A Machine Learning Framework for Road Safety of Smart Public Transportation

Shengda Luo[1] , **Alex Po Leung**[1*] , **Xingzhao Qiu**[1]
[1]Macau University of Science and Technology, Taipa, Macau, China
pleung@must.edu.mo

Abstract

To monitor road safety, billions of records can be generated by Controller Area Network bus each day on public transportation. Automation to determine whether certain driving behaviour of drivers on public transportation can be considered safe on the road using aritificial intelligence or machine learning techniques for big data analytics has become a possibility recently. Due to the high false classification rates of the current methods, our goal is to build a practical and accurate method for road safety predictions that automatically determine if the driving behaviour is safe on public transportation. In this paper, our main contributions include 1) a novel feature extraction method because of the lack of informative features in raw CAN bus data, 2) a novel boosting method for driving behaviour classification (safe or unsafe) to combine advantages of various statistical learning methods with much improved performance, and 3) an evaluation of our method using a real-world data to provide accurate labels from domain experts in the public transportation industry for the first time. The experiments show that the proposed boosting method with our proposed features outperforms other six popular methods on the real-world dataset by more than 5.9% and 5.5%.

1 Introduction

1.1 Motivations

Traffic accidents not only bring huge financial losses to society but also cause great physical and mental damages to citizens [1]. Millions of people die from traffic accidents in 2018 [2], and most traffic accidents are caused by human mishandling [3]. Analyzing the behaviour of drivers (especially public transportation drivers) is important to protect the road safety of citizens [4–6]. For public transportation driver management and monitoring, massive data is collected from the driving vehicles using state-of-the-art technologies of sensors (e.g. MobilEye from Israel). In the public transportation control center, thousands of real-time events and alarms are produced from the sensors in vehicles every day. Although it is difficult to handle the huge amount of data manually, accurate predictions with machine learning techniques to analyze behaviour with massive data collection from the vehicles become feasible. Machine learning techniques have been applied to analyzing behaviour in different tasks with various kinds of data collected using sensors in moving vehicles [7,8]. The industry needs efficient and accurate machine learning methods to classify whether the driving behaviour of public transportation drivers is safe, and the drivers with unsafe behaviour will have to be re-trained.

1.2 Challenges

There are three challenges to classify whether certain driving behaviour of drivers on public transportation can be considered safe on the road using machine learning techniques. First, the industrial need for a high classification performance cannot be satisfied using existing methods using computer vision as the false classification rates are too high with existing methods. Due to the high false classification rates, it makes it hard to compare to evaluate the performance of drivers on public transportation. Second, the lack of features in the standard Controller Area Network (CAN) bus data does not provide a lot of information for driving behaviour analysis to train an accurate machine learning classifier. There is no existing method for road safety predictions with CAN bus data to extract extra useful information from features. Last, because of the high cost of labeling, there is no public data with labels for the evaluation of the machine learning models to predict whether certain driving behaviour of drivers on public transportation is safe. The lack of urgently needed labels in datasets for road safety makes the evaluations of machine learning models impossible to achieve.

1.3 Contributions

In this paper, our contributions include 1) we proposed a method to compute extra time-series features to extract richer information, 2) we propose a boosting method to make various statistical learning complementing each other, 3) we evaluate methods using a new real-world dataset with labels from experts in the public transportation domain.

There are very few available features in the data collected using the standard CAN bus system. With the lack of sufficient useful features, it is hard to find patterns in the data

*Corresponding Author

to determine driving behaviour from the driver. Feature engineering is very an important tool to extract useful information in time series for machine learning methods to get better performance in this case. We, thus, propose a method to compute extra time-series features from the raw data of the CAN bus system to extract extra information.

In order to obtain the best possible performance, in this paper, we propose a boosting method for classifying whether the driving behaviour of drivers on public transportation is safe. To combine the advantages of various statistical learning methods, we use boosting to make the methods complementing each other. We consider that the ensemble of methods can outperform the particular method, and our boosting method combines six state-of-the-art methods: support vector machine (SVM), random forest (RF), k-nearest neighbour (KNN), discriminant analysis, naive Bayes classifier, and adaptive boosting (AdaBoost). It is shown that the proposed boosting method outperforms the other six methods regardless of whether or not the proposed feature extraction method is used.

