CS 5350/6350 Final Project Report

Jesus Zarate & Madeline MacDonald

2018-05-03

1 Project Overview

This semester we've focused our research on analyzing data from UFO sightings reported to the National UFO Reporting Center (NUFORC). Initially our main objective for this project was to identify similar patterns of alien visitations in the hope of isolating characteristics of different alien civilizations. Our original aim was predicting which type of alien was visiting based off of location, time, UFO shape, and other characteristics. As our analysis continued we discovered that because our dataset is so unconventional and the UFO reports so unreliable, we needed to transform our input data to find meaningful results.

This lead to a two part project: First, we applied machine learning techniques from class to analyze our data and make predictions for future instances. Second, we used clustering and other data analysis techniques to prune our data, in hopes of reducing the noise of our dataset and finding more meaningful predictions. Together these two aims form a multifaceted analysis of UFO activity throughout America, and have provided a lot of fun and some interesting results, while also giving us an opportunity to learn and practice.

2 Motivation

2.1 Real World Implications

Throughout this project we've had a lot of fun joking about the idea of aliens and UFO sightings, but the reality is that UFO sightings can actually contain important and useful information. Because of our First, UFO sightings can occur when highly unusual weather happens unexpectedly. Meteor showers, unexpected lighting, and odd clouds can all account for UFO sightings (Zimmer). Additionally, UFO sightings are frequently symptoms of mental illness and drug usage (Katharine J. Holden & Christopher C. French). From that perspective our UFO dataset is, in many aspects, a public health data set.

With this perspective, a hypothetical discovery that UFO sightings increase in a specific rural area during the summer could indicate that the warming air

is causing interesting weather phenomenon, and lead a weather researcher to analyze that city more closely. That same UFO sighting pattern could also mean that drug use increases with the warm weather, or that mental illness is affected by the longer daylight. For the purposes of our project we didn't conduct analysis on how our UFO data relates to weather or public health datasets, but we hope that someday this data could be used to help learn more about other public health and environmental issues.

3 Motivation for ML Techniques

Machine learning is the best approach for this project for a number of reasons. First, the dataset we have is incredibly large, spread out, and contains a lot of noise. We needed a way to analyze the data without over-fitting that could adapt to noise, while clustering together our data points on a number of different attributes. We also had to have the ability to map our categorical and empirical data into geometric and mathematical interpretations, like using written UFO descriptions as input to our algorithms, and traditional methods aren't up to the task.

Most importantly, our goal for this project is to use a large dataset to learn patterns of UFO activity and draw assumptions about the underlying causes and indicators, which is virtually the dictionary definition of a machine learning problem. Cluster lots of data highly dimensional data together, finding relationships between the data, and drawing big picture conclusions from thousands of individual data points makes this an ideal question for machine learning.

4 Our Solution

4.1 Sightings Credibility Scores

Because of the size of the data set, and the messiness and noise inherent to any self reported dataset, one solution we developed was to begin the project by analyzing the credibility of each sighting. To accomplish this, we assigned each sighting a credibility score, which was a heuristic determined by the number of similar reports, the time elapsed between the sighting and submitting the report to NUFORC, and how long the sighting was (a quick glimpse vs watching for 5 minutes).

We started by grouping together highly similar UFO sightings. We clustered our data across timestamp, latitude, and longitude, using the DBScan algorithm with a minimum group size of 3. This gave us clusters of sightings that occurred within the same time frame (usually between 2 weeks and 2 months), and within the same general town/city. From here, we took the size of the cluster of similar sightings, time elapsed before reporting, and duration time of the sighting, and we created a heuristic using a linear combination of these attributes. From this,

we assigned each report a credibility score, then filtered the data so that we only kept highly credible reports.

We manually checked a few of these clusters to ensure that our algorithms were working as expected, and found some entertaining and informative results. The first cluster examined was five sightings of unusual green lights flashing above Seattle during a 1 week span in the winter. This got us excited, and we read the comments left by the reporters, which all described exactly the same thing, up until the last sighting report which stated "I know they said there were going to be northern lights visible this week, but trust me, this green light was different." It gave us a laugh, but it also lends credit to the fact that our clustering algorithm was effective enough to cluster together similar weather phenomenon.

4.2 Analysis of Sightings

The initial problem we proposed to address was determining the number of distinct alien civilization that were involved visiting earth. By using the reported shape of the UFO as a label distinguishing alien civilizations, we ran an analysis over our filtered data to find a solution. First, we manually observed the data and discovered that even in highly similar sightings happening at the same time and the same place, people labeled UFOs slightly differently, e.g. a disk vs a saucer vs a sphere. To reduce the effect of human error we decided to cluster all of our credible data in two dimensions, the first being the reported UFO shape, and the second being the location and time cluster assignment from the credibility scoring. With these new clusters we were able to easily count the different reported UFO shapes and find the most prevalent ones that best described that cluster of sightings. We analyzed the number of sightings credited to each major UFO shape, and used this to determine the number of Alien Civilizations that we believe visit the US on a regular basis.

