Developing Crash Models with Supporting Vector Machine for Urban Transportation Planning

Xiaoduan Sun, PhD and PE
Professor
University of Louisiana
Lafayette, Louisiana 70504, U.S.A
xsun@louisiana.edu

Subasish Das, PhD
Associate Transportation Researcher
Texas A&M Transportation Institute
U.S.A.

Nicholas Broussard, MS and PE
Traffic Planner and Engineer
Neel-Schaffer, Inc.
U.S.A.

ABSTRACT

Effectively incorporating roadway safety into transportation planning requires robust safety models that can quantitatively predict the safety performance of future planned roadway development options. Although various safety models have been developed including the models introduced in the first edition of Highway Safety Manual (HSM) by American Association of State Highway Transportation Officials (AASHTO), these models try to link roadway design features, such as lane with, shoulder, width, horizontal curve and vertical grade design with crash occurrences at disaggregated level and require the detailed inputting data and complex application procedures. Transportation planning mainly deals with type and functionality of roadway or roadway network. The HSM crash prediction models for urban and suburban roadway are complex involving several sub-models for different types of collisions, which makes it hard for transportation planning applications.

This paper introduces an innovative crash prediction model with so-called Support Vector Machines (SVM). Being a branch of machine learning, SVM focuses on the recognition of patterns and regularities in data. The dramatic growth in practical applications for machine learning over the last ten years has been made possible by many important developments in the underlying algorithms, techniques and readily available open-source programming code. Motivated by lack of suitable safety models for transportation planning, this study used the SVM with crash data from Louisiana urban roadways to develop safety models for urban 2-lane roadway, multi-lane roadway and freeways with satisfactory results. Comparing with parametric statistical regression models, the SVM model produces results can not only reach the same level of accuracy but also be straightforward for practical applications in urban transportation planning.
1. INTRODUCTION

Transportation planning is the act of evaluating the existing transportation system of an area, projecting its future growth, identifying its current and future deficiencies, and selecting transportation projects to address the deficiencies under the budgetary constraints. The United States Transportation Bill, Moving Ahead for Progress in the 21st Century (MAP-21) provides the funding for roadway projects across the country and basic guidance for what a Metropolitan Transportation Plan (MTP) should encompass. One of the important aspects of the Bill states that the transportation planning must increase the transportation system’s safety for all users, which requires that any transportation improvement plan must make an effort to improve the safety of the area.

Due to the lack of quantitative analysis method in the past, the safety benefit for the future transportation network was often determined by engineering judgment, or that of an experienced transportation planner in the past transportation planning practice. The first edition of Highway Safety Manual (HSM) published by AASHTO in 2013 (1) does tools for safety analysis in quantitative terms. However there are several issues that make the HSM analysis largely inconvenient and incompatible with transportation planning. For example, the required roadway data for HSM models are lane width, shoulder width and type, median width, side slopes, lighting, density of roadside fixed object and driveway, presence of auto speed enforcement and on-street parking, and etc. At planning level, the roadway information needed are limited to roadway functionality, number of lanes, and daily traffic volume. It is hard, if not totally unfeasible, to obtained the detailed roadway data at the planning stage for the safety analysis.

Furthermore, the HSM models for urban and suburban roadway are the most complex compared to the models for rural roadways as shown in Figure 1. In addition to models for five different roadway
type and four different intersection types, there are sub-models for five collision types on segment and
four collision types for intersection, which makes the safety analysis very time-consuming. The data
requirement and complicated application process make use of the HSM methodology in transportation
planning to be an inadequate approach and “too umbrella” to be used properly.

Thus, the objective of the research was to establish a crash prediction model for roadway planning
purposes with data readily available from the state Department of Transportation, local government
with jurisdiction of the transportation project(s), or another reliable source in order to determine the
safety impact of a transportation project on the roadway network.

