Project Status Report

Abstract

On average, the state of California loses over 100 million dollars and 218,000 acres of land each year due to wildfire damages. Currently California only employs fire prevention methods such as restricting certain kinds of fuels, controlled fires, and fire education to curb the damage of wildfires; however, using machine learning to learn fire patterns can help fire departments properly allocate resources and take targeted measures to preventing large wildfires. Our project aims to predict the likelihood and severity of wildfires by location in California using the Maximum Entropy classifier. .

1. Status Report

1.1. Data Selection Progress

We chose our feature space based on factors that we knew or guessed would affect the likelihood and severity of wildfires, based on previous work that has been done in this field. In the past, research has been done to create separate models for weather and human factors; however, no significant work has been done to create a model using both of these features at once. Previous work has also mostly focused on large wildfires (with regards to weather) or relatively small structural fires (with regards to human factors, for instance those that affect arson rates). Instead, we gathered instance data for a range of wildfire sizes and matched them with environmental and human factors to create a cohesive model. Some examples of the features we have collected data for include temperature, precipitation (immediate and seasonal trends), drought status, foliage type/density, population density in the area, nearby camp sites, and wealth of the area, among others.

For obtaining an actual data set we found that the National Wildfire Coordinating Group has created and maintained a database of wildfire instances since 1973. This database logs the start date, coordinates, fire type, and cause of the fire. Using this database we were able to obtain over 4,800 wildfire instances. Many of the features that we proposed for our project (especially human factors) were not included in existing fire databases. As a result, we had to utilize a number of unrelated databases to collect the data we needed and then synthesize it with the information from the NWCG database. For instance, to collect information related to wealth we consulted the California Department of Finance, and to find information about terrain and foliage we utilized the Multi-Resolution Land Characteristics Consortiums National Land Cover Database.

Because we were gleaning data from other sources, we restricted the resolution to the county level and year. Features such as population density, unemployment rates, average income, terrain type, etc., were calculated as averages by county for each year. These features were then assigned to each wildfire based on the county and year in which it occurred.

1.2. Algorithm Selection

We chose to use the Maximum Entropy classifier because it excels at training models on presence only data. All of our instances correspond to actual wildfire occurrences, and the available data for non-occurrences is both overwhelming in size and not very meaningful to our model. After reading through other presence only data papers, we feel confident we will be able to obtain reliable results through MaxEnt.

1.3. Steps Going Forward

Thus far, our efforts have mostly been focused on obtaining as much data as we can in order to make our personal database as complete as possible. Since we have not been able to find an existing dataset that fits our needs, we have had to create one ourselves. After gathering a subtantial amount of data, we have set up python to read in our data into testing and training arrays. We still need to compare different implementations of the max entropy classifier and determine which returns the most reliable results, and then train, test, and evaluate our model.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.