Research on short-time Prediction of driving behaviour at unsignalized intersections based on machine learning

Yunfan Zhang1a, Xuedong Yan1b*

1 School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044, China

a18120957@bjtu.edu.cn
bxdyan@bjtu.edu.cn

*Corresponding author’s e-mail: bxdyan@bjtu.edu.cn

Abstract. The unsignalized intersections are the important nodes in the road network, and also the sections with a high incidence of serious collision accidents. Based on the driving simulation experiment, the interactive driving behaviours of multiple drivers were collected. The data were input to the machine learning model, training for predicting the short-time driver’s decision-making. The results showed that the Random Forest model has a better prediction effect than the Support Vector Machine (SVM) model, and the shorter prediction interval also has a better prediction effect. This study was helpful to give timely warning to drivers in danger and provided a new idea for the research of collision prevention warning strategies and control methods.

1. Introduction
The traffic situation is extremely complicated at intersections. Due to the lack of traffic control facilities, there are more traffic conflicts at unsignalized intersections. The traffic statistics showed that there were about 40% traffic accidents occurring at the intersections, which 28% were unsignalized intersections accidents and 12% were signalized intersections accidents[1]. When the vehicle approaches the unsignalized intersection, drivers dynamically decide to accelerate or decelerate[2]. This game process is fraught with insecurity. If the driving behaviour and decisions of drivers could be predicted accurately, then the driver could be alerted in advance, and the insecurity would be reduced.

The use of machine learning to categorize driving behaviour from raw data was not uncommon. Naiwala P.et.al[3] presented a machine learning approach to classifying the skill at maneuvering by drivers’ behaviour data. On that basis, data-driven approaches with advances in machine learning techniques were utilized to predict driving behaviour. Pongtep Angkititrakul et.al[4] employed a Bayesian framework to calculate the probability of a driver decelerating. Linsen Chong et.al[5] proposed a rule-based neural network model to simulate driver behaviour in terms of longitudinal and lateral actions in the car-following situation and safety-critical events. Shun Yang et.al[6] investigates the contribution levels of 13 features of driver's behaviour prediction at unsignalized intersection using a Random Forest model.

Compared with synchronous recognition, the short-time prediction model of driving behaviour state predicts the driving behaviour state of the next period by inputting relevant data of the current period, to obtain the driver's driving status information earlier than the classification model of driving behaviour state. Therefore, when the state of dangerous driving is predicted, it is helpful for the
intelligent driving assistance system to warn the driver in advance, and ensure driving safety. In this paper, driving simulation equipment would be used to collect driving behaviour data at unsignalized intersections. On this basis, the Random Forest and the Support Vector Machine (SVM) models, two kinds of classification models would be trained to make short-time prediction drivers’ decision-making behaviour.

2. Experiment and data collection

2.1. Apparatus
The equipment used in this experiment was the Multi-User Driving Simulation (Hereinafter referred to as: MDS), which was developed by Beijing Huiyitiancheng Digital Technology Co., Ltd. MDS system was a driving simulation system with high integration, excellent performance and realistic effect. It could support driving simulation experiments on various urban traffic environments under laboratory control according to different experimental purposes, and conduct corresponding experimental results. The system consisted of ten single driving simulators, which were connected by a server, as shown in Figure 1. This experiment used two of these linked driving simulators. Each single simulator was equipped with a driver seat, an automatic gearbox, a gas pedal, a brake pedal and three screens to project the front-view driving scenarios with a 120-degree view etc.

![Figure 1. A part of driving simulators.](image)

2.2. Scenarios design
The scene where the two straight-moving vehicles would encounter in an unsignalized intersection was designed. The road network is shown in Figure 2.
Considering the adequacy data, three consecutive unsignalized intersections were set up. The scenario was designed as a bidirectional straight rural road with a speed limit of 40 km/h, and the total length of the road network was 3000 m (the blue vehicle gone along the blue line, and the red vehicle gone along the red line) as shown in Figure 2. There was no traffic flow in the whole scene, and there was no obstacle blocking the driver's view in the scene. Each driver was required to stop at 250m in front of each intersection for the encounter of the two vehicles at the intersection, and then two subjects in each group would start driving according to the instruction from the corresponding experimenter for each subject.

