Interpreting Random Forest Predictions for Firearm Identification Using LIME

Katherine Goode and Heike Hofmann

February 08, 2019

1 Introduction

(need to finish adding sources)

The discipline of firearm identification examines bullets to determine the likelihood that a bullet found in a criminal case was fired from a particular gun. To do this, the bullet from the crime will be compared with a bullet that was known to be fired from the gun under evaluation. Traditionally, this is a procedure that has been performed by hand. Specially trained examiners visually compare the microscopic bullet striations that were created when the bullets passed through the gun barrel. Often a comparison microscope is used that allows the examiners to view both bullets at the same time (National Research Council 2009). The examiners use this ability to determine whether the two bullets were fired from the same gun barrel.

Recently, the scientific community has been encouraging the inclusion of more data driven techniques to be used in forensic investigations. These methods would allow for the reporting of a measure of uncertainty in addition to the conclusion drawn from the analysis. This led Hare, Hofmann, and Carriquiry (2017) to propose a new computer automated method of bullet matching that could supplement the visual inspection by the firearm examiners. Their method involves obtaining bullet signatures of the striations from the scans of the two bullets, computing variables that measure the similarity of the two signatures, and using a trained random forest model to determine the probability of a match based on the similarity variables. They trained their model on a set of known bullet matches and non-matches from the Hamby study. They demonstrated how the random forest model can be used to make prediction on a new set of bullet signature comparisons. The results from the paper suggest that the random forest model leads to highly accurate bullet matching predictions. The authors found that when their model was applied to a testing dataset (?), the resulting error rate was 0%.

While random forest models often result in good predictions as seen in Hare, Hofmann, and Carriquiry, it is well known that a disadvantage of random forest models and other machine learning techniques is that it is difficult to interpret the models (reference?). For example, it is not possible to tell which variables played an important role in the creation of individual predictions. This issue led to the development of LIME (reference), which is an algorithm that examines the behavior of the complicated model on a local scale around a new prediction using a linear regression model. This allows for the ability to understand which were the driving variables that led to a prediction of interest.

(should probably list other current methods for interpreting random forests)

Since firearm identification identification is commonly used as evidence for convictions in court cases, it is important to be able to understand and assess the model that is being used to quantify the probability that a bullet was fired from a gun. LIME provides the ability to understand which were the key variables used by the random forest model to make a prediction, which would allow firearm examiners to check whether or not the predictions created by the random forest are based on reasonable variables.

This paper provides an example of the application of LIME to a bullet matching problem. (provide more details about what is contained in the papers - i.e. section 2 describes the Hamby data; section 3 describes the random forest model and LIME; . . .)

2 Data

2.1 Training Data: The Hamby 173 and 252 Datasets

2.2 Testing Data: The Hamby 224 Clone

The Hamby 224 Clone is organized as a test set of a cloned (sub-)set of the Hamby 224 bullets. As with all Hamby sets (Hamby, Brundage, and Thorpe 2009), Hamby set 224, is a collection of 35 bullets, organized as 20 known bullets and 15 questioned bullets. The known bullets are fired in pairs of two through one of ten consecutively manufactures P-85 barrels. Clone set 224 is arranged as a test set of fifteen tests, one for each questioned bullet. Each test set is arranged as a combination of three bullets: two known bullets and a questioned bullet. The test asks for a decision on whether the questioned bullet comes from the same source as the two known bullets or from a different source. This situation is similar to what a firearms and toolmarks examiner might encounter in case work.

3 Methods

3.1 Random Forest Model

3.2 Overview of LIME

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin developed LIME in response to the interpretation issue of many machine learning techniques. Many models used in machine learning are referred to as black box prediction models since they are complex models that may provide accurate predictions but are not able to be interpreted. That is, it is not possible to understand the relationship between the features and the response variable. Without being able to interpret the models, it becomes difficult to diagnose why a model is producing the predictions that it is. The developers claim that LIME allows users to check which variables are driving the predictions made by the model. This further allows users to determine whether their model is trustworthy.

LIME was designed to be a generalized algorithm that can be applied to any predictive model to return an "explanation" for an individual prediction of interest. This explanation is meant to convey which features from the model are driving the prediction created by the complex model. In the next section, I will provide a detailed description of how the LIME algorithm obtains the explanation when applied to a random forest model with numeric features. For now, general overview of the LIME algorithm will be explained.

Suppose that a complex model has been fit on a training dataset with k features and a response variable that may either be numeric or categorical. This trained model will be applied to a testing dataset containing the k features to obtain predictions. It is of interest to understand which variables played an important role in the predictions. LIME will be applied to one of the cases in the testing dataset to obtain an explanation. To create the explanation, LIME starts by approximating the distribution of each of the k features in the training dataset. It then simulates a large data frame from these distributions. The simulated data are used to perform feature selection to determine the m most important features. Then a weighted ridge regression model is fit to standardized versions of the selected m features from the simulated data. The model assigns higher weights to observations closer to the observed case of interest. A linear model is used since it has a well known form and coefficients that can be interpreted. The coefficients from the model are extracted and used to "explain" the importance of the variables. Features with a larger absolute value of their coefficient are implied to play a larger role in the creation of the prediction.

The developers of LIME wrote a Python package to implement LIME, and Thomas Lin Pedersen adapted their work to create an R package called `lime`. This paper will use version 0.4.1 of the R package `lime` in the analysis of the bullet matching data.

3.3 Applying LIME

(this is the section where I will describe how lime was applied and which input values were used)

3.4 Development of Shiny App for Visualizing the LIME Explanations

4 Results

- show examples of using the app and what we have learned from the LIME explanations

5 Discussion

Ideas for improvement of LIME: - change the binning method: - lime is not good with linear relationships with classifiers due to the inside-outside binning - it would be better to use a cumulative approach - consider interactions in the ridge regression model - learn lambda in the ridge regression - could add a penalty for the number of parameters in the model (or the number of bins) - could start with a fine grid of bins and then go backwards and fit the models on this sampled data - could try a type of ANOVA test if we can have nested models or if we can assume nesting - could compute something like an R^2_adj - could force the intercept to be 0.5 - try out the subsampling

References

Hamby, James E., David J. Brundage, and James W. Thorpe. 2009. "The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries." *AFTE Journal* 41 (2): 99–110.