Because of the high cost of sample labeling, there is no publish real-world dataset with labels for analysis driver behaviour of drivers on public transportation. To completely evaluate methods in the real world, the experiments are conducted using a real-world dataset collected using the CAN bus system. The samples in the dataset are labeled by the experts of Transportes Urbanos de Macau (TransMac).

2 Related Work

In this work, we focus on analyzing the data been made available by CAN bus systems. With growing data collected using CAN bus systems, machine learning plays an important role in building analytics models to handle the massive data. Methods based on statistical learning are successfully used to solve many related behaviour analytics problems. In [8, 9], Bayesian learning techniques are used to predict braking behaviour and model speed of drivers. KNN is employed to classify driving styles in [7]. SVM, and decision trees (DT) are applied to predict driving behaviour and accident risk prediction [10, 11]. These existing methods are all designed using one particular machine learning technique. We argue that the ensemble of machine learning methods can outperform, most of the time, one particular technique, and we propose a heterogeneous boosting method to obtain better performance. In our experiments, our boosting method is compared with six state-of-the-art methods (see Section 1.3).

3 Our Dataset to Evaluate Road Safety

For any application domain of machine learning, one of the most objective evaluation methods is to see how prediction models perform in real-world datasets. However, to the best of our knowledge, there is no published real-world dataset with labels provided for behaviour analysis of public drivers. In this work, we build a new dataset collected using the CAN bus system by one of the biggest public bus companies in Macao called TransMac.

Every three seconds, one record is produced from one sensor in a moving public vehicle, and there are totally 6451 records with 24 features in the new dataset. All 6451 records are labeled by the operators in Transmac. In total, the recording time for our CAN bus data is $6451 \times 3 = 19353$ seconds long which respond to 5.38 hours of driving by professional bus drivers in the company. Although each sample contains 24 features, some features cannot be used to train the machine learning model (see Table 1). As shown in Table 1, features recorded the information related to the identifier are useless for training machine learning methods. Some features contain too many missing data also cannot be used to train. For example, most entries of the 'CANALARMSTATE' feature and the 'CANALARMSTATE' are N/A (Not Available). For further research, Table 1 is listed to provide a reference for whether features used to train our boosting method.

Feature Name	Meaning	Used to train our method?
LOGID	Bus identifier	No
GPSDATE TIME	Time	No
VELOCITY	Instantaneous speed	Yes
MILEAGE	GPS mileage	Yes
TOTAL MILEAGE	Total mileage	Yes
FRONT PRESSURE	Front pressure	Yes
REAR PRESSURE	Rear pressure	Yes
ENGINESPEED	Engine speed	Yes
ENGINETEMP	Engine temperature	Yes
CARSWITCH	Switches of the bus	No
CARLIGHT STATE	Switches of light	No
CANALARM STATE	Switches of alarm	No
CREATETIME	Time	No
GPS VELOCITY	Instantaneous speed	Yes
DRIVERID	Driver identifier	No
LONGITUDE	Longitude	Yes
LATITUDE	Latitude	Yes
DIRECTION	Turn	Yes
STATIONID	Station identifier	No
ROUTEID	Route identifier	No
BUSSTATE	Bus status	No
ALARM STATE	Alarm light status	No
STATION MILEAGE	Mileage each station	Yes
UPDOWN	Up and down	No

Table 1: Features in our dataset to evaluate road safety.

4 The Proposed Methods

We propose a feature extraction method (see Algorithm 1) for extracting richer information from the change of feature vectors against time, and propose a boosting method (see Algorithm 2) to classify whether driving behaviour of drivers on

public transportation can be considered safe on the road. The feature extraction method is a general method. It can be used with any other machine learning classification method. In the experiments, it is shown that our feature extraction method can be used to improve the performance of any classification method. In addition, it is also shown that our boosting method outperforms other six machine learning methods (see Section 1.3) whether or not our feature extraction method is used.

