Traffic Accident Analysis Extension Using Machine Learning

Jeff Conway
Software Engineering, Drexel University

Date: May 18, 2014

This paper draws on the data and methods used in a previous paper, "Traffic Accident Analysis Using Machine Learning Paradigms" (2005) by Chong, Abraham, and Paprzycki. In this paper, the authors attempted four different methods to analyze the probability of injury types (fatal, incapacitating, ...) using the NASS GES machine learning data sets, which include hundreds of variables on tens of thousands of vehicle accidents over dozens of years in the United States. The goal of this current paper is slightly different -- rather than directing information towards traffic engineers, these learning algorithms can now be directed towards autonomous vehicles to increase their safety. In particular, we proceed by producing data in a manner nearly identical to that of the original authors, and then replicating their methods and extending them by working around an error which could not be solved by the time of their initial publication. Results were generally promising in the sense that algorithms with even limited data can do, in some cases, an excellent job of predicting injury given surrounding circumstances.

1 Introduction

This paper was motivated by the realization that two somewhat disparate advances in recent years complement each other to a great degree -- namely, autonomous vehicles and machine learning. This introduction will explore why these two disciplines are so important when paired together, and the rest of the paper will begin some work on using them side by side.

In recent history, much of the research done on increasing the safety of transportation systems (in this case, specifically road systems) has been done by expensive, commissioned studies that are relatively narrow in their scope. One study may find the effect of speeding on accident rates; another study may investigate the effectiveness of marking certain areas as school zones, etc. Though this information is undoubtedly useful, it always leaves behind the infamous quote of academia: "More research must be done."

However, this has begun to change. The National Automotive Sampling System has compiled a massive amount of data into a repository referred to as the General Estimates System, called NASS GES, which contains information on hundreds of thousands of accidents from the past few decades. Although data is sometimes incomplete or the format has changed, it is relatively well documented and has proven extremely useful for traffic researchers. The NASS GES data for 2012 (the most recent year available) served as the basis for this paper. Using this data, researchers are able to investigate the effects of *hundreds* of variables on nearly any situation they'd like and, of most important to traffic engineers, accidents. While congestion may seem more present in one's everyday life, it is nearly self-evident why accidents and specifically fatal accidents are of utmost importance to traffic researchers.

The flaws in NASS GES data are apparent in a few ways, even just from the paragraph above, though. Firstly, data comes out very slowly -- as of this writing in May 2014, the data from 2013 is still not available. Data from 2014 will surely not be available for at least another year. Some data is incomplete, as accident reports may be done in haste by emergency responders who have more urgent

things on their minds than data collection. The accuracy with which the data is recorded is limited by what a human responder can provide -- namely, things which are reported or easy to investigate, and even reported items could be false. So what can be done about it?

This is where autonomous vehicles come into play. Actually, the self-driving part is not particularly important - what *is* important is that such vehicles carry extremely powerful sensors with them which record perhaps billions of data points per minute. The benefits of this are numerous.

First, all accident data will be recorded accurately, not just that which responders deem important enough to record. Second, data from *before* an accident, stored in the sensor's buffer, will also be recorded accurately (rather than reported by a human or found through an investigation). Thirdly, enormous amount of information on circumstances that *don't* lead to accidents will be reported, giving examples of good behavior. Fourth, data will be available instantly, since data is useless to an autonomous vehicle if it comes late. Some downsides, however, include variables that a machine may be bad at measuring, such as the driver's blood alcohol content. Although it is possible to measure this, it would sure be seen as impractical and invasive. Furthermore, what's the point if the car drives itself?

Regardless, hopefully by now the motivation is clear: the powerful sensors shipped with autonomous vehicles provide an incredible wealth of information that can be used to teach all autonomous vehicles how to behave appropriately on the road to avoid serious accidents. Perhaps rather than worrying about the effect one some subset of variables on some subset of outcomes, researchers can simply collect it all and analyze it all, more or less throwing hypotheses to the wind (unless they choose to revisit them later). But will machine learning be good enough to keep us safe? Or do we need human intelligence to guide us? Let's find out.

