Accident Analysis and Prevention 120 (2018) 250–261

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data

Heejin Jeong[a,b,*], Youngchan Jang[b], Patrick J. Bowman[a], Neda Masoud[c]

[a] University of Michigan Transportation Research Institute, 2901 Baxter Road, Ann Arbor, MI, 48109, USA
[b] Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109, USA
[c] Department of Civil and Environmental Engineering, University of Michigan, 2350 Hayward Street, Ann Arbor, MI 48109, USA

ARTICLE INFO

Keywords:
Vehicle crashes
Injury severity classification
Imbalanced data
Machine learning
Data analytics
Automated vehicle safety

ABSTRACT

This study aims to classify the injury severity in motor-vehicle crashes with both high accuracy and sensitivity rates. The dataset used in this study contains 297,113 vehicle crashes, obtained from the Michigan Traffic Crash Facts (MTCF) dataset, from 2016–2017. Similar to any other crash dataset, different accident severity classes are not equally represented in MTCF. To account for the imbalanced classes, several techniques have been used, including under-sampling and over-sampling. Using five classification learning models (i.e., Logistic regression, Decision tree, Neural network, Gradient boosting model, and Naïve Bayes classifier), we classify the levels of injury severity and attempt to improve the classification performance by two training-testing methods including Bootstrap aggregation (or bagging) and majority voting. Furthermore, due to the imbalance present in the dataset, we use the geometric mean (G-mean) to evaluate the classification performance. We show that the classification performance is the highest when bagging is used with decision trees, with over-sampling treatment for imbalanced data. The effect of treatments for the imbalanced data is maximized when under-sampling is combined with bagging. In addition to the original five classes of injury severity in the MTCF dataset, we consider two additional classification problems, one with two classes and the other with three classes, to (1) investigate the impact of the number of classes on the performance of classification models, and (2) enable comparing our results with the literature.

1. Introduction

According to the National Highway Traffic Safety Administration Research Note, motor-vehicle crashes in 2015 led to 35,092 fatalities in the United States, an increase of 7.2% from 2014 (National Center for Statistics and Analysis, 2015). A crash analysis report published in the U.S. Department of Transportation indicates that the total economic cost of motor vehicle crashes occurred in 2010 was more than $200 billion, including the costs due to approximately 33,000 fatalities, 3.9 million non-fatal injuries, and 24 million damaged vehicles (Blincoe et al., 2015). About 31 percent of the total economic costs were in form of property damage costs, while 10 percent of the total costs were in medical costs. Based on these statistics, a better understanding of the relationship between crash risk factors and the injury severity can help enhance driving safety, curb the economic impact of crashes, and reduce the number of fatal crashes.

In classifying crash injury severity, the classification power of a model cannot be simply captured by its correct classification rate. Accident severity datasets are typically imbalanced, with the non-fatal class containing disproportionally more data points compared to the fatal class. If untreated, such a structure could lead to training models that look promising on the outside with high accuracy rates (the accuracy is defined as the ability of the model to correctly predict accident severity classes on a test set; see Eq. (1)), but fail to be informative in reality. An extreme example of a weak model is a trivial model that predicts all accidents to be non-fatal, in a 2-class problem with fatal and non-fatal classes. Such a model would have a very high accuracy rate, while the value of a crash classification model lies mostly on correct classifications of higher severity classes (e.g., fatal crashes), typically referred to as “sensitivity” (i.e., the ability of the model to correctly classify the severity level as ‘fatal’ (Farchi et al., 2007; Parikh et al., 2008; see Eq. (2)). On the other hand, a model that classifies all accidents as fatal would produce a high sensitivity, but a low accuracy score. Hence, there is a clear trade-off between accuracy and sensitivity scores of crash severity models that can only be resolved through appropriate handling of imbalanced data. This imbalanced data structure,

* Corresponding author.
E-mail address: heejinj@umich.edu (H. Jeong).

https://doi.org/10.1016/j.aap.2018.08.025
Received 14 December 2017; Received in revised form 16 August 2018; Accepted 26 August 2018

Table 1
Confusion Matrix and the Four Measurements for 2-Class Classification.

	Predicted Positive (fatal)	Predicted Negative (non-fatal)
Actual Positive (fatal)	True Positive (TP)	False Negative (FN)
Actual Negative (non-fatal)	False Positive (FP)	True Negative (TN)

therefore, necessitates additional steps in model training and evaluation: (*i*) using appropriate evaluation metrics, and (*ii*) balancing the dataset before training.