2.3. Subjects and experiment procedure

48 subjects were participated in this experiment. The experiment intentionally recruited drivers who were the same gender (all of them were male), and age (average age was 24, ranging from 22 to 30 years old) for the purpose of eliminating the interference of the internal differences of the subjects. Every subject held a valid driver’s license in China, and the driving experience of each subject was between 2 and 4 years. Subjects were divided into pairs of 24 groups and enter the same scene of encountering at the unsignalized intersection. The experiment lasted about 60 min in total, and each subject received 300 Chinese RMB (about 51 U.S. dollars) after the experiment was completed. Subjects should be trained before the experiment, including attempting to change speed, turn, start and stop, and other driving skills. They were then advised to drive and behave as they normally would and to adhere to traffic laws as they would in real-life situations. And subjects were also notified that during the experiments if they had strong dizziness and other physical discomforts, the driver could quit the experiment at any time. Two subjects cannot communicate with each other during the whole process. After completing the driving tasks, the subjects were organized to fill in the questionnaire.

2.4. Data preprocessing and labeling

During the experiment, the simulator data were sampled at 30 Hz. The data within 200m from the intersection are selected as the analysis objects. A total of 653 valid sample data were collected in the experiment, and all the data were recorded in time series. As the distance between the vehicle and the intersection dwindled, when two drivers found themselves in conflict with each other's vehicles, they accelerated or decelerated in response. Some drivers also chose to travel at a uniform speed. Literature showed that the average prediction accuracy of the model was relatively high when the prediction interval was 1.0s, and research showed that if drivers could get the risk warning 1.0s before
the collision risk, about 90% of the accidents could be avoided[7]. Therefore, considering the accuracy and real-time requirements of the prediction model, t = 1.0s were finally chosen as the prediction interval of the model. In order to study whether a shorter time interval had a better prediction effect, another 0.5s time interval (t=0.5s) was introduced as a comparison. The relevant variables of driving behaviour in each time interval of t were extracted in the form of the average value. The interpretation of variables is shown in Table 1. According to the average acceleration of each time period t, the period of every sample was labeled by 0 for no response (driving at a constant speed), 1 for acceleration, and 2 for deceleration, as shown in Table 2.

Besides, in the whole process of the vehicle approaching the intersection, the decision (acceleration, deceleration, no response) of the vehicle and the average acceleration (In the whole game process of vehicle driving, the average acceleration from the starting point of speed change to the ending point) were extracted for the purpose of describing the overall driving process.

### Table 1. Interpretation of variables

| Variable types                      | Definitions of behavioral measures                        | Notation |
|-------------------------------------|----------------------------------------------------------|----------|
| Variables related to the ego vehicle| Speed(km/h)                                              | S        |
|                                     | Acceleration(m/s²)                                       | A        |
|                                     | Distance between the vehicle and intersection(m)         | D        |
| Variables related to the other vehicle| Difference of two vehicles’ distance between the vehicle and intersection(m) | DD       |
|                                     | Difference of two vehicles’ speed (km/h)                 | SD       |
|                                     | Difference of two vehicles’ TTC (s)                      | TD       |
|                                     | Direction of the other vehicle (left or right)           | DIR      |
|                                     | The speed of the other vehicle(km/h)                     | OS       |
|                                     | Acceleration of the other vehicle(m/s²)                  | OA       |
|                                     | Distance between the other vehicle and intersection(s)   | OD       |

### Table 2. Conditions for labeling

| Acceleration                      | label |
|-----------------------------------|-------|
| Acceleration <-0.1m/s²            | 2     |
| Acceleration >0.1m/s²             | 1     |
| -0.1 m/s²< Acceleration<0.1m/s²   | 0     |

### 3. Methodology

These prediction methods aim to divide the driving behaviour state (acceleration deceleration and no respond) of the next period by inputting the relevant driving behaviour data (From Table 1) of the current period.

