Using Satellite Imagery to Predict Wildfires (Natural Sciences)

Andrew Lazar
atlazar@stanford.edu

Rahul Srivastava
srivastr@stanford.edu

June 2, 2021

Abstract

For the CS229 project, we use publicly available information like satellite imagery, historical wildfire incidents, and other information to explore whether it can be used to help predict future wildfires.

1 Introduction and Problem Statement

Wildfires are an increasingly common phenomena that impact both human and natural ecology. Studies show that over the past 10 years (2011-2020), there were an average of 62,693 wildfires annually and an average of 7.5 million acres impacted annually [1] in the United States. Further, the top three years with largest wildfire acreage burnt were 2020, 2015, and 2017 respectively [1]. The ability to predict potential wildfires using open-source satellite imagery and other data can help public safety and emergency services prepare for fire mitigation (e.g., controlled burn, slash piles, etc.) as well as guide firefighting resource deployment and evacuation.

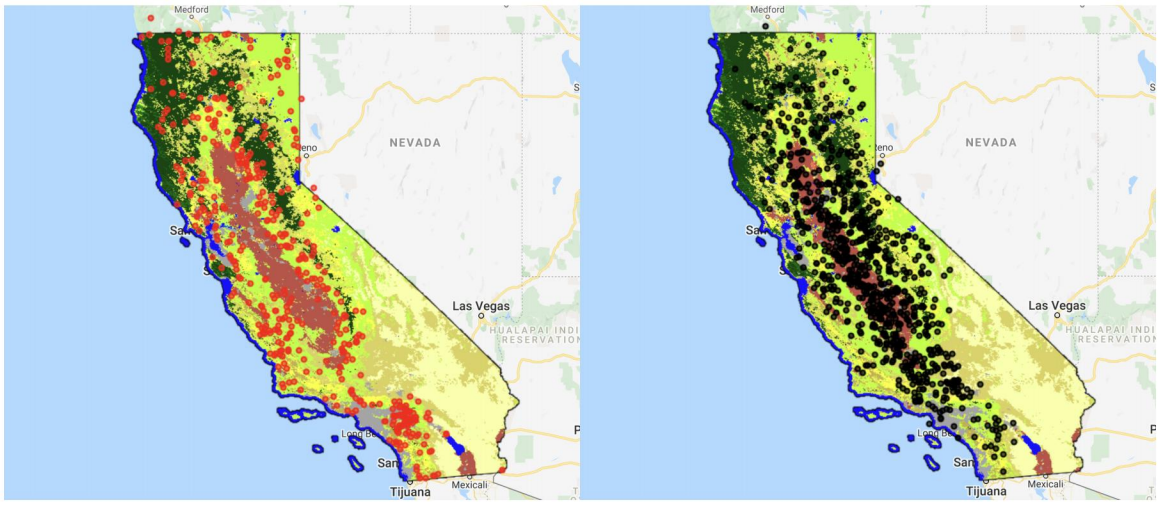
Formally, we pose our problem as a classification model. We sample different coordinates in California and predict the possibility of a fire occurring in the area enclosing that coordinate[1] within the next year. We will use historical wildfire data to determine the labels for each location. Training features will be a variety of open source satellite data. We will use these labels and features for training, validating and testing a machine learning model. With this model we hope to be able to infer the possibility of future wildfires using existing data.

[1]For simplicity, we consider 1000m x 1000m geographical areas, with the coordinate as the centroid of this area.

2 Related Work

There has been research on similar subjects utilizing satellite imagery to predict poverty [2], oil spill detection [3], and various natural science mappings [4]. These studies formulated the problem in a regression framework using small variations within each image. We intend to build upon previous work to find if there are efficient and successful ways to predict natural disasters, specifically wildfires, using similar techniques.

3 Datasets



(a) Wildfire locations (b) "Not Wildfire" locations

Figure 1: Map of the dataset labels for Wildfires in 2017.

