US Wildfire Prediction

Ethan Tucker (eht2122), Ritvik Khandelwal (rk3213), Dieter Joubert (dj2574), Xinyu He (xh2562), Keli Wang(kw3015)

With global warming increasing over time, a number of studies have shown resultant changes to the features of wildfires, including increases in season length, frequency and burned area.[1,2] With such progression, there are likely to be severe economic and social consequences.[3] Available data on wildfires may be useful in clarifying the details and scope of these consequences, while also informing on what methods will prove most effective in counteracting the adverse effects.

In particular, analysis of available wildfire data, may be able to answer relevant questions such as:

- Is the cause of a wildfire predictable, given the limited features in available data?
- What level of risk do various geographic areas have for wildfires throughout each year?
- Over time, is wildfire frequency and intensity increasing in all areas, or is there variability? Is there a predictable temporal dependence to the increase?
- With respect to intensity and frequency, how quickly are different regions able to get fires under control, and how has this changed over time?

To attempt answering these questions, there is data available on Kaggle (and collected through the national Fire Program Analysis (FPA) reporting system) that catalogs the features describing 1.88 million US wildfires that occurred between 1992 and 2015. This dataset includes 51 features for each wildfire, some of the most notable of which are: fire cause, discovery date and time, containment date and time, fire size (in acres) and class, latitude and longitude, owner/entity responsible for managing land where the fire originated, National Wildfire Coordinating Group unit ID and associated details.

As there are a number of questions to be answered with this data, it will prove helpful to use various ML techniques, and in exploring the data, it may become apparent that one technique is more ideal. Prior to training there will be a need for feature engineering, and in particular feature creation in the case of answering the second question above. How can a risk level feature be synthesized from available features? Considering the first question above, there will also be a need for multiclass classification, for which there are a number of options: Decision trees (i.e. Random Forest), Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Naive Bayes, etc. For the third question, it may be informative to do time-series forecasting (using exponential smoothing, ARIMA, RNNs etc.) to predict the likelihood of fires occurring over time, and to understand quantitatively the temporal-dependence of increasing frequency and scale.

References

[1] Reidmiller, D. R., et al. "Fourth national climate assessment." *Volume II: Impacts, Risks, and Adaptation in the United States, Report-in-Brief* (2019).

[2] Westerling, Anthony LeRoy. "Increasing western US forest wildfire activity: sensitivity to changes in the timing of spring." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1696 (2016): 20150178.

[3] Gill, A. Malcolm, Scott L. Stephens, and Geoffrey J. Cary. "The worldwide “wildfire” problem." *Ecological applications* 23.2 (2013): 438-454.

[4] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA*FOD*20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4 (Kaggle source: https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires?datasetId=2478&sortBy=voteCount)