2. LITERATURE REVIEW

Modelling roadway safety performance has been a popular research topic for the past two decades
while practitioners look for tools to link roadway design features to crash occurrences. The analytical
methods are the traditional modelling tools because they can provide end user with an equation to
predict the roadway safety performance for various purposes. However, analytical solutions to complex systems can often become inadequate since those analytical solutions provide a concise numerical solution to a complicated system that is hard to capture. In highway safety, the most representable analytical models are presented in the first edition of Highway Safety Manual published by American Association of Highway Transportation Officials (AASHTO), which discusses crash prediction models for 3 types of roadways based on 50 years of research results from not only the United States but other countries as well.

Another modeling method that has been developed rapidly in real world during the past 5 years is called machine learning algorithms that “learns” from the available data and determines how to perform the given task(s) by generalizing from example and mimicking the expected results (2). There are several machine learning algorithms types, which include clustering, support vector machines (SVR), fuzzy algorithms, and kriging methods. The biggest advantages of these models are capable of recognizing sophisticated patterns through data analysis algorithm and self modifying to increase predictability of complex relationships. The drawback to using the machine learning algorithms is that they do not provide an equation, but rather a “black box”.

In recent year, more and more research in transportation have utilized non-conventional machine learning methods in modeling because of the demonstrated modeling benefits. Support Vector Machines (SVP) as one of machine learning-vector regression techniques has been successfully used in Annual Average Daily Traffic (AADT) estimation (3 and 4). There are little studies in utilizing the SVP in safety modeling.

3. METHODOLOGY

Three data sets were first prepared for the modeling, crash, roadway attributes and AADT. The three most recent years of crash data (2011-2013) in Louisiana are used, which also include AADT rom 4(12)
the LADOTD. The use of three years of data also allows for the establishment of more recent traffic trends while avoiding a regression to the mean (a statistical event that makes natural variation between samples look like real change) bias. The Louisiana crash records contain data on crash number, highway number, AADT, control section, functional classification, highway classification, logmiles, latitude and longitude, milepost, and more. Table 1 lists the initial data items considered in modeling.

<table>
<thead>
<tr>
<th>Data Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTROL_SECTION</td>
<td>LADOTD Control Section number</td>
</tr>
<tr>
<td>LOGMILE_FROM</td>
<td>Control Section Logmile Starting point</td>
</tr>
<tr>
<td>LOGMILE_TO</td>
<td>Logmile Ending point</td>
</tr>
<tr>
<td>LENGTH</td>
<td>Length of segment</td>
</tr>
<tr>
<td>AADT</td>
<td>Annual Average daily traffic</td>
</tr>
<tr>
<td>FUNCTIONAL_CLASS</td>
<td>Highway function classification</td>
</tr>
<tr>
<td>HIGHWAY_CLASS</td>
<td>Type of highway</td>
</tr>
<tr>
<td>MEDIAN_WIDTH</td>
<td>Width of the median, if any,</td>
</tr>
<tr>
<td>NUM_LANES</td>
<td>Number of lanes in both directions</td>
</tr>
<tr>
<td>PAVEMENT_WIDTH</td>
<td>Total pavement width of the segment</td>
</tr>
</tbody>
</table>

To minimize the regression-to-the-mean effect, three years of crash data were used for 2-lane highway, multiple lane highway and freeway. The exploratory data analysis has yield the basic results shown in Figures 2, 3 and 4.
Figure 2. Results of EDA for Urban 2-Lane

Figure 2. Results of EDA for Urban Multiple-Lane
Two variables were removed based on the initial EDA, pavement width and number of lanes since lane width is directly linked with pavement width and the analysis is separated by type of roadway (indicating number of lanes). Totally more than 450,000 crashes in 15,400 segments used in the modeling. The 70% of data were used in trainset for model development. The SVR models used in this study are dependent upon the kernel type, value of the penalty for excess deviation during training ($C$, $Gamma$), and error-term value ($\epsilon$, Epsilon) for the $\epsilon$-insensitive loss function. The number of support vectors to be used in modeling is determined before running the SVR analysis. The models were run using an open-source software programming language, R, to predict the average yearly crash frequency. The model parameters are as follows:

- SVM-Type, eps-regression
- SVM-Kernel, radial
- Cost, a value of 100 in the study
- Gamma, a value of 1
- Epsilon, a value of 0.1
The SVM results are displayed as predicted crashed vs. observed crashed in Figure 6, 7 and 8 for three types of roadways.