For labeling, we used wildfire data from 2017 to establish a training/validation/test set of known events. For simplicity, we use the label 0 for non-incidents (i.e., no fire at the location) and label 1 for incidents (i.e., fire at this location). To collect fire incident labels, we scraped the CAL FIRE website [5] for fire reports and captured the latitude/longitude information for each incident. We generated the non-incident labels by randomly sampling locations within the state of California and ensuring that they were at least 1 km away from the closest fire incident location.

Available learning features included open source satellite data hosted by various organizations:

- MODIS land cover collection provides global land cover types at yearly intervals (2001-2016) derived from six different classification schemes [6].

- MODIS land surface temperature collection provides daily land surface temperature (LST) and emissivity values in a 1200 x 1200 kilometer grid [7].
- USGS ground elevation image contains elevation data for the globe collected from various sources [8].

We split the labels/learning features into a 70/20/10% split for training/validation/test datasets respectively.

4 Methods

We used Google Earth Engine [9] as a cloud computing platform to process geographical information as well as satellite imagery. Earth Engine hosts datasets and also provides APIs and other tools to analyze these datasets. To fully leverage these features we decided to use the Earth Engine Code Editor which is a web-based IDE for the Earth Engine JavaScript API.

For this project we followed the following workflow:

1. Collected training data. This includes scraping data from primary sources, assigning labels based on incident reports, and ingesting these labels into Google Earth Engine.
2. Within Google Earth Engine, we assemble features (from previously mentioned datasets) which have a property that stores the known class label and properties storing numeric values for the predictors.
3. Next, we instantiate a classifier (discussed below) and select associated hyperparameters if necessary using the validation set.
4. Train the classifier using the training data.
5. Classify each location for fire incidents.
6. Estimate classification error with independent test data.

The primary feature selected for the learning algorithm is the land cover classification for each location. Intuitively, land cover should be a good indicator of fire hazard, as Shrublands and Grasslands will have a higher risk of fire, whereas Barren and Permanent wetlands will have a lower risk of fire.

We compare the performance of four supervised classification algorithms in order to select the machine learning framework to predict wildfires. In the next section we compare results and find the most efficient and effective model. A description of each is as follows:

- **General Classification and Regression Tree (CART) algorithm**: This basic algorithm uses a tree representation to denote attributes and class labels based on each internal node of the "tree". The advantages of this method include less effort for data preparation, and a minimal effect from missing values in the data. It's main disadvantage is that a small change in the data can cause a large change in the structure of the decision tree causing instability, as well as being potentially expensive with increasing complexity [10].

- **Support Vector Machine (SVM)**: As shown in class, SVMs choose the hyperplane that represents the largest separation between the two classes. The SVM algorithm we implemented utilized a Radial Basis Function (RBF) kernel. The advantage of this kernel is that it shows better generalization performance than general RBF networks when the training data size is relatively small [11].

- **Random Forest**: This is an "ensemble classifier" that consists of many decision trees and outputs the majority vote of individual trees. The method combines bagging idea (a method for generating multiple versions of a predictor and using these to get an aggregated predictor) and the random selection of features. The advantages of this algorithm include a highly accurate classifier which runs efficiently on large data sets. Additionally, it has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing. It's main disadvantage is that random forests are prone to over-fitting for some datasets [11, 12].

- **Gradient Boosted Trees**: This method uses "boosting" which is "an optimization algorithm on a suitable cost function. In particular, the boosting algorithms can be abstracted as iterative functional gradient descent algorithms". This algorithms chooses a function that points in the negative gradient direction. While an effective and efficient model, it is also prone to over-fitting [11, 12].

5 Results and Discussion

The JavaScript code for all the experimental results provided here is hosted at `https://code.earthengine.google.com/?accept_repo=users/srivastr/cs229`. Defining the *error matrix* as a matrix with entry (i, j) as the number of observations with true label i and predicted label j[2].

