24 May 2021 Report

Brad Burkman

23 May 2021

Contents

1 Jargon to Understand

- Naturalistic Driving Data - Data collected from sensors installed in the driver's own car, trying to get as close as possible to the driver's "natural" behavior.
- Heterogeneity

2 IRB, SHRP Database

Eleven of the papers in *Accident Analysis and Prevention* used the Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS), which put sensors in 3400 cars and recorded five million trips, including crashes. To get "Qualified Researcher Status" with "full access to data that has been made available through the SHRP 2 NDS Data Access Website," I had to submit a certificate of training on research with human subjects. I did the training through the UL Institutional Review Board (IRB). I now have access.

3 NGSIM Database

Three papers us the Next Generation Simulation dataset from the US Dept of Transportation, and it's available for download with no restrictions.

4 Ideas for a Paper

4.1 Comparing Image Transformation Techniques

In "A deep learning based traffic crash severity prediction framework" by Rahim (LSU), they transformed each data point into an image, and used the images as input for a Convolutional Neural Network (CNN) for classification. There are several ways to make such a transformation. Write a paper where I compare different image transformation methods, and their effect on different metrics (precision, recall, accuracy, sensitivity, f1, false alarm rate) of the classification of the test set.

4.2 Comparing Data Cleaning Techniques

In "A deep learning based traffic crash severity prediction framework" by Rahim (LSU), they just deleted any records with missing or inconsistent data. The *Titanic* Kaggle sites you showed me use several other methods for filling in incomplete data. Write a paper where I compare different methods for dealing with missing data, and their effect on different metrics (precision, recall, accuracy, sensitivity, f1, false alarm rate) of the classification of the test set.

Rahim's article took out 37% of the records for missing or inconsistent data, but only 21% of the fatal crashes; could that imbalance in the data cleaning skew the model prediction? It makes sense that police would be more meticulous in their record keeping for fatal crashes, but 21% and 37% are huge.

Would we get a better model if we found a good way to fill in missing data?

4.3 Missing Data Analysis

This isn't a ML paper, but somebody should write it. In what situations is the crash data most likely to be incomplete or inconsistent? Rather than a, "How can we fix the roads to make them safer?," this would be a "How can we get better records, so we can know how to fix the roads to make them safer?"

4.4 Fill Holes in Osman et al

"A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data."

A group at LSU has a 2019 paper using the SHRP-2 data to predict, based on the driving behavior whether the driver was making a phone call, texting, or having a conversation with a

passenger, but "the effect of roadway and geometric characteristics is not within the scope of this paper." They say that in the introduction and devote the entire last paragraph of the conclusion to laying out that hole that needs to be filled.

The data is not just crashes, but any time a person was using a phone, texting, or talking.