Construction of Injury Prediction Models for Car Occupants using Gradient-Boosting Decision Trees

Keita Takahashi[1)] Yusuke Miyazaki[1)] Koji Kitamura[2)] Fusako Sato[3)]

1) Tokyo Institute of Technology, School of Engineering
2-12-1 Oookayama, Meguro, Tokyo, 152-8550, Japan (E-mail: ???)
2) National Institute of Advanced Industrial Science and Technology
Central 1, 1-1-1 Baien, Tsukuba, Ibaraki, 305-8568, Japan
3) Japan Automobile Research Institute
2530 Karima, Ibaraki, 305-0822, Japan

KEY WORDS: Safety, Injury prediction, Accident analysis, Machine learning, Gradient-boosting decision tree [C1]

While developing automated driving systems, it is necessary to be able to quantitatively evaluate their effectiveness in reducing injuries. To achieve this, the following methods have been proposed to achieve this: First, the probability density distribution of the severity of occupant injuries is calculated for each accident scenario, and the relationship between various crash conditions, such as the crash speed and injury rate, is clarified for each scenario. Next, using accident reproduction simulations, the ability to avoid accidents and the quantity of crash speed reductions are estimated by replacing the accident vehicle in each scenario with a vehicle equipped with an autonomous driving system. Finally, the results of the above two steps are synthesized to estimate the reduction in the frequency of injuries as a whole when automated vehicles are widely used. When employing such a method, elucidating the relationship between the crash conditions and injury rates for each accident scenario is directly related to the reliability of the evaluation method.

Injury prediction models employing logistic regression have been proposed in an attempt to predict occupant injuries from occupant and crash condition information. However, logistic regression models have the problem that it cannot express the nonlinear relationship between explanatory variables and objective variables. In addition, in recent years, the traffic accident database used for the construction of the model may not be able to cope with the rapidly changing traffic conditions.

In this study, we constructed injury prediction models which are more current and have a better prediction performance in predicting the maximum MAIS3+ in vehicles. To implement these models, we carried out three new attempts: First, we changed the acquisition period of the data used for training and validation from 2010 to 2019. The data from 2010 to 2015 were obtained from the National Automotive Sampling System – Crashworthiness System (NASS-CDS), and data from 2016 to 2019 were obtained from the Crash Investigation Sampling System (CISS). Second, we constructed the injury prediction models with LightGBM, one of Gradient-Boosting Decision Tree models. Third, we included some new explanatory variables which can influence the injury risk, such as the crash scenario, body mass index, weight ratio of colliding and collided vehicles, vehicle damage extension. To separately observe the influence of the machine learning method and variable selection on the prediction accuracy, we constructed two injury prediction models, one using the same variables of URGENCY algorithm (Malliaris et al., 1997) and the other using our own selected variables.

For accuracy test, we performed classifications on the prepared data set using a total of three models, URGENCY, and the two models constructed in this study. We calculated the ROC-AUC scores. As a result, the ROC-AUC score for URGENCY was 0.8644, that of the model with the same variables was 0.8445, and that of the model using the selected variables was 0.8967, which performed the best among the three models. This result showed that the LightGBM model using properly selected variables exhibited better performance than URGENCY model. The advantage of the LightGBM model with selected variables is shown by its ability to use explanatory variables which have non-linear relationships with the objective variable.

The results of the gain importance of the LightGBM model with the selected variables (in Figure 1) shows that some variables such as BMI, weight ratio of vehicles, and weight of vehicles, are more important than the variables used in the URGENCY model, such as PDOF and seat belt usage. The more important variables should therefore be incorporated into future injury prediction models.

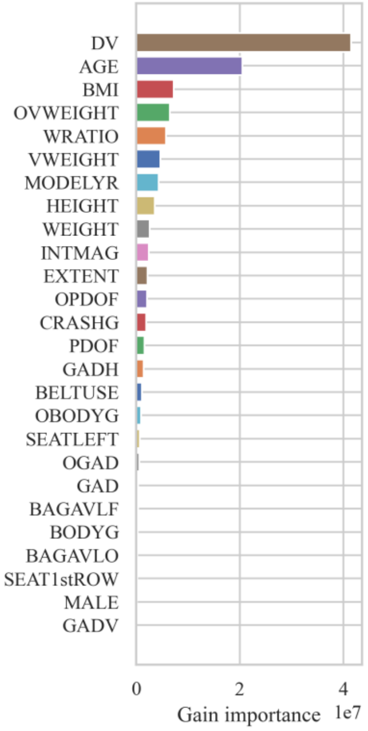

Fig. 1 Gain importance of the model with own selected variables