#### 3.1. The Random Forest model

The Random Forest model combines Brieman’s bagging idea[8] and Ho’s Random subspace methods[9] to construct a collection of decision trees with controlled variations[10]. It has been widely used in the research field of driving behaviour. Rachel Bycroft et.al captured the driver's intention of other vehicles using a Random-Forest classifier and prevent motion predictions that were too conservative[12]. Random Forests model was also used to identify whether the driver was driving distractedly, and the accuracy was highest[13]. Cheng[11] proposes a robust Random Forest method to analyze travel mode choices for examining the prediction capability and model interpretability.
Furthermore, the Random Forest model could investigate the contribution levels of different features of human driving behaviour[6].

![Random Forest Model Diagram](image1)

**Figure 3. The Random Forest model simplified**

The Random Forest model is an algorithm that integrates multiple trees. Its basic unit is the decision tree, while its essence belongs to a large branch of Machine Learning, the Ensemble Learning method. It could build a multitude of weak decision tree classifiers in parallel and then combine them to form a single, strong learner by averaging/voting their individual predictions, as illustrated in Figure 3. A training set for growing trees is Randomly selected from a sample set, and the remaining samples termed an out-of-bag (OOB) set, are used to estimate the Random Forest model's goodness-of-fit. Then, trees are grown to the maximum extent possible without pruning.

Instead of using a generic performance index, an out-of-bag (OOB) error was introduced to represent a Random Forest error. During the training, about one-third of the data was excluded from the bootstrap sample and was not used to reconstruct the decision tree. These data were called OOB samples and were used for model testing. The estimated error on these OOB samples is the OOB error. Figure 4 shows the OOB error as the number of decision trees (n_estimators). From Figure 5, when the number of trees is 60, the error is about 0.24. Finally, we choose a tree number of 100, so that the error converges well.

![OOB Error Graph](image2)

**Figure 4. The OOB classification error of the Random Forest model**

The importance of each characteristic variable was output through the Random Forest, as shown in Figure 5. The importance of DIR was lowest. Therefore, DIR was eliminated during the training of the Random Forest model.

![Variable Importance Graph](image3)
3.2. The Support Vector Machine (SVM) model

In machine learning, SVM is a supervised learning algorithm. D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_m, y_m)\}, y_i \in \{1, -1\} is a training sample set. The core of the algorithm is to find a partition hyperplane in the sample space based on training set D and separate samples of different categories. The partition hyperplane can be expressed as \( \vec{w} \cdot \vec{x} - b = 0 \). Where, \( \vec{w} \) is the normal vector that determines the direction of the hyperplane, and b is the displacement that determines the distance between the hyperplane and the origin. When the training sample set is linearly separable, two parallel hyperplanes \( \vec{w} \cdot \vec{x} - b = \mp 1 \), that separate the two classes of data can be selected. The sample data on the two parallel hyperplanes is called the support vector, and the sum of the distance between the two heterogeneous support vectors and the hyperplane is \( \frac{2}{\|\vec{w}\|} \), which is called the "interval". In order to maximize the distance between the two planes, \( \frac{1}{2} \|\vec{w}\|^2 \) should be maximized, that is, \( \frac{1}{2} \|\vec{w}\|^2 \) should be minimized. Meanwhile, the data points of the sample data should be outside the interval area of the hyperplane. Therefore, the objective function and constraint conditions of SVM algorithm are shown as follows:

\[
\begin{align*}
\min & \quad \frac{1}{2} \|\vec{w}\|^2 \\
\text{s. t.} & \quad y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1, i = 1, 2, \ldots, m
\end{align*}
\]

