Project Proposal

Authors: Mengjie Zhang, Yilin Wang, Xinning Wang

Title: Using machine learning techniques to analyse the dataset of traffic accidents in UK

Introduction

With fast developing urbanization during the past decade, automobile traffic volume is generally increasing in major cities. In the field of transportation planning, annual average daily traffic (AADT) is a measure widely used to present how busy the road is. AADT means the total volume of vehicle traffic of a highway or road for a year, or by time of day and day of week. In the United Kingdom AADT is a measure of traffic used by local highway authorities to forecast maintenance needs and expenditure.

One of the goals of government transportation departments is to respond to traffic accidents and reduce the occurrence of accidents by maintenance. So it is important to gain as much information from historical records of traffic accidents. The change of traffic volume, along with road conditions are also expected to affect the occurrence of accidents. This project aims to apply machine learning techniques to model the accidents occurrence and severity according to road conditions and traffic volumes (mainly AADT)

Description of the dataset

The data were downloaded from Kaggle, named “1.6 million UK traffic accidents”. It has more than 60,000 observations and over 30 features (location, accident severity, weather conditions and date, etc). Thus, we suppose that this dataset is large and good enough to train a deep network. The other dataset were also from Kaggle, it recorded the annual average daily traffic conditions in minor and major roads, including over 25 variables and 10,000 observations. The links for these two datasets are as below:

https://www.kaggle.com/yesterdog/eda-of-1-6-mil-traffic-accidents-in-london/data

https://www.kaggle.com/coolcoder001/predicting-car-count/data

Methods

First, we will clean and preprocess the “accidents” and “trafficAADF” datasets. We will filter “the traffic AADF” data according to values of location variables from the “accidents” data and combine the two dataframes. The response variables are accident severity and casualties, and predictor variables are road type, speed limit, light condition, weather condition, road surface, urban or rural, AADFs of different types of vehicles, etc. We will of course label the categorical variables with integers and split the dataset to training and test dataset.

Then we will perform different machine learning techniques on the preprocessed dataframe. We will first use random forest to select important features from all variables. Then neural network and support vector machine and random forest will be useful in classifying accident severity. We will also use neural network to perform regression on occurrence and casualties of accidents.

Finally we will evaluate the performance of networks according to metrics including Gini and information gain, confusion table, ROC curve etc and determine the best network for this problem.

Schedule of the Project

Finish preprocessing the data by August 6th.

Finish building the networks by August 8th.

Discuss the result and write the report August 8 - 11th.

Submit the report August 12th.

Reference

Hoang Nguyen , Chen Cai, Fang Chen(2017). Automatic classification of traffic incidents' severity using machine learning approaches. *IET Intell. Transp. Syst.*, Vol. 11 Iss. 10, pp. 615-623.

UK Department of transport https://www.dft.gov.uk/traffic-counts/index.php
Kaggle https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales
Neural Network Design (2nd Ed), by Martin T Hagan