San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Team Project Proposal

San Jose Traffic Accidents Prediction

Ping Chen, Mandy Wong, Nihanjali Mallavarapu, Dhruwaksh Dave

1. Abstract

Traffic Accidents are everywhere and happened every time all over the world. According to the statistics showing in the Association for Safe International Road Travel, nearly 1.25 million people die in road crashes each year and 20-50 million people are injured or disabled. There are so many factors to make this happen. The Bad weather conditions, for example, rain, snow, ice, and windy, etc, is one of the major factors and will be increasing the possibility of a road crash than sunny days. This project will be focusing on weather-related factors to predict the possibility of traffic accidents will occur in San Jose, CA under certain weather conditions. In order to do that, we need to analyze the relationship between weather conditions and the number of car accidents. We need to find out the common causes and how the weather impact. Then, we can predict the possibility of traffic accidents occurred in different weather conditions.

2. Motivation for the project

Road Safety is an important issue nowadays and also recognized as a major public health concern. It is a shared responsibility for all of us. When traffic accidents have happened around you, it would ruin your day. We all like to avoid or prevent any kind of traffic accidents happened. Apart from having a good driving attitude and fastening your seatbelt while you are driving. It is also better to know if you are in good condition to drive in different weather conditions to prevent any traffic accidents happened. Traffic accidents are a serious problem. It is not just about car damage, but could also lead to injury or even death. On the other hand, the weather is a natural condition. Even now that we are more precise about predicting the weather, we still cannot predict what will happen on the street. Therefore, the weather condition is still one of the major factors of traffic accidents. As a human, we cannot control the weather, but we can control our behavior. We are hoping this project could help to reduce traffic accidents in any weather conditions, through to minimize the major uncontrollable factor. Also, traffic accidents are not just affecting the party, it is also affecting others behind the accident. Every traffic accident is causing a certain level of traffic jams. As a student commuting to San Jose State University for classes, we always spare more time more than the GPS estimated time to arrive. We all have wasted so much time stuck in a traffic jam on the way to school. When this project helped to reduce the accidents, we are also hoping it could help save us time for commuting.

3. Brief literature survey

There are several papers have demonstrated the methods to predict traffic accident . Some of the papers we reviewed are outlined briefly below:

In [1], the authors collected huge heterogeneous urban datasets to predict whether an accident will occur or not for each road segment in each hour. The datasets included all the motor vehicle crashes in the state of Iowa from 2006 to 2013, detailed road network, and hourly weather data such as rainfall, temperature, census data which gives the population corresponding to a sub-area. Then they preprocessed the datasets by interpolating the missing values in weather related features and matched crash data with road networks, etc. They compared four classification models: linear SVM, Decision Tree, Random Forest and Deep Neural Networks. Based on the datasets which includes 415,000 crashes containing 40 features, the results showed that DNN got the highest accuracy 0.9512.

In [2], the authors using two supervised learning models(ANN and Decision Tree) to predict traffic accidents. They divided the features into four key factors: driver factors, road factors, vehicle factors and climate factors. To reduce the complexity of the model, they did dimension reduction based on domain knowledge and other techniques. Then they split the datasets into training set and test sets, using ANN and Decision Trees to build the model. The experiment conducted on 4861 crash records and 14 attributes. The accuracy of ANN model was 79.8% and accuracy of Decision Tree is 77.7%.

Chang, et al employed a negative binomial regression model and an ANN model to analyze accident data for National Freeway 1 in Taiwan. They investigated the relationship between vehicle accidents and highway geometry, traffic characteristics and environment conditions. The number of sections used for model estimation is 1500, and the number of sections used for testing is 492. For the negative binomial regression model, the overall model prediction accuracy for the training data is about 58.3%, while that for the testing data is about 60.8%. For the ANN model, the overall model prediction performances for the training data and the testing data are 64% and 61.4%, respectively. The author concluded that ANN is a consistent alternative for analyzing freeway accident frequency by comparing the prediction performance with negative binomial regression analysis.

[4] presented a two steps methods to prediction roadway traffic crash. The SSM(state-space model) was developed in the first step to identify the dynamic evolution process of the roadway systems that are caused by the changes of traffic flow and predict the changes of impact factors in roadway systems. Using the predicted impact factors, the SVR(support vector regression) model was incorporated in the second step to perform the traffic crash prediction. This model was evaluated in a five-year dataset that obtained from 1152 roadway segments. The proposed models result in an average prediction MAPE of 7.59%, a MAE of 0.11, and an RMSD of 0.32.

4. Methodology

This section presents how we will conduct the experiment design to predict San Jose traffic accident based on historical data. First we will briefly introduce data preparation, then we will talk about four machine learning algorithms to conduct the experiment design. At last, we will present how we verify the results of each model.

4.1 Data preparation

Motor Vehicle Crash Data: We obtained crash data in the City of San Jose DOT[5]. This data set shows the location of individual crashes where one or more fatalities and/or severe injuries occurred during the five-year period of 2013 to 2017. The data including 940 crash data. Since these crashes are mapped to the nearest intersections, each crash contains the following information: Crash Location/Intersection, Date/Time, Injured Party, and Number of Fatal or Severe Injuries per crash.

Road Networks: We collected road datasets from San Jose DOT with basic information in San Jose, street name, nearest intersection, speed limits, street segment length and the most recent average daily traffic.

Climate Data: We also obtained the historical weather data in the website of California Agriculture & Natural Resources [6]. The weather information we retrieved including observation time, precipitation amount, max temperature and min temperature.