Limitations of classification accuracy rate in evaluating model performance could be addressed through using additional statistical measures, namely, true positive, true negative, false positive, and false negative (see Table 1 for a detailed description) to create more informative metrics. Using these measurements, accuracy, sensitivity, and specificity, as defined in Eqs. (1)–(3), respectively, can be easily computed for a 2-class classification problem. These metrics collectively help depict a more comprehensive picture of the overall model performance (Parikh et al., 2008). Ultimately, geometric mean (or G-mean) of sensitivity and specificity can be used as a compact evaluation metric to compare the general performance of different models. The G-mean is calculated as the square root of the product of sensitivity and specificity and will have high values when both sensitivity and specificity are high and the difference between the two metrics is small (Kubat et al., 1997). Finally, while reporting a variety of metrics that can provide a comprehensive picture of model performance is necessary, measures need to be taken to produce high-performance models in the first place. This can be obtained by generating a balanced dataset based on the original imbalanced dataset on which models can be trained.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{3}$$

Using a different terminology, the G-mean in binary classification can be the square root of the product of class 1's accuracy (former sensitivity measure) and class 2's accuracy (former specificity measure). As such, in a C-class classification problem the definition of G-mean can be expanded to C classes as the following:

$$\text{G-mean} = (\textit{Class 1 accuracy} \times \textit{Class 2 accuracy} \times \cdots \times \textit{Class C accuracy})^{1/C}$$

Traditionally, motor vehicle crash injury severity has been modeled using statistical methods, with the goal of identifying the significance of each potential factor in the severity of the outcome of a crash. This type of analysis is valuable in informing future safety-focused planning efforts. Although the proposed methodology here can be used for the same purposes through conducting sensitivity analysis, our main goal is to learn an ensemble of models that can predict the severity of crashes in a fraction of a second and with a high degree of accuracy. Such a model is needed by autonomous vehicles, where fast evaluation of the circumstances and decision making is critical. An autonomous vehicle, having access to such prediction, can take the necessary cautionary measures to avoid (or mitigate the impact of) any potential accidents.

Although autonomous vehicles are equipped with a variety of sensors to help them browse through the surrounding environment, they are highly susceptible to low-quality sensor readings as well as anomalous sensor readings, either due to faulty sensors or malicious cyber-attacks. As such, redundancy in information plays a great role in increasing the reliability of autonomous vehicles under various scenarios, and in the absence of (reliable) sensor readings. With focus on factors that are easy to measure (e.g., whether the driver is under the influence of alcohol using in-vehicle cameras), and can be reliably obtained from other sources (e.g., the roadway alignment using high definition maps, the weather conditions), we provide a prior model on the severity of potential accidents, should they occur. This information can help the autonomous system make better choices, e.g., take over the control of the motor vehicle from the human driver in case the driver is perceived to be under the influence (in semi-autonomous vehicles), or drive at lower speeds in the presence of adverse weather conditions, such as fog, which may reduce the precision of sensors.

The contributions of this paper to the literature are three-fold. First, we use a total of 5 machine learning techniques to model crash injury severity levels. These models are trained in isolation, and as ensemble models, providing insights on the degree to which various machine learning techniques are appropriate for crash injury severity classification. Second, we use under-sampling and over-sampling to treat the inherent imbalance of the crash dataset before learning and discuss the effects of these treatments. Finally, we provide various performance statistics and show that several of our models out-perform models in the literature by achieving both high accuracy and sensitivity rates at the same time.

2. Literature review

To date, many previous studies have modeled the traffic crash injury severity with potential risk factors, using statistical and machine learning methods (e.g., Abdelwahab and Abdel-Aty, 2001; Chang and Wang, 2006; Eluru et al., 2008; Zhu and Srinivasan, 2011; Castro et al., 2013; Xu et al., 2013; Yu and Abdel-Aty, 2013; Lee and Li, 2015; Chen et al., 2015, 2016). For example, Abdelwahab and Abdel-Aty (2001) used two neural network models (namely, multilayer perceptron and fuzzy adaptive resonance theory) to classify driver injury severity with driver, vehicle, roadway, and environmental factors. Chang and Wang (2006) estimated the effect of several risk factors (e.g., driver/vehicle, highway/environmental variables) on injury severity (i.e., fatal, injury, and no-injury) using classification and regression tree (CART). They analyzed crash data from police records collected in Taiwan and found that vehicle type is the most important factor associated with injury severity. Chen et al. (2015) used a hybrid method that combines multinomial logit and Bayesian network to classify the driver injury severity, with crash data from New Mexico. They identified several risk factors for motor vehicle crash fatalities, including environmental factors such as windy weather and inferior lighting conditions.

Among the studies on classifying motor vehicle crash injury severity in the literature (e.g., Abdelwahab and Abdel-Aty, 2001; Chang and Wang, 2006; Eluru et al., 2008; Castro et al., 2013; Chang and Chien, 2013; Xu et al., 2013; Yu and Abdel-Aty, 2013; Lee and Li, 2015; Chen et al., 2016), some report only the classification accuracy (e.g., Abdelwahab and Abdel-Aty, 2001). Other studies report both the classification accuracy and the classification accuracy of the class of interest by reporting statistics such as sensitivity or specificity, but only the classification accuracy has been the focus of discussion (e.g., Chang and Chien, 2013; Chang and Wang, 2006). Table 2 summarizes studies in the literature that report more statistics than just the general classification accuracy. Moreover, none of these studies have taken measures to address data imbalance, although two studies (i.e., Chang and Wang, 2006; Chen et al., 2015) have pointed out this issue as a limitation of their work.

3. Data description

Crash data used for fatality analysis were obtained from the Michigan Traffic Crash Facts (MTCF) database that contains official Michigan year-end crash data (Office of Highway Safety Planning, 2017). In this study, two years of crash data (restricted to vehicle crashes only; neither pedestrian nor bicycle) were collected (from 2016 to