4.1 Our Method for Richer Information With Feature Extraction

Missing data is common in industrial data collected from CAN data systems. Features with too many N/A (Not Available) entries cannot be used to train the machine learning methods. Also, features irrelevant for driving behaviour analysis like the identifier of the vehicles are excluded. Therefore, there are only a few useful features left for classification without irrlevant features (see Table 1). The low dimensionality of the feature space of training data severely limits the descriptive power of the samples. The lack of descriptive power makes it is very difficult to obtain accurate machine learning models. We argue that richer information can be extracted from the change of feature values against time and we, hence, propose a feature extraction method to provide extra useful time-series features to deal with the lack of information in the original features. For example, the acceleration of the car is important for driver behaviour analysis, but this information is not recorded in the original data. The acceleration of the bus can be obtained by calculating the gradient of the velocity of the bus. The proposed feature extraction method is shown in Algorithm 1.

Algorithm 1 Our Method for Richer Information With Feature Extraction

Input: n samples $\{s_1, ..., s_n\}$ where $s_i = [f_{i,1}, ..., f_{i,m}]^T$
Output: n samples $\{s_1, ..., s_n\}$ with time-series features, $s_i = [f_{i,1}, ..., f_{i,m}, t_{i,1}, ..., t_{i,m+7}]^T$

Divide n samples into p periods, $\{P_1, ..., P_p\}$, by the recording time.
for $j = 1, ..., p$ **do**
  for each sample $s_i \in P_j$ **do**
    $[t_{i,1}, ..., t_{i,m}]^T = \frac{1}{|P_j|} \times \sum_{s_z \in P_j} [f_{z,1}, ..., f_{z,m}]^T$
    $s_i = [f_{i,1}, ..., f_{i,m}, t_{i,1}, ..., t_{i,m}]^T$
  end for
end for
for each sample s_i **do**
  $t_{i,m+1}$ and $t_{i,m+2}$ are the differences in the feature values of s_i and s_{i-1}, for the latitude and the longitude respectively.
  $\{t_{i,m+3}, ..., t_{i,m+7}\}$ are the gradients of the feature values of s_i related to the velocity, the mileage, the tire pressure, the engine speed and the engine temperture.
  $s_i = [f_{i,1}, ..., f_{i,m}, t_{i,1}, ..., t_{i,m+7}]^T$
end for

The input data of Algorithm 1 contains n samples, $\{s_1, ..., s_n\}$, with m features $s_i = [f_{i,1}, ..., f_{i,m}]^T$. Especially, the features irrelevant for training are excluded in these m features. For example, for our proposed dataset, the m features are the twelve features used to train the classification model (see Table 1). It is common in time-series analysis to use moving averages. For example, the average driving speed of a driver in two minutes (a period) is useful information for analyzing his driving behaviour. Motivated by this, some time-series features are calculated for this particular reason (see Algorithm 1). The n samples are divided into p periods by the recording time. The value of p is a tunable parameter, and it depends on the time interval between two samples. In our boosting method, one period covers two minutes. For sample s_i, m time-series features, $\{t_{i,1}, ..., t_{i,m}\}$, are extracted from raw features. The latitude and longitude features, $f_{latitude}$ and $f_{longitude}$, are obtained from GPS information. The difference in the latitude/longitude values of the adjacent samples can be used to measure the velocity of the bus. In Algorithm 1, $t_{i,m+1}$ and $t_{i,m+2}$ of sample s_i are the differences in the values of sample s_i and sample s_{i-1}, for latitude and longitude features respectively. The gradient of a feature is used to describe how fast the feature values change. The velocity, the mileage, the tire pressure, the engine speed and the engine temperture are important for accurate classification. These features can reflect the different behaviour of drivers, and $\{t_{m+3}, ..., t_{m+7}\}$ are calculated to find the rates of change.