2 Data Preparation

Data for this paper was prepared, as per the assignment's instruction, in almost exactly the same way that the authors of the original paper prepared theirs. The only difference to be aware of is that the original paper used information from multiple years. The most recent data, 2012, is formatted differently from the years that the original authors drew from and even from 2009 - 2011. 2012's data would most likely prove to be sufficient regardless, since it contains more than 64,000 records, and using just one year's data would provide absolute certainty that the data was consistent and clean.

The first step taken by the authors (and duplicated here) was to consider only accidents which were head-on (front-to-front) collisions, since those had the highest probability of fatal injury. After including only these records, the dataset contained roughly 2,000 records -- still definitely sufficient for machine learning sample sizes.

The original authors' data set included speed at time of accident as well as the speed limit of the area, which could of course be used to calculate speeding percentage and absolute speeding overage (or underage). However, the 2012 data did not include these variables, and thus this step was essentially skipped because it was replicated by default.

From here, the data was split into a training set and a validation set, where each one was roughly 50% of the total number of records available at that point. To be specific, the training data had roughly 989 records and the testing data had roughly 930 records. The original paper used 60%, 20%, and 20% for training, cross-validation, and testing respectively. Here, the testing portion has been reduced, and rather than worry about validation steps, we will simply immediately test any model that we're interested in. Additionally, modified data sets were created for each feature of interest in an "all against one" fashion - for example, while testing for fatal accidents, the labels would be 1 for fatal accidents and 0 elsewhere. This allows for greater accuracy using discrete classifiers.

3 Implement and Extend

At this point, the authors decided on four different machine learning paradigms to use: neural networks with hybrid learning, decision trees, support vector machines, and a hybrid decision tree (DTANN). While reading through the results of the original paper, notice that the authors were unable to finally implement any support vector classification algorithms. This presented a perfect opportunity to implement their data model while extending a method which they had been unable to try, and so the goal now is to investigate the effectiveness of SVM classifiers on this data. Slightly more ambitiously, the goal is to beat their accuracies. Their accuracies are as follows:

Table 1 - Original Accuracy

Injury Type	% Accuracy
No injury	67.5
Non-incapacitating	62.2
Incapacitating or fatal	81.4
Fatal	90.0

When it comes to modern computers and machine learning, it turns out that brute force actually isn't such a bad method in certain cases. Since the goal here was simply to provide the best SVM classifier possible, this provided such an opportunity - the data would be fit and predicted using a large number of parameters, and the best would be recorded and returned. Since the data set size was quite reasonable, about 2,000 samples in all, the time necessary for this was trivial - just a few seconds at most.

The first task is to import the data, which is easy enough using NumPy. Secondly, it turns out that a second variable called *imputed_maximum_severity_injury* also exists in the data set, so each learner also removes this variable to ensure that the classifier doesn't just use one column unfairly as a guide. Next, a parameter grid was set up for any SVM classifier that was desired - in particular, NuSVC, generalized SVC, and LinearSVC from SciKit Learn's API were attempted. The parameter grid was then passed to a function which iterated through every possible combination of parameters that could be generated from the grid, fit the data, tested it, scored it, and returned the operator with the best score on the testing data.

The NuSVC operator, for some reason, refused to cooperate with the data provided and always produced an exception with the message that "the nu value was infeasible." However, other operators were met with much more success, as described below.

4 Results

The first problem to be tackled should of course be the most important and most costly problem to society, especially if the cost of addressing that problem is equal to the cost of addressing less important problems. In this case, that's true, and so the first fit attempted was for fatal accidents. After playing with parameters for a bit of time, a classifier with very promising results was found. In particular, it was a generalized SVC classifier with the following parameters:

Table 2 - Fatal Accident Classifier

Name	Value
C	10
Cache size	200
Class weight	auto
Coefficient 0	0
Degree	2
Gamma	0
Kernel	Sigmoid
Max iterations	Infinity
Probability	False
Shrinking	True
Tolerance	0.001

With these parameters, a prediction accuracy of 94.6% was achieved. Of course, the true accuracy of the model is not exactly 94.6%, but instead probably somewhere in the interval of about [93%, 96%]. Regardless, this classifier has already accomplished two significant goals. Firstly, it implemented a support vector machine classifier, which the original authors were unable to do. Secondly, it achieved a greater accuracy with regard to fatal accidents than the original

authors could by a few percentage points. This may seem small, but even tenths of a percentage point translate into thousands of lives saved or affected by the model.

