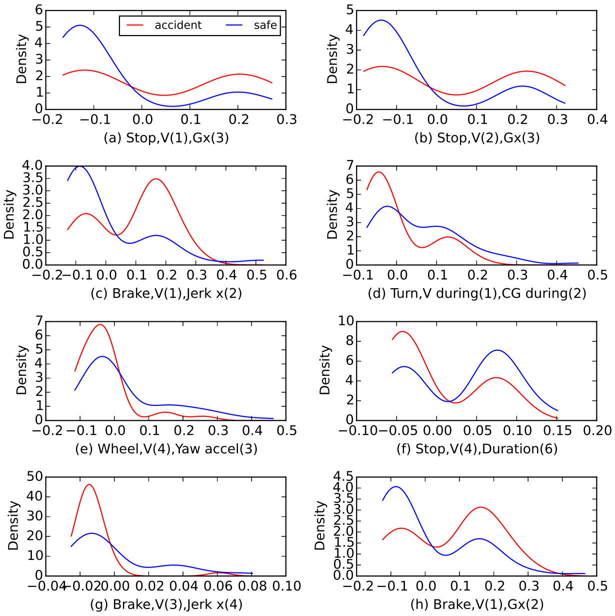


Fig. 2. Classification result of $-p\log(q) \times sign$ method



(a) Stop,V(1),Gx(3)

(b) Stop,V(2),Gx(3)

(c) Brake,V(1),Jerk x(2)

(d) Turn,V during(1),CG during(2)

(e) Wheel,V(4),Yaw accel(3)

(f) Stop,V(4),Duration(6)

(g) Brake,V(3),Jerk x(4)

(h) Brake,V(1),Gx(2)

Fig. 3. Selected features of $-p\log(q) \times sign$ method

2) Selected informative features: Figure 3 shows the selected features and its value distribution of two driver groups using entropy-like method. Each panel represents the value density distribution of one feature, estimated by KDE. Red and blue lines indicates two driver groups. Features are represented as (operation type, basic property (bin ids), property (bin ids)). Larger bin id represents the bin that has larger value of the property. For example, "(Stop, V(1), Gx(3))" means the bin with the lowest velocity and 3rd lowest Gx during stopping operation. These features are expected to have some relationship with the driver's accident history.

In Figure 3, (a), (b), (c), (f) and (h) features have different shapes within two groups. Different peak positions in these features will help to separate drivers, therefore these features seems to be informative for considering drivers' safety.

IV. Conclusion

We try to understand drivers' properties by examining a large scale log of vehicle data recorder. Experiments show that our methods successfully find some informative driving operation behaviors that might cause accidents.

This is the first step to understand the relationship between safety driving and driver's behavior. This work only discuss about the past accidents, however the acquired knowledge will help to investigate the driver's safety and prevent future accidents. Our approach might be helpful to analyze other properties such as driver's skill, fatigues, and so on.

Our approach that is focusing on the differences of driving operations works well. The frequencies at rare bins are small, and the operation will not occur in short term; several days, for example. This means that daily review of vehicle recorder data may not have the ability to distinguish an unsafe behavior. We plan to keep archiving the vehicle data recorder log to see long-term behavior of drivers. We also plan to combine other information such as geo-location of the operation or weather.

Acknowledgment

We thank Sagawa Express Co., Ltd. and Datatec Co., Ltd. for providing and collecting precious vehicle recorder data. This work was supported by the "Vehicle Big Data Platform for Driver-Centric Services", National Institute of Information and Communications Technology, Japan.

References

[1] B. Higgs and M. Abbas, "A two-step segmentation algorithm for behavioral clustering of naturalistic driving styles," in *Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on*, Oct 2013, pp. 857–862.

[2] R. Dang, F. Zhang, J. Wang, S. Yi, and K. Li, "Analysis of chinese driver's lane change characteristic based on real vehicle tests in highway," in *Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on*, Oct 2013, pp. 1917–1922.

[3] G. Castignani, R. Frank, and T. Engel, "Driver behavior profiling using smartphones," in *Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on*, Oct 2013, pp. 552–557.

[4] D. Johnson and M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, Oct 2011, pp. 1609–1615.

[5] W. Wu, W. S. Ng, S. Krishnaswamy, and A. Sinha, "To taxi or not to taxi? - enabling personalised and real-time transportation decisions for mobile users," in *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, July 2012, pp. 320–323.

[6] M. V. Ly, S. Martin, and M. Trivedi, "Driver classification and driving style recognition using inertial sensors," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*, June 2013, pp. 1040–1045.

[7] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92. New York, NY, USA: ACM, 1992, pp. 144–152. [Online]. Available: http://doi.acm.org/10.1145/130385.130401

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 03 1951. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177729694

[9] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, ser. Studies in Fuzziness and Soft Computing, I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, Eds. Springer Berlin Heidelberg, 2006, vol. 207, pp. 315–324. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-35488-8_13