4.2 Our Boosting Method

Ensemble learning is a machine learning technique, which is used to combine multiple models and to get better performance than that of a particular model. The proposed boosting method combines six state-of-the-art machine learning methods. The six methods are SVM [12], KNN [13], R-F [14], navie Bayes [15], discriminant analysis [16] and AdaBoost [17]. The proposed boosting method is in Algorithm 2.

Algorithm 2 The Proposed Boosting Method

Input: Traning data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $x_i \in \mathbf{R}^m$, $y_i = 0, 1$ for safety and unsafety samples respectively.
Output: Final strong classifier $H(x)$.

Initialize weights $w_{1,i} = \frac{1}{2c}, \frac{1}{2(n-c)}$ for safety samples and unsafety samples, respectively, where c is the number of safety samples.
for t = 1 to U **do**
  For each sample, s_i, normalize its weight, $w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$.
  Train all g weak classifiers, $\{h_1(x), ..., h_g(x)\}$, using the training data D with time-series features.
  Predition using g classifiers, and choose the classifier, $h_u(x)$, with the highest correctly rate a.
  Updata the weights: $w_{t+1,i} = w_{t,i} \times B_u^{1-e_i}$, where $e_i = 0$, if the sample s_i is correctly predicted, otherwise $e_i = 1$.$B_u = \frac{a}{1-a}$.
end for
The boosting classifier combines the U classifiers: if $\sum_{t=1}^{U} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{U} \alpha_t$, $H(x) = 1$, otherwise $H(x) = 0$.

As shown in Algorithm 2, there are n samples in the training data, and the dimensionality of them is equal to m. There are g weak learners used in the algorithm. In the proposed

boosting method, g is equal to six. U is the number of the weak classifiers which are chosen to form final strong classifier $H(x)$. The value of U is a tunable parameter, and it is equal to five in our method. Each of g classifiers is trained based on one particular machine learning method.

5 Experiments

Given that there is no published real-world dataset with labels, the experiments are conducted using our own data which is collected and labeled with the help of the experts at TransMac. In the experiments, the proposed boosting method is compared with other six popular machine learning methods: SVM, KNN, RF, National Bayes, Discriminant Analysis, and AdaBoost. We use MATLAB to implement these six methods and the proposed methods. The classification accuracy is obtained by using

$$Accuracy = 100 \times \frac{e}{c}, \tag{1}$$

where e is the number of test samples which get the correct prediction, and c is the number of all samples in test set. To clearly show the comparison for the accuracy (see Equation 1), the experiments are divided into two parts. In the first part, in order to demonstrate that the proposed boosting method can outperform other machine learning methods, all methods are trained using the raw data without using our feature extraction method (see Algorithm 1). In the second part, to determine whether our feature extraction method can be used to improve the performance of machine learning methods, methods are trained using the data with the time-series features provided by our feature extraction method. By comparing the accuracies of these two experiment parts, it is shown that the performance of methods is improved using our feature extraction method. To avoid the overfitting issue of machine learning, in both parts of the experiments, there are two scales, 70% of the dataset and 90% of the dataset, of the training set. The samples in the training set are randomly extracted from the whole dataset.

The performance comparison of methods are listed in Table 2 and Table 3. As shown in the two tables, our boosting method outperforms the other methods in any case. It shows that the performance of any method is improved using our features extraction method. Using our boosting method with our feature extraction method can outperform other methods without our feature extraction method by more than 5.9% and 5.5%, with 70% and 90% of the whole dataset randomly selected for training respectively.

Classifying Our Dataset Without Our Features		
Methods	70% percent data for training	90% percent data for training
Our Method	**91.04**%	**92.86**%
AdaBoost	78.06%	80.12%
Simple Bayes	60.62%	62.01%
Discriminant Analysis	58.44%	60.04%
KNN	75.09%	75.19%
RF	89.70%	91.16%
SVM	57.67%	61.08%

Table 2: The comparison among seven methods on our real-world dataset without the proposed feature extraction method. It is shown that our boosting method outperforms the other state-of-the-art methods by more than 1.7% and 1.5%, with 70% and 90% of the whole dataset randomly selected for training respectively.