Data preprocessing: After we got this data, we will evaluate the data quality of the dataset we acquired. We plan to perform below steps in data preprocessing [7]:

Data Cleaning: Since dirty data can cause confusion for the mining procedures for dealing with incomplete or noisy data, they are not always robust [7]. Instead, they may concentrate on avoiding overfitting the data to the function being modeled. Therefore, it is essential to run the data through some data cleaning routing. Clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. For example, in our scenario, the road network features contain many missing values (eg. the average daily traffic data is not available for some roads). Therefore, we need to fill in the missing values either by using a global constant or using mean/median, etc.

Data Integration: Merge multiple datasets into a coherent data store. This can improve the accuracy and speed of the subsequent data mining process. When matching attributes from one database to another during integration, special attention must be paid to the structure of the data. And we also need to do correlation analysis in order to avoid redundancy and data value conflict detection to avoid different attribute values from different sources.

Data reduction: This techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Mining on the reduced data set should be more efficient yet produce the same analytical results. Data reduction strategies include dimensionality reduction, numerously reduction and data compression. We can pick up one or two methods to process our data.

4.2 Algorithm

In this project, we need to find the machine learning algorithm which can get the highest accuracy for classification. We are going to talk about four machine learning models, points out the difference and find the most suitable model for solving our problem.

Logistic Regression: Logistic Regression is commonly used to estimate the probability that an instance belongs to a particular class. If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the positive class, labeled “1”), or else it predicts that it does not (i.e., it belongs to the negative class, labeled “0”). This makes it a binary classifier. Logistic Regression model computes a weighted sum of the input features (plus a bias term), and it outputs the logistic of this result [8]. We need to train the model to find the best weights in order to get the smallest cost.

Support Vector Machines: SVM is another algorithm for classification by finding a hyperplane in an N-dimensional Space to classify the data points. The hyperplane is selected to find the maximum distance between the classes [8]. The hyperplane is learned from training data using an optimization procedure that maximizes the margin. SVM has three important parameters: C, gamma and kernel. In this project, we plan to fine tune the parameters to find the most desirable outcome of this model.

Random forests: A random forest multi-way classifier consists of a number of trees, with each tree grown using some form of randomization. The leaf nodes of each tree are labeled by estimates of the posterior distribution over the image classes. Each internal node contains a test that best splits the space of data to be classified [8].

Artificial Neural Network: An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the brain. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning largely involves adjustments to the synaptic connections that exist between the neurons[8]. Both [2] and [3] concluded that they got the highest accuracy using ANN, we will try this model in our project.

4.3 Evaluation

In this project, we will label the dataset into two categories: 0 for non-accident and 1 for accident. They we will split 80% of total datasets to be training data and the remaining 20% to be test data. We will train our training set using the four models mentioned above then evaluate each model on test set. Finally we will evaluate each model based on the following factors:

Compare the training accuracy and test accuracy, decide whether overfitting occurs or not;

Find the model delivers the highest accuracy.

5. Deliverables

The milestones we have for each phase of the project are

Initial Phase

Collecting project relevant data related to motor vehicle crashes, road networks, fatalities and weather data from reputed and trusted sources which can be used for feature extraction and eventually as a training data set after processing and pruning.

Intermediate Phase

1) Pruning the data and reduction of dimensions so as to ensure that maximum results can be predicted from the least amount of relevant features and attributes.

2) Development of model generated by each algorithm which is being evenly distributed to each team member so as to ensure equal participation and to maximize efficiency of the team.

Final Phase

Presenting the models which are generated by the different algorithms and comparing predictions on the basis of accuracy to settle on the best fit for a potential research paper in the future, provided we come up with enough noteworthy developments in the same.

Detail Timeline:

Milestone	**Projected End Date**
Data preparation	9/27
Data preprocessing: Data Cleaning	10/4
Data preprocessing: Data Integration	10/11
Data preprocessing: Data Reduction	10/18
Development: Machine Learning Algorithms	11/1
Model Evaluation	11/15
Final paper report and presentation	11/26

6. Team members and their roles

Team Member: Ping Chen, Dhruwaksh Dave, Nihanjali Mallavarapu, Mandy Wong

Team Roles:

- Data Preparation: Since we have three data resources, each of us will choose one feature to do data preparation and the rest of us will do data integration.
- Algorithms: Each of us will choose one algorithm to train the data and then compare the results of the four mode:

Name	Data Preparation	Algorithms
Ping Chen	Road Network Dataset	ANN
Mandy Wong	Climate Data	SVM
Nihanjali Mallavarapu	Motor Vehicle Crash Data	Logistic Regression
Dhruwaksh Dave	Motor Vehicle Crash Data	Random forests

Reference:

[1] Yuan, Zhuoning, et al., "Predicting traffic accidents through heterogeneous urban data: A case study." 6th International Workshop on Urban Computing (UrbComp 2017). 2017.
[2] Roop Kumar R, et al. “DATA ANALYSIS IN ROAD ACCIDENTS USING ANN AND DECISION TREE.” International Journal of Civil Engineering and Technology (IJCIET). 2018
[3] Chang, Li-Yen. "Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network." Safety science 43.8 (2005): 541-557.
[4] Dong, Chunjiao, et al. "Roadway traffic crash prediction using a state-space model based support vector regression approach." PloS one 14.4 (2019): e0214866.
[5]http://gisdata-csjdotgis.opendata.arcgis.com/?geometry=-123.635%2C37.161%2C-121.523%2C37.489
[6] http://ipm.ucanr.edu/WEATHER/wxactstnames.html
[7] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
[8] Géron, Aurélien. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.", 2017.