3.3. Evaluation indicators of the trained model

The ROC's full name is the Receiver Operating Characteristic curve. According to the prediction results of the learner, the threshold was changed from 0 to the maximum, that is, in the beginning, each sample was used as a positive example for prediction. With the increase of the threshold, the learner predicted fewer and fewer positive samples until no samples were positive samples at last. In this process, the ROC curve was obtained by calculating the values of two important quantities each time and plotting them as horizontal and vertical coordinates respectively. The ROC curve has the True Positive Rate (TPR) on the vertical axis and the False Positive Rate (FPR) on the horizontal axis. As shown in Figure 6, the ROC curve can easily detect the influence of the arbitrary threshold on the generalization performance of the learner, which is helpful to select the best threshold. The closer the ROC curve is to the upper left corner, the higher the model's recall. The point on the ROC curve closest to the upper left corner is the best threshold with the fewest classification errors and the fewest total number of false positives and false negatives. This method is simple, intuitive, the accuracy of the analysis method can be observed by the diagram, and can be judged by the naked eye. ROC curve combines true case rate and false-positive case rate with the graphical method, which can accurately reflect the relationship between true case rate and false-positive case rate of a certain learning device and is a comprehensive representation of detection accuracy.
4. Results and analysis

4.1. Analysis of whole driving process

The average acceleration and decision during the whole driving process of the vehicle were analyzed. Figure 7 illustrates that the value of the average acceleration was mainly distributed between -5 m/s\(^2\) and 5 m/s\(^2\), which is a reasonable range.

From Table 3, the decision of the ego vehicle and the acceleration pattern of the other vehicle were proved to be highly correlated (Value=34.997, Sig=0.000). That verified the reliability and rationality of the experiment and. When the other vehicle acceleration, the ego vehicle had a high probability of deceleration. On the contrary, when the other vehicle deceleration, the ego vehicle was more likely to accelerate.

| Decision of the ego vehicle | Acceleration | Deceleration | No response | Value | Sig. |
|-----------------------------|--------------|--------------|-------------|-------|------|
| Acceleration pattern of the other vehicle | 24.0% | 50.0% | 26.0% | 34.997 | 0.000 |
| Deceleration | 51.0% | 29.7% | 19.3% |
| No response | 37.9% | 42.1% | 20.0% |
4.2. Evaluation of the trained Random Forest model
The above has introduced the meaning of the ROC curve. Figure 8 shows the ROC curve of the prediction result of the Random Forest model. When $t=0.5s$, AUC areas are 0.93, 0.94, 0.94, when classifications are 0, 1, 2, respectively, and the micro-average is 0.94. When $t=1s$, AUC areas are 0.88, 0.89, 0.87, when classifications are 0, 1, 2, respectively, and the micro-average is 0.88. The predictions were more accurate when $t=0.5s$.

![ROC curve for Random Forest model at $t=0.5s$](image)

![ROC curve for Random Forest model at $t=1.0s$](image)

Figure 8. The ROC curve and AUC area of RF model

4.3. Evaluation of the trained SVM model
The above has introduced the meaning of the ROC curve. Figure 9 shows the ROC curve of the prediction result of the Random Forest model. When $t=0.5s$, AUC areas are 0.83, 0.86, 0.85, when classifications are 0, 1, 2, respectively, and the micro-average is 0.80. When $t=1s$, AUC areas are 0.76, 0.74, 0.73, when classifications are 0, 1, 2, respectively, and the micro-average is 0.71. The predictions were more accurate when $t=0.5s$. 
From Table 4, the F1-score of the SVM model was relatively low, and the model's predictions were not accurate.

| class | Precision t=1.0s | Precision t=0.5s | Recall t=1.0s | Recall t=0.5s | F1-score t=1.0s | F1-score t=0.5s |
|-------|------------------|------------------|---------------|---------------|----------------|----------------|
| 0     | 0.64             | 0.76             | 0.40          | 0.57          | 0.49           | 0.65           |
| 1     | 0.45             | 0.52             | 0.89          | 0.91          | 0.60           | 0.66           |
| 2     | 0.64             | 0.82             | 0.13          | 0.30          | 0.22           | 0.44           |

By comparing the ROC curves of the two models, the following conclusions can be drawn:

(1) ROC curves of the prediction results of the Random Forest model are closer to the upper left corner, compared with the SVM model. It indicates that the prediction results of the Random Forest model are more accurate and the classification effect is better.

