Look before you Leap :

Leveraging Predictive Models to Improve Automotive Safety and Travel Time

Isaac Lapides, Hannah Do, Kamil Sachryn

ML Fall 2020

Problem Description

Recent growing amount of vehicles and traffic has been slowing down intensive road users, which costs over 10% of the logistic drivers' working hours.

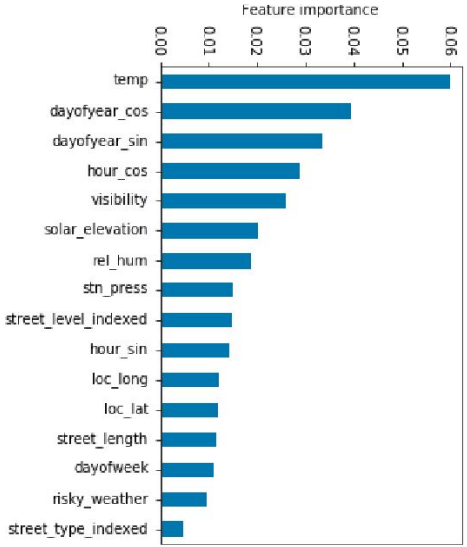
Previous attempts of predicting traffic incidents has been difficult since human-based reports have high labor cost with significant delays. And many automatic incident detection algorithms that have been developed have difficulties in reaching certain accuracy due to unexpected circumstances and its expensive cost.

Finding an optimal algorithm to predict the traffic would be useful in avoiding the congestion and optimizing the time cost for many people in daily lives.

State of the art

"High-Resolution Road Vehicle Collision Prediction for the City of Montreal" (Antoine Hébert et al., 2019)

Predicted risk of accidents with high spatiotemporal resolution over a large area and time, allowing for useful identification of significant risk factors. This extended the state of the art by dynamically combining three large datasets, doing feature engineering for imbalanced classes, and implementing BRF in Spark.



"From Twitter to detector: Real-time traffic incident detection using social media data" (Gu et al., 2016)

Processed real-time tweets in order to retrieve possible traffic accident areas via NLP methods

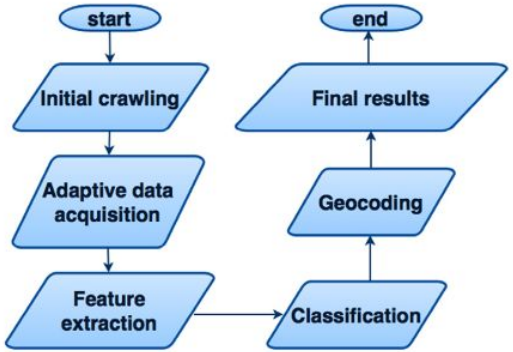


Fig. 1. The work flow of Twitter data acquisition, processing and analytics.

The methodology was applied in two regions, Pittsburgh and Philadelphia Metropolitan Areas, and the discrepancy between public Twitter accounts and individual accounts was measured.

Approach : Algorithms to be investigated

We plan to integrate the previous two approaches by adding the tweets as an additional feature set to the existing traffic record-based prediction. Following are the algorithms that will be used in our project.

1. Processing of tweets from public Twitter accounts to retrieve possible accident-prone geolocations

Adaptive data acquisition, Semi-Naive-Bayes classifier, and sLDA classifier models to collect and classify the tweets, with geocoding to predict the accident locations

2. Addition of Twitter data as a new feature set to the previous traffic record

The existing traffic record will be compared with the given Twitter geolocations. For each location (instance) in traffic record, if any of the Twitter geolocation is in a proximity of less than certain distance, the instance would be given according feature value. Whether this would be a binary classification or multinomial is to be determined.

3. Addressing Data Imbalance

Given the large imbalance between classes (street-hours with accidents and street-hours without accidents), we must carefully rebalance these classes to use for our models. In the past, this has been done through random sampling of the negative class.

4. Comparison of Decision Tree-based Algorithms

Previous work has shown that road accident risk is best predicted by models built with either decision tree-based algorithms or deep learning. As we have decided to focus on classical machine learning techniques, we will investigate the use of the former, namely Random Forest and its balanced variants.

Team Roles

Isaac

Implementation of the Random Forest algorithm and variants, e.g. Balanced Random Forest and XGBoost.

Analysis of qualitative differences between location of focus datasets and datasets referenced in previous research.

Kamil

Implementation of logistic regression, SVM and other alternative ML models to compare with our main ML models - RF and XGBoost.

Visualization of the result on arcGIS map for evaluation.

Hannah

Collection of tweets from Twitter API and work on pre-processing the data via Adaptive Data Requisition, Semi-Naive Bayes classification, and geocoding for location retrieval.

Add the pre-processed Twitter data as an additional feature set to the existing traffic accident record.

Evaluation

In our project, we will be mainly measuring the difference between the predicted probability of traffic accidents and the actual traffic accident records.

We are expecting that adding custom features such as weather, road infrastructure, and twitter geolocations would increase the accuracy in predicting the accident probability.

We will be using ArcGIS software to visualize the outcome on a map.

Wilson, D. (2018) *End result: an accident risk heat map*

To measure performance of each classifier at all thresholds, we will use **ROC curve, AUC value, and precision-recall scores** for each models used in the project

- Random Forest algorithm with further under- sampling (RF),
- Balanced Random Forest algorithm (BRF)
- imbalanced-learn RF
- imbalanced-learn BRF
- XGBoost algorithm (XGB)

Our goal is to obtain AUC value of over 90% as suggested in the paper by He´bert et al. (2019)

Timeline

- Retrieve tweets through Twitter API (Adaptive Data Requisition)
- Perform initial sanity checks on algorithm implementations from papers and *sklearn*

- Semi-Naive Bayes classification on tweets and geocoding
- Engineer feature sets for sourced datasets, identify and drop redundant features

- Addition of the pre-processed tweet data as a new feature set
- Begin modeling accident risk with implemented models and completed feature sets

- Finish modeling, begin evaluation
- Visualization of the predicted and previous record on a map

- Evaluation of the data and discussion
- Preparation of the final paper and presentation

Project completion