Predicting Car Collisions on Highways

Diego Rodriguez

Introduction

According to the California Highway Patrol, there are on average over 400,000 vehicle collisions per year just in California. About less than 1% of those collisions end in fatalities and roughly 35% of those collisions result in injuries. Most collisions are due to risky behavior driving [1] because drivers are simply unaware of their risky driving [2]. In California, there is an ongoing statewide effort to reduce the number of collisions through education and there has also been an ongoing decline of collisions since 2005[3]. Unfortunately, there has not been any implemented technological measure to date that helps prevent car collisions.

Currently, there are no systems that attempt to predict highway accidents. The closest to this is a built-in car technology that can predict collisions just milliseconds before it happens in order to activate an emergency safety system [4]. Another similar system was also recently developed to predict the locations of deer for discovering car collision hot spots in a specific area [5]. However, none of these predict car collisions on the highway.

The goal of this project is to implement a system to predict car collisions by the hour. Predicting car collisions can be used to help drivers prevent accidents on the highway, especially if collision prediction information can be used in conjunction with the driver's current driving information. However, this project only tries to predict collisions without any driver information.

Prediction System

To be able to predict car collisions on the hour, a support vector machine (SVM) was implemented (libsvm[6]) to learn from previous collision and weather history. The SVM used a Gaussian kernel; a linear kernel was first attempted but the preliminary results were abysmal. Feature normalization was performed and data set examples were split into three groups: 60% training, 20% cross validation, and 20% for testing. Parameters for the SVM Gaussian kernel were varied for the C, the cost parameter, and γ for regularization.

Feature Selection

Data on vehicle collisions was collected through 2005 to November 2011 from the California Highway Patrol Statewide Integrated Traffic Records System, which provided the time and day of the collision. It was also possible to add latitude and longitude as a feature but that data was inconsistent in the collision dataset; some data points had the latitude and longitude of the collision, but many did not contain such information. Figure 1 displays a summary of the number of collisions per hour over the years. This dataset was chosen as the label since the goal of this project is to predict car collisions on the hour.

For the weather, historical data was provided by the National Climatic Data Center from the National Oceanic And Atmospheric Administration for the years 2005 to November 2011. Features from the weather data set chosen were temperature in Fahrenheit, wind speed in miles per hour, visibility in miles, and the amount of precipitation in inches. Temperature was chosen because it can affect the tire pressure of a car, which can affect breaking, cornering, and stability of the car. Wind speed was also chosen as it can also affect the safety of a driver because the wind can push the car off course, thus giving less control to the driver. Visibility indicates how far

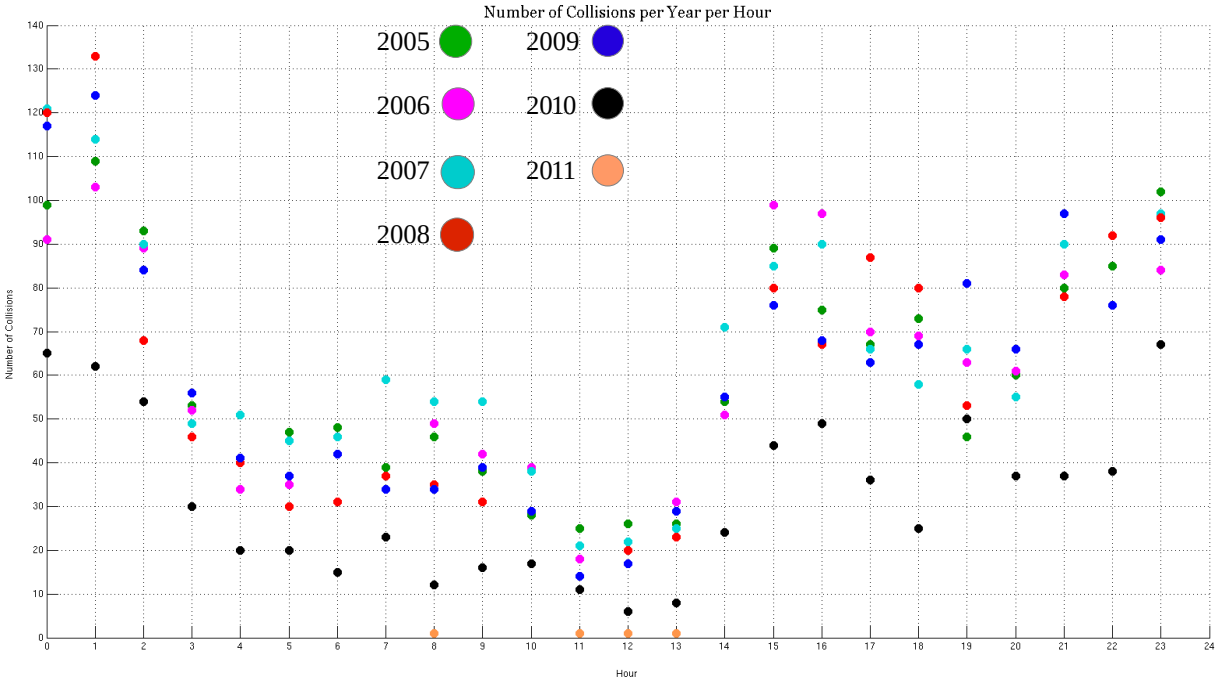

Fig.1 : The total collisions for a given hour for the years 2007 to 2011. Notice that most collisions are from the 23rd hour to 2nd hour.