Figure 6. Results for Urban 2-Lane SVM Model (Trainset)

Figure 7. Results for Urban 4-Lane SVM Model (Trainset)
It is clear that the SVR models are capable of predicting the crashes close to the observed values and have an acceptable amount of variation. Additionally, the models are not predicting values in a manner that is either consistently higher or consistently lower than the observed crashes.

4. MODEL VALIDATION

Model validation is used to determine if adjustments are necessary in order for a model to replicate the base model conditions as closely as possible. This means trying to get the predicted values as close to the observed values as possible, which can be done by changing the model parameters, the application of external factors for corrections, or other means through calibration if necessary. The validation process is meant to show that the model performs within a range of values that simulates the observed values shown in everyday application.

Validation of the crash predictions models is done at the individual model level using the trainset data. In order to validate the model an analysis of the $R^2$ of the data plots (observed vs predicted), average absolute difference between observed and predicted values, percent deviation, RMSE, and Percent RMSE was conducted.
The use of RMSE and Percent Root Mean Square Error (Percent RMSE) was chosen due to the fact that a raw aggregate sum and percent deviation comparison can be misleading. This is due to the fact that the total sums of the observed and predicted crashes can be very close, but individual segments can have a high amount of variation between them, resulting in what appears to be a good overall model performance but a weak performance at the segment level. The RMSE is a representation of the standard deviation of the differences between the observed field values and predicted values within the model sample. However, the RMSE does not provide information about the magnitude of the error relative to the observed values, leading to why the Percent RMSE is also computed. This measure expresses the RMSE as a percentage of the average count value. The Percent RMSE is defined as below:

\[ \%RMSE = \frac{\sqrt{\sum_{j} (Model_{j} - Count_{j})^2 / (\text{Number of counts})}}{\sum_{j} Count_{j} / \text{Number of counts}} \times 100 \]

Table 2 displays the model validation statistics with the Transet and Testset. The Testset uses model developed by Trainset.
However, the results from Testset, i.e., comparing the observed crashes with results from the model developed by the Trainset, are not significant, which is common in both conventional statistical methods and SVM method. The prediction of the SVM model (results of the Testset) can be improved by eliminating the outliers in the Trainset. Another way to improve the model is to determine the optimized loss-function by simulation that is the future direction of this research.

5. DISCUSSIONS

Urban transportation planning and metropolitan transportation plans are complex, regulated, and vitally important undertakings that impact the long range growth and health of an area. As such, the Federal government ensures that where possible the planning processes are adequately defined and the desired outcomes are well known. Historically the safety element scoring in these transportation planning documents has relied solely upon the judgment and knowledge of experienced professionals. The performance-based measures require using quantitative terms in cooperating safety onto transportation planning. While the use of HSM modeling is available to transportation planning, the applications are cumbersome, consuming and requires the detailed data that are hard to obtain at the
planning stage. The models developed in this project aimed to fulfill the safety analysis needs at the planning level with the data readily available at the planning level or easily obtainable. This research have shown that using local crash data and limited number of variables can lead to better safety models for local transportation planning. These models can be easily used to predict the future crash frequencies on roadways based upon their future conditions. Using the predicted crash frequencies, the relative change between two different improvement projects can be easily obtained for “what if” analysis. This will allow transportation planners to use the results to rank and score transportation projects based on safety and other factors to satisfy the current Transportation Bill requirements.

REFERENCE