Classifying the Our Dataset With Our Features		
Methods	70% percent data for training	90% percent data for training
Our Method	**95.60**%	**96.74**%
AdaBoost	83.72%	85.27%
Simple Bayes	60.41%	66.82%
Discriminant Analysis	77.82%	78.81%
KNN	77.67%	78.02%
RF	94.17%	94.26%
SVM	77.93%	77.20%

Table 3: The comparison among seven methods on our real-world dataset with the proposed feature extraction method. After comparing with the accuracies in Table 2, it shows that the performance of any method is improved using our feature extraction method.

6 Conclusion

Automation to determine whether certain driving behaviour of drivers on public transportation can be considered safe on the road using A. I. or machine learning techniques has become a possibility recently. However, the industrial need for a high classification performance cannot be satisfied using existing methods using computer vision as the false classification rates are too high with existing methods. Due to the high false classification rates, it makes it hard to compare to evaluate the performance of drivers on public transportation.

Our goal is to build a practical and accurate method for road safety predictions that automatically determine if the driving behaviour is safe on public transportation. In this paper, our main contributions include 1) a novel feature extraction method because of the lack of informative features in the data, 2) a novel boosting method for driving behaviour classification (safety or not) to combine advantages of various statistical learning methods with much improved performance, and 3) evaluating methods using real-world data to provide accurate evaluations from labels from experts in the public transportation industry for the first time. The experiments show that the proposed boosting method with the proposed features outperforms other six popular methods on the real-world dataset by more than 5.9% and 5.5%.

Acknowledgement

We would like to express our gratitude to Transportes Urbanos de Macau (Transmac) for the help of data collection. This work is supported by the Faculty Research Grant (No. FRG-18-020-FI) at Macau University of Science and Technology and funding from the Hong Kong Applied Science and Technology Research Institute (No. OSO-18-002-FI).

References

[1] J. Tison, N. Chaudhary, L. Cosgrove, P. R. Group *et al.*, "National phone survey on distracted driving attitudes and behaviors," United States. National Highway Traffic Safety Administration, Tech. Rep., 2011.

[2] W. H. Organization *et al.*, *Global status report on road safety 2018*. World Health Organization, 2018.

[3] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, 2015.

[4] D. Hallac, A. Sharang, R. Stahlmann, A. Lamprecht, M. Huber, M. Roehder, J. Leskovec *et al.*, "Driver identification using automobile sensor data from a single turn," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016, pp. 953–958.

[5] O. Moll, A. Zalewski, S. Pillai, S. Madden, M. Stonebraker, and V. Gadepally, "Exploring big volume sensor data with vroom," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1973–1976, 2017.

[6] E. Massaro, C. Ahn, C. Ratti, P. Santi, R. Stahlmann, A. Lamprecht, M. Roehder, and M. Huber, "The car as an ambient sensing platform [point of view]," *Proceedings of the IEEE*, vol. 105, no. 1, pp. 3–7, 2016.

[7] V. Vaitkus, P. Lengvenis, and G. Žylius, "Driving style classification using long-term accelerometer information," in *2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR)*. IEEE, 2014, pp. 641–644.

[8] A. Mudgal, S. Hallmark, A. Carriquiry, and K. Gkritza, "Driving behavior at a roundabout: A hierarchical bayesian regression analysis," *Transportation research part D: transport and environment*, vol. 26, pp. 20–26, 2014.

[9] J. C. McCall and M. M. Trivedi, "Driver behavior and situation aware brake assistance for intelligent vehicles," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 374–387, 2007.

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

[11] J. Paefgen, T. Staake, and F. Thiesse, "Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach," *Decision Support Systems*, vol. 56, pp. 192–201, 2013.

[12] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011.

[13] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] J. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of naive bayes classifiers (pdf)." ICML, 2003.

[16] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2006.

[17] P. Viola, M. Jones *et al.*, "Rapid object detection using a boosted cascade of simple features," *CVPR (1)*, vol. 1, pp. 511–518, 2001.