Look! UFO Proposal

Jin Yan, Lei Tian, Yao Yao, Jing Qian

VISION

UFO reports have been around us for decades, and people tell their UFO encounters in thousands of different ways. Predicting anything with UFO is intriguing since these tiny fickle light dots leave us little trace behind. In this proposal we will investigate this topic in three steps. Firstly, we will figure out correlations between UFO sightings and other source of information such as geometry, weather and population. Secondly, based on statistic analysis and machine learning techniques, we can develop models to detect fake reports. Finally, we will build a web application to visualize our result, enabling users to report their own sightings.

DATA

Data to use:

We collect UFO reports data from http://www.nuforc.org/, GIS data from Google's API https://maps.googleapis.com/maps/api/geocode/, and weather data from https://darksky.net/ and U.S population data from https://www.census.gov/.

How to collect data:

1. We write python scripts to scrap UFO reports data from NUFORC's online database,
2. use Google's API to get latitude and longitude data,
3. use Darksky's API to get weather data
4. download CSV files from U.S.'s Census Bureau to get population data.

How big is data:

UFO data: about 96000; GIS data: about 19000; Weather data: about 89000; Population Data: about 200

How to clean, merge and load data:

To clean UFO reports data, we use regular expressions to remove useless HTML tags, thus remove rows containing columns of “Unknown” or “Unspecified” and transform date format to yyyy, mm, dd, hh:MM:ss. We use UFO reports data event ID as a key to merge UFO data and location data. We create a sqlite3 database with 4 relational tables “events”, “weathers”, “populations” and “areas”, and load data from CSV files into database.

METHODOLOGY

Model training:

Machine learning is the central technology to use. Since our data can be divided into two parts, numerical features and text description, we can train models based on these different sub-features. We use SVM to classify each sample for numerical features and vectorize text as input for SVM and Logistic Regression models for event description. Since our data are really imbalanced (minority of our data is labeled as fake), we need to assign class weight separately. Currently, this weight is chosen by comparing cross validation score. For now, we have reached around 80% judge score in average.

Another label we will investigate is the duration of UFO appearance. UFOs may appear for a different duration of time: some are transient while others last for hours, if not days. Is it possible for us to tell their duration based on the weather conditions? To answer this question, we can use machine-learning algorithms described above to determine their relationships. Using different features from the weather, population, and geography data we can train our models to predict possible duration time in the event UFOs appear.

Statistic analysis:

Sometimes, machine learning algorithms alone is insufficient to tell the relationship among various data. We need a direct way to review the relationships them. Using statistic methods such as descriptive statistics and pearson correlation we can generate an overall review of the over 90,000 samples of data we have. A descriptive statistics reports a mean, mode, standard deviation, percentile, kurtosis and skewness of our collected data, therefore helping us further investigate distribution of data. Pearson correlation finds the relationship among different arrays of data and yields a number range from 0 to 1 for relatedness (0 as not related, 1 as very related). Consequently with various statistical tools, we can analyze our data from a different perspective that possibly gives useful insight to our future works.

Visualization:

Data is abstract. Visualizing data enables researchers to see the relationship “on the fly” instead of crawling through numbers. There are numerous visualization technologies we can incorporate to show our data. D3, for example, can serve as the core technology for visualizing our result, as well as Python matplotlib and GoogleMap. We have now created a US map figure reflects how many sightings each year. This figure may change year by year. Some other statistical figures, such as histogram to show distribution of UFO sighting numbers according to hours, pie charts to show distribution of different shapes, are also created and posted. We also use GoogleMap to show heatmaps about accumulated number of sightings around US.

Web application:

Web application integrates all our results together, and offers a way to interact with users. For the back-end, we use sqlite3 database to store reported data and Node.js as server. Every time a new report is submitted, Node,js executes fake_detection.py to generate probability of truth.

For the front-end, there are two functions: let user to submit a form of report and view our statistical analysis. After user submits report form, Node will first update it to our database, and then send probability of truth back, and we give it to users as a feedback. In viewing statistical analysis, we display linked D3 charts to user.

FUTURE WORK:

Create a big linked D3 chart. There are maps, histogram and other charts to show statistic information of UFO sightings. All these charts are correlated to each other and would change simultaneously. Then, adding this visualization into our web application. We believe users will get to understand UFO sighting better in this way.

We will create a descriptive stats for the UFO appearing duration data, shape data, geolocation and demographics data, and search for possible significant results and observe patterns. By looking at kurtosis, skewness, mean and standard deviation of data, we will be able to read patterns such as if UFOs like certain weather more than other or they have a tendency to appear in certain areas. If enough statistical significance is determined, we can even further push for prediction of future events. Other static methods such as pearson correlation will be used to find any link among all the data features we have, therefore further deduce from the findings above.