The error matrix for **CART** algorithm on the test set is: $\begin{bmatrix} 181 & 18 \\ 120 & 13 \end{bmatrix}$ and the test accuracy was 0.5843.

[2]As noted previously, label 0 is assigned to locations that have not seen fire in the observed year and label 1 are the locations that have had a fire incident.

The error matrix for **Random Forest** algorithm of size 100 on the test set is: $\begin{bmatrix} 195 & 4 \\ 119 & 14 \end{bmatrix}$ and the test accuracy was 0.6295.

The error matrix for **SVM (RBF)** algorithm with $\gamma = 0.5$ and cost $= 10$ on the test set is: $\begin{bmatrix} 142 & 57 \\ 48 & 85 \end{bmatrix}$ and the test accuracy was 0.6837.

The error matrix for **Gradient Tree Boost** algorithm with 10000 decision trees on the test set is: $\begin{bmatrix} 143 & 56 \\ 47 & 86 \end{bmatrix}$ and the test accuracy was 0.6898.

As expected, the CART algorithm performed the poorest due to its lack of complexity, while the SVM and Gradient Tree Boost algorithms performed the best. Even with the best performing predictors, the results are not yet ideal for a production-ready predictor. Given the limitations of the datasets and classifiers, the test accuracy was better than expected and are expected to improve with the addition of more features and datasets in future tests.

6 Conclusion and Future Work

Considering the simplicity of the datasets and models used, and the accuracy noted thus far, we see great potential in these algorithms for wildfire prediction. Some of the datasets that we used contain multiple images over a period of time. For instance, we can develop a time-series statistic that tracks the change in landcover as an additional learning feature. There are a number of other datasets that contain features like temperature, wind, human settlements, roads etc., that might improve the quality of the model.

On the model side, with additional features, we can add more sophisticated classifiers that use neural networks with hidden layers. In addition, we can also add multiclass prediction models by categorizing the severity/size of a fire. Finally, with new datasets/images available daily, there is a potential to develop a ML pipeline that ingests data daily and provides near-realtime prediction for wildfires.

Both of us contributed equally to this project.

References

[1] K. Hoover and L. A. Hanson, "Wildfire Statistics." Congressional Research Service Report, Jan. 2021. [Online]. `https://crsreports.congress.gov/product/pdf/IF/IF10244/49`.

[2] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using Publicly Available Satellite Imagery and Deep Learning to Understand Economic Well-being in Africa," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[3] M. Krestenitis, G. Orfanidis, K. Ioannidis, K. Avgerinakis, S. Vrochidis, and I. Kompatsiaris, "Oil spill identification from satellite images using deep neural networks," *Remote Sensing*, vol. 11, no. 15, 2019.

[4] S. Bumm, "Using Satellite Images to Determine AQI Values in California." CS229 Project Report, 2019. [Online]. `http://cs229.stanford.edu/proj2019spr/report/22.pdf`.

[5] California Department of Forestry and Fire Protection, "Cal fire homepage." `https://www.fire.ca.gov/incidents/`, 2021.

[6] M. Friedl and D. Sulla-Menashe, "Mcd12q1 modis/terra+aqua land cover type yearly l3 global 500m sin grid v006 [data set]." NASA EOSDIS Land Processes DAAC, 2019. [Online]. `https://lpdaac.usgs.gov/products/mcd12q1v006`.

[7] Z. Wan and S. Hook and G. Hulley, "Mod11a1 modis/terra land surface temperature/emissivity daily l3 global 1km sin grid v006 [data set].." NASA EOSDIS Land Processes DAAC, 2015. [Online]. `https://doi.org/10.5067/MODIS/MOD11A1.00`.

[8] NASA JPL, "Nasa shuttle radar topography mission global 1 arc second.." NASA EOSDIS Land Processes DAAC, 2013. [Online]. `https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL1.003`.

[9] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017.

[10] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. Taylor & Francis, 1984.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics, Springer, 2009.

[12] H. Li, "Smile." `https://haifengl.github.io`, 2014.