(2) ROC curves of the prediction results when t=0.5s are closer to the upper left corner, compared with t=1.0s. It means that more accurate prediction results could be obtained by selecting the time interval of 0.5s.

5. Conclusion

This paper investigated the predicting effect on the driver’s decision-making of two machine learning methods: the Random Forest model and the SVM model. From ROC results, the Random Forest model had a higher accuracy (94%) than the SVM model, and shorter time intervals were more
advantageous. On the whole, the method of machine learning was feasible and accurate to predict the driver’s decision-making at unsignalized intersections. The variables related to the prediction mainly included acceleration and speed of the ego vehicle in its current state, the distance from the other vehicle to the intersection, and the acceleration of the other vehicle. The limitation of this study is that the experimental scene was monotonous. However, this research can provide ideas and theoretical support for the design of an intersection collision warning system.

Acknowledgments
This work is financially supported by National Key R&D Program of China (2019YFF0301403).

Reference
[1] Z. Wu, J. Yang, and L. Huang. (2013) Study on the Collision Avoidance Strategy at Unsignalized Intersection based on PreScan Simulation. Procedia - Social and Behavioral Sciences, vol. 96, pp. 1315–1321.
[2] M. Liu, G. Lu, Y. Wang, and Z. Zhang. (2014) Analyzing drivers’ crossing decisions at unsignalized intersections in China. Transportation Research Part F: Traffic Psychology and Behaviour, vol. 24, pp. 244–255.
[3] Chandrasiri, N.P., Nawa, K., Ishii, A., 2016. Driving skill classification in curve driving scenes using machine learning. J. Mod. Transp. 24, 196–206.
[4] Angkititrakul, P., Miyajima, C., Takeda, K., 2012. Analysis and prediction of deceleration behavior during car following using stochastic driver-behavior model, in: 2012 15th International IEEE Conference on Intelligent Transportation Systems. Presented at the 2012 15th International IEEE Conference on Intelligent Transportation Systems - (ITSC 2012), IEEE, Anchorage, AK, USA, pp. 1221–1226.
[5] Chong, L., Abbas, M.M., Medina Flintsch, A., Higgs, B., 2013. A rule-based neural network approach to model driver naturalistic behavior in traffic. Transp. Res. Part C Emerg. Technol. 32, 207–223.
[6] Yang, S., Wang, W., Jiang, Y., Wu, J., Zhang, S., Deng, W., 2019. What contributes to driving behavior prediction at unsignalized intersections? Transp. Res. Part C Emerg. Technol. 108, 100–114.
[7] EINEL H. (1998)Automotive millimeterwave radar history and present status [C]. IEEE European Microwave Conference, 1998: 619 – 629.
[8] Breiman, L. (1996). Bagging predictors. Mach. Learn. 24, 123–140.
[9] Ho, T.K. (1995). Random decision Forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE, pp. 278–282.
[10] Washington, S.P., Karlaftis, M.G., Mannering, F. (2010). Statistical and Econometric Methods for Transportation Data Analysis. CRC Press.
[11] Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a Random Forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10.
[12] Driver Intention-based Vehicle Threat Assessment using Random Forests and Particle Filtering, 2017. IFAC-Pap. 50, 13860–13865.
[13] Osman, O.A., Hajj, M., Karbalaei, S., Ishak, S., 2019. A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. Accid. Anal. Prev. 123, 274–281.