the driver can see and was chosen as it affects how much the driver can see on the road and finally, precipitation was chosen because it make road more slippery for cars and can also one of

the factors of the collision. All historical weather information was recorded on an hourly basis, providing over 60,000 data points.

For the purposes of this project, the scope of the data was limited to information observed from the lower California Bay Area and highway US-101 ; the areas covered were San Francisco county, San Mateo county, and Santa Clara County. The system can be scaled to include other areas by simply using a different data set from another location.

From this data, four different data sets were devised. All of the following data sets included the weather information, but information on the time of the collision was varied to see if more time information would help the SVM learn better. The following list shows what each of the four sets contained besides the weather data:

Set A: Hour
Set B: Day of week and hour
Set C: Day of year and hour
Set D: Day of year, day of week, and hour

The hour was represented in 24 hour format, the day of the week was a number from 0 to 6 with 0 being Sunday, and the day of the year as a number from 1 to 366.

Results

After having the SVM learn from the data, the best results came from data Set D where all of the time information was provided. Set A performed the worst, while the other data sets had similar performance to Set D. The results are summarized in the tables below:

Data Set	C (cost parameter)	γ	Accuracy	F1 Score
Set A	50	4.0816	80.49%	0.125100
Set B	50	2.0000	80.31%	0.173070

Set C	35	3.1250	79.10%	0.177049
Set D	70	1.0204	76.64%	0.189411

Conclusion And Discussion

Although the accuracy was generally around 80%, the results for the SVM were unimpressive. The best F1 score achieved was from set D with a score of 0.1774. By looking closer at the test results, the recall rate was about 12% and the precision was generally around 20%, except for Set D which had its best precision at 12.21%. However, Set D was able to attain a higher recall rate at 18.71%. Since most of the time there are no car collisions occurring, the SVM attained a high rate of accuracy by simply predicting that a collision would not occur within the next hour. However, it was hitting mostly true positives on the weekends and during the early hours of 12:00 AM to 4:00 AM. Thus by simply adding the day of the week, as can be seen by comparing the F1 scores of Set A and Set B, the precision of the SVM is increases by about 8 points. Having the day of the year, the day of the week, and the hour as part of the feature set helped increase the recall percentage in Set D, but did not increase the precision at all. The hour and weather information at which the collision occurred was simply not enough for the SVM to attain a higher precision.

Given the results from the SVM, it is not enough to predict a car crash within the next hour from only collision history and weather. Perhaps if data from the drivers whom had collisions, such as care make, year, and driver gender, were provided, then more accurate results would come. This system could then be combined with real-time data from a drivers smartphone to help increase the precision of the SVM machine and then with enough precision it could be used to help prevent collisions. Since most American citizens today have a smartphone, gathering the realtime information such as speed, acceleration, and other data would not be impossible.

Nonetheless, after thorough testing, it seems that the current features for the SVM machine are not enough to accurately predict a collision within the next hour, more features are need to achieve higher precision predictions.

Resources

[1] : “*Traffic Safety Facts: September 2009”* p. 1. U.S. Department of Transportation’s National Highway Traffic Safety Administration

[2]: Currie, Donya. “Drivers Unaware of Risky Behaviors”. The Nation’s Health. June 29 2011.

[3]: “2009 Annual Report of Fatal and Injury Motor Vehicle Traffic Collisions” p. 4. California Highway Patrol.

[4]: Simonite, Tom. "Crash-predicting Car Can Brace Itself for Impact." *Science News and Science Jobs from New Scientist - New Scientist*. Copyright Reed Business Information Ltd, 27 May 2008. Web. 16 Nov. 2011. <http://www.newscientist.com/article/dn13973-crashpredicting-car-can-brace-itself-for-impact.html?feedId=online-news_rss20>.

[5]: "Predicting Locations for Deer vs. Car Collisions." *Science Daily: News & Articles in Science, Health, Environment & Technology*. Science Daily, 30 Jan. 2011. Web. 19 Nov. 2011. <http://www.sciencedaily.com/releases/2011/06/110630131826.htm>.

[6]: Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm