DS-GA-1001
Fall 2016 Introduction to Data Science
Project Proposal

Yiqiu Shen(ys1001), Xinsheng Zhang(xz1757), Zemin Yu(zy937)

October 30, 2016

1 Business Problem

Nowadays, issues about terrorism has gain increasing attention from the general public as it becomes the greatest threat to the public security. In lots of cases including the appalling Explosion of Eypt Air Flight 804, no claim has been made from any group, leaving the actual perpetrator at large.

Most of the investigations on terrorist attacks are performed by human professionals based on domain knowledge and experts' opinion. We would like to propose an empirical model that learns patterns from the past terrorist attacks and identifies the most likely perpetrator given an unseen attack. A matured product would be a platform in which the human experts can enter information about an unclaimed terrorist attack, and the model would give out the most suspicious perpetrators. Such a platform would bring real value to the society because it helps the law enforcement detain the perpetrators before they could commit another crime and deters other potential attackers.

2 Data Mining Formulation

Our data mining problem is a supervised learning problem in which we take a dataset and produce an empirical model $f(x)$ which takes in a data entry x describing a terrorist attack and output the most likely perpetrator c from a set of known terrorist group denoted by C. More formally, we formulate our problem as following:

Given training set $D = \{X, Y\}$, where X is a $n \times d$ matrix, and $Y = \{y_1, y_2, ..., y_n | y_i \in C\}$. **Find** a function $f(x) : \mathbb{R}^d \mapsto c, c \in C$.

In our case, we will use the Global Terrorism Dataset(GTD) provided by University of Maryland, which documents most of the major terrorist attacks from

1970 to 2015. For GTD, $n = 156772$, $d = 137$, and $|C| = 3290$. GTD includes columns that documents the time (year, month, date), location (country, province, city), attack type (bombing, hijack, assassination, etc.), weapon (poisonous chemical, AK47, animals, etc.), target (nationality, category, etc.), consequence (casualty, property loss, wounded, etc.), and perpetrator identity (terrorist group name) of each attacks.

We will like face these following technical difficulties in our data mining problem:

- Most of the columns (106) contains null value, and 79 of them has more than 50% null. We need to figure a reasonable way to clean null columns and impute null values.
- Most of our columns are categorical with large number (> 100) of unique values. If we represent them in integer, we would potentially mislead our model by assuming order in these categories. If we represent each unique value as a binary feature, we will likely introduce too many dimensions.
- The label itself is categorical with > 3000 unique values. To treat label index as integer would assume order in terrorist group. To treat label in binary form would require us making > 3000 binary classifiers. One possible solution is to use deep learning model with a soft-max output layer. However, the data may be insufficient for deep learning models.
- Lots of columns (58) are in natural language which requires NLP techniques to extract features.