4.3 Prediction and Analysis

Prediction is at the heart of machine learning, and it was an integral part of our course project. For prediction, we chose to attempt to solve the problem of predicting what UFO shape would be seen, given a written description of the sighting and a location (latitude and longitude). We used two techniques taught in class, decision trees (via a Random Forest) and SVMs.

4.4 SVM

Following the discussion of SVM's in class, we used an SVM library to implement UFO shape prediction with SVM's. We began by vectorizing the written comments for each sighting, by converting each comment to a bag of words, and then using PCA to analyze the most relevant words. From this we selected the

top 20 words from each written comment and combined that vector with the latitude and longitude to get the input for our SVM. We used the UFO shapes as a label, and evaluated how accurately we could predict for an unknown future sighting the type of Alien ship that would arrive.

4.5 Decision Trees

Similarly to the SVM, we vectorized each comment using a bag of words and PCA, to select the top twenty words. Again, we combined the latitude and longitude of the sighting with this vector, to get the input for our Random Forest model. We ran our algorithm, using the UFO shapes as the label, and collected our results.

5 Experimental Results

5.1 Credibility Scores

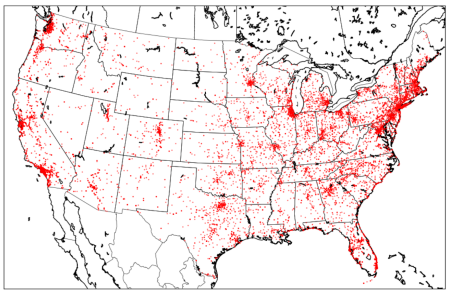

Figure 1: Input to credibility scoring: All UFO Sightings in the US

- Total US Sightings: 64896
- Initial [lat, long, time] clustering
  - 2452 unique clusters
  - Unclustered sightings: 17159
  - Clustered sightings: 47737
  - Average sightings per cluster: 19.483
- Credibility Score Filtering
  - Number of Removed Rows: 17685
  - Number of Credible Sightings: 47211

5.2 Sightings vs Shapes

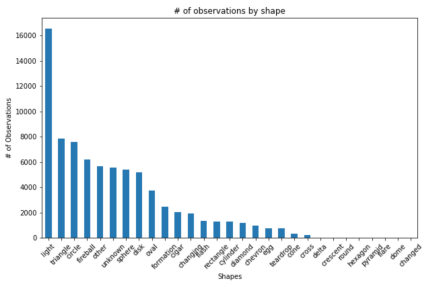

Figure 2: Shapes

5.3 SVM

Table 1: SVM

Approach	Accuracy
No Clustering	0.208397534669
w/ Clustering	0.211373504183
Clustering + Comments(PCA top n = 20)	0.269568567026

5.4 Decision Tree

Table 2: Random Forests

Approach	Accuracy
No Clustering	0.130277349769
w/ Clustering	0.17123795404
Clustering + Comments (PCA top n = 20)	0.31155624037

6 Analysis

By our measurements, clustering the data and filtering by credibility played a major role in improving our results. In all cases removing noisy and unreliable data lead to more accurate predictions, and when the credibility filtering was combined with vectorized comments as an input, we was huge improvements on accuracy. This it's highly interesting to see the improvement on Random Forests versus the SVM models used, but the near tripling in accuracy on the Random Forest model indicates that noisy data was having a far greater impact on the Random Forest model than it was on the SVM, which makes sense considering the underlying logic behind the algorithm design.

The reason we believe our experimental results were so inaccurate, even with clustering and vectorized comments, is that our data comes from unreliable sources. Our sighting info is entirely self reported, and because UFO sightings are so stigmatized, uncommon, and highly correlated with other concerning factors like drug use or mental illness, we can't trust that even our most highly processed data will be accurate. Our predictions were a a great learning and process and very entertaining, and performed better than expected, so we feel pleased with our results.

7 Future Work

Following this, we'd be interested in exploring further how we can transform our input data to be more descriptive. We'd like to consider even more factors when performing our initial data cleaning and credibility scoring, such as using vectorized comments and euclidean distances between comment vectors within [lat, long, time] clusters to produce more meaningful results.
We'd also like to explore the intersection of this dataset with other datasets related to public health, or to unusual weather forecasts. We believe there are many interesting discoveries to be made concerning the correlation between those two factors.

7.1 Works Cited

"Alien Beings." Mutual UFO Network, MUFON, www.mufon.com/alien-beings.

Katharine J. Holden & Christopher C. French (2010) Alien abduction experiences: Some clues from neuropsychology and neuropsychiatry, Cognitive Neuropsychiatry, 7:3, 163-178, DOI: 10.1080/13546800244000058

Zimmer, Troy A. "Social Psychological Correlates of Possible UFO Sightings." ProQuest, Journal of Social Psychology, 1 Aug. 1984, search.proquest.com/openview/.