Next up was the set of car accidents which resulted in *either* an incapacitating injury or a fatality. The logic behind this selection is that while all accidents are unfavorable, there is clearly a large gap in the favorability between "a car accident from which you never recover" and "a car accident from which you quickly recover fully." The latter is a small disturbance in one's life; the former is a permanent change to the way one's life is lived. Thus, it makes sense to also consider ways to avoid specifically this subset of accidents.

A similar method was used for this set as was used for fatal accidents only. In this case, however, generalized SVC didn't deliver especially promising results. The best operator fit at about 68%, which is not bad, though it does not beat that achieved by the original authors. Thus, another attempt was made, this time using a Linear SVC model, and this worked out much better:

Table 3 - Fatal or Incapacitating Classifier

Name	Value
C	1
Class weight	None
Dual	False
Fit intercept	True
Intercept scaling	1
Loss	L2
Multi-class	OVR
Penalty	L1
Tolerance	0.0001

Using these parameters, predictions turned out to be roughly 97.1% accurate. Again, this is probably a reflection of some reasonable margin of error, perhaps as low as 95% or as high as 99% percent. This also turned out to be better than what the authors had achieved for such accidents, which was 81.4%. In this case, actually, it was *much* better, a result which could save or greatly enhance many, many thousands of lives if the implications of this result could be implemented meaningfully in autonomous vehicles.

Next up were the non-incapacitating injuries, where the original authors were able to predict them with about 62.2% accuracy. Using methods extremely similar to the above - similar enough to not be worth repeating - good results were again achieved using a LinearSVC fit, where the top reported accuracy was roughly 95.5%.

Table 4 - Non-incapacitating Classifier

Name	Value
C	1
Class weight	None
Dual	False
Fit intercept	True
Intercept scaling	1
Loss	L2
Multi-class	OVR
Penalty	L1
Tolerance	0.0001

Lastly came concentration on what might be called "good behavior," an accident which led to no injury at all. Though accidents are never favorable, it's better for a vehicle to be in position for a no injury accident than in position for a fatal or incapacitating injury accident. Thus, to investigate the causes of such accidents is still worthwhile. Noticing that the two classifiers used in the previous two accident types were identical, it was decided that time might be saved by simply trying the same classifier on this last set of data. Sure enough, it produced excellent results -- 100% accuracy with the sampled data. Again, there's no guarantee that it will predict perfectly forever, and there are always outliers, but out of the more than 900 predictions that it made in the test case, every single one was correct. Thus, in conclusion, the table of results for the SVM classifiers is as follows on the next page.

Table 5 - Updated Accuracies

Injury Type	% Accuracy
No injury	100.0
Non-incapacitating	95.5
Incapacitating or fatal	97.1
Fatal	94.6

5 Concluding Remarks

From even just this small dataset, it seems that machine learning and the dearth of information provided by modern vehicle sensors do make a very good pair. Using quite simple algorithms and relatively limited data, we were able in this paper to predict the most costly types of accidents with extremely high probability, which could translate into saving an enormous number of lives over time, or preventing a great number of permanent handicaps. There is no doubt that collection of more variables from more vehicles and incorporating and analyzing that data more quickly - even continuously - would result in significant safety gains. With some hard and clever work up front, hopefully the predictive models could get to a point where accidents are avoided in all circumstances and that dreaded academic adage, "More research must be done," can be abandoned in this very important domain.

6 References

[1] Chong, Abraham, and Paprzycki, Traffic Accident Analysis Using Machine Learning Paradigms. *Informatica*, 2005.

[2] National Automotive Sampling System - General Estimates System. *National Highway Traffic Safety Administration*, 2013. http://www.nhtsa.gov/NASS

[3] SciKit Learn Developers, SciKit Learn, 2013. scikit-